Effets audionumériques adaptatifs : théorie, mise en œuvre et usage en création musicale numérique.

Vincent Verfaille

Université Aix-Marseille II

Doctorat ATIAM – Acoustique, Traitement du signal et Informatique Appliqués à la Musique
École Doctorale 353 – Mécanique, Physique et Modélisation

Effets audionumériques adaptatifs : théorie, mise en œuvre et usage en création musicale numérique.

THÈSE

présentée et soutenue publiquement le Vendredi 12 Septembre 2003

pour l'obtention du

Doctorat de l'Université Aix-Marseille II

(spécialité ATIAM)

par

Vincent VERFAILLE

Composition du jury

Présidente	Myriam Desainte-Catherine	LABRI, Université Bordeaux 1
Rapporteurs	Philippe Depalle	SPCL, Université McGill, Montréal
	Xavier Serra	MTG, Université Pompeu Fabre, Barcelone
Invités	Emmanuel Favreau	INA-GRM, Paris
	Patrick Boussard	GENESIS S.A., Aix-en-Provence
Directeur de thèse	Daniel Arfib	CNRS-LMA, Marseille

Laboratoire de Mécanique et d'Acoustique — UPR 7051
Centre National de la Recherche Scientifique

Résumé

Ce travail de thèse porte sur la théorie, la mise en œuvre et les applications musicales des effets audionumériques adaptatifs. Dans la première partie, nous plaçons le sujet dans le contexte des transformations sonores. Un grand nombre de techniques de traitement du signal sonore numérique se complètent et fournissent un ensemble d'algorithmes permettant de transformer le son. Ces transformations sont appliquées selon les dimensions perceptives du son musical, à savoir la dynamique, la durée, la hauteur, la spatialisation et le timbre. Pour quelques effets, les contrôles évoluent de manière automatique ou périodique, et ce contrôle est intégré à l'algorithme. Le contrôle offert à l'utilisateur porte sur les valeurs de certains paramètres de l'algorithme. Il se réalise à l'aide de contrôleurs réels, tels des potentiomètres, des interrupteurs, ou à l'aide de contrôleurs virtuels, telles les interfaces graphiques sur écran d'ordinateur. En synthèse sonore, l'un des sujets majeurs d'étude à l'heure actuelle est le *mapping* : il s'agit de savoir comment mettre en correspondance les paramètres d'un contrôleur gestuel et les paramètres d'un algorithme de synthèse. Notre étude se situe à l'intersection entre les effets audionumériques, le contrôle adaptatif et gestuel, et la description de contenu sonore.

Dans la seconde partie, nous présentons les effets audionumériques adaptatifs tels que nous les avons formalisés et développés. Ce sont des effets dont le contrôle est automatisé en fonction de descripteurs sonores. Nous avons étudié puis utilisé de nombreux algorithmes de traitement, certains en temps-réel et d'autres hors temps-réel. Nous les avons améliorés afin de permettre l'utilisation de valeurs de contrôle variables. Une réflexion a été menée pour choisir une classification des effets qui ait du sens pour le musicien ; elle a logiquement abouti à la taxonomie perceptive. Parallèlement, nous avons étudié les descripteurs sonores et les moyens de contrôle d'un effet, par le son et par le geste. Nous avons rassemblé de nombreux descripteurs sonores, utilisés en psychoacoustique, en analyse-synthèse, pour la segmentation et la classification d'extraits sonores, et pour la transcription automatique de partition. Nous proposons un contrôle généralisé pour les effets adaptatifs, hiérarchisé en deux niveaux. Le premier niveau de contrôle est le niveau d'adaptation : le contrôle de l'effet est effectué par des descripteurs du son, à l'aide de fonctions de *mapping*. Nous indiquons des fonctions de conformation (fonctions de transfert non linéaires) permettant de transformer la courbe d'évolution temporelle d'un descripteur, des fonctions de combinaisons des descripteurs ainsi que des fonctions de conformations spécifiques des paramètres de contrôle. Le second niveau de contrôle est celui du contrôle gestuel : le geste agit sur les fonctions de *mapping*, soit sur la combinaison, soit sur la conformation spécifique des contrôles. De cette étude, il ressort non seulement une généralisation du contrôle des effets audionumériques, mais aussi la réalisation d'outils pour la composition, et leur utilisation en situation musicale. De nombreuses expériences et illustrations sonores ont été réalisées, parmi lesquelles une spatialisation adaptative contrôlée par une danseuse, et un équalisateur stéréophonique adaptatif. Les expériences confirment l'intérêt d'un tel contrôle adaptatif et gestuel, notamment pour modifier l'expressivité d'une phrase musicale, ou pour créer des sons inouïs.

Mots-clefs : effets audionumériques, traitements sonores, transformations sonores, analyse–transformation–synthèse, contrôle automatique, contrôle adaptatif, contrôle gestuel, descripteurs sonores, *mapping*, perception, jeu musical, expressivité.

Abstract

This PhD thesis addresses the theory, the implementation and the musical use of adaptive digital audio effects. In the first part, we situate the subject in the context of sound transformations. There exist a great number of signal processing techniques that complete each other and provide a complete set of algorithms for sound transformations. These transformations are applied according to the sound perceptive dimensions, namely dynamics, duration, pitch, spatialisation and timbre. For some effects, the control evolves in an automatic or periodic way, and this control is integrated to the algorithm. The control let to the user is about some parameters of the algorithm. It is given by real controllers, such as knobs, switches, or by virtual controllers, such as the graphical interfaces on computer screens. A main interest in sound synthesis today is the mapping: the topic is to find how we can map the gesture transducer data to the parameters of the synthesis algorithm. Our study is situated at the intersection between digital audio effects, adaptive and gestural control, and sound features.

In the second part, we present adaptive digital audio effects, in the way we formalised and developed them. These effects have their controls automated according to sound features. We studied and used a lot of processing algorithms, some in real-time and some out of real-time. We improved them in order to use varying control values. A reflexion was carried out in order to choose a meaningful classification to the musician: the perceptive taxonomy. In parallel, we studied sound features and descriptors, and the ways to control an effect, by the sound and by gestures. We brought together numerous sound features that are used in psycho-acoustics, for analysis-synthesis, for sound segmentation, for sound classification and retrieval, and for automatic transcription of music. We propose a generalised control for adaptive effects, structured with two levels. The first control level is the adaption level: sound features control the effect with mapping functions. We give a set of warping functions (non-linear transfer functions) allowing transformations of the evolution of sound feature curves; we also give feature combination functions and specific warping functions used to warp a control curve according to specific rules. The second control level is the gesture control, which is applied onto the mapping functions between sound features and controls, during combination or during specific warping. This study provides a generalisation of the control of digital audio effects, as well as the conception of toolboxes for composition, and their use in musical context. Numerous experiments and sound examples have been made, among which an adaptive spatialisation controlled by a dancer, and an adaptive stereophonic equaliser. The experiments confirm the interest of such an adaptive and gestural control, for example to change expressiveness of a musical sentence, or to create new sounds.

Key Words: digital audio effects, sound processing, sound transformations, analysis-transformation-synthesis, automatic control, adaptive control, gestural control, sound features, mapping, perception, musical expressiveness.

Remerciements

Exercice bien difficile que celui des remerciements. Je demande l'indulgence à tous, en espérant n'oublier personne.

Je tiens à remercier Daniel Arfib pour m'avoir proposé ce sujet, et encadré avec l'intelligence et le cœur dont il sait faire preuve, ainsi que Jean-Claude Risset pour l'intérêt porté à mon travail. Je tiens aussi à remercier le personnel du Laboratoire de Mécanique et d'Acoustique et l'équipe APIM ; chercheurs, ingénieurs, administratifs, doctorants et stagiaires que j'ai côtoyés et avec qui j'ai beaucoup appris. De même, merci aux chercheurs avec lesquels j'ai pu collaborer, à Paris, Barcelone ou Hambourg, pour la richesse des échanges.

Je remercie ma famille, qui a su malgré la distance m'encourager, me supporter et m'envoyer des ondes positives depuis la Mayenne, la Charente-Maritime et la Nouvelle-Calédonie. Il m'a semblé que la période de la thèse était par moments aussi peu facile à vivre pour eux que moi.

Je remercie le Centre National de la Recherche Scientifique et la Région Provence-Alpes-Côte-d'Azur pour avoir cofinancé la bourse de thèse qui m'a permis de réaliser cette étude dans d'excellentes conditions.

Merci aux étudiants en Electroacoustique du Conservatoire Nationale de Région de Marseille ainsi qu'au compositeur et enseignant Pascal Gobin pour les nombreuses réflexions, discussions sur la musique, qui m'ont ouvert encore plus à l'électroacoustique.

Je n'oublie pas les amis que j'ai rencontrés à Marseille et dans ses environs, notamment par la pratique de la musique irlandaise avec Derry Liam et de l'improvisation théâtrale avec la Ligue d'Improvisation Phocéenne (LIPHO). Merci encore aux amis de plus longue date, pour qui le temps et l'éloignement ne sont pas des raisons valables pour que l'oubli s'installe.

Merci à Christian Puech de l'Association Pour l'Emploi des Cadres (APEC) de Marseille, aux organisateurs des Doctoriales du Grand Sud en 2002, à Gary Burkhart, dont les enseignements et réflexions sur le doctorat ont dépassé de loin le cadre fixé au départ.

Merci à tous ces musiciens, ces artistes qui ne me connaissent pas, mais que j'ai pris plaisir à découvrir, tant dans le jazz et les musiques improvisées que dans la musique contemporaine et dans la musique électroacoustique, durant ces années de réflexion et d'expérimentation, et durant les mois de rédaction du présent document. Leur inspiration m'a été d'un grand secours.

Merci enfin à tous ceux que je ne cite pas ici mais que je n'oublie pas, avec qui mes rapports furent aussi divers qu'enrichissants.

Je dédicace cette thèse
à cette part de rêve
que nous gardons tous en nous,
plus ou moins enfouie.

AVERTISSEMENT CONCERNANT LA VERSION ÉLECTRONIQUE :

La version électronique de ce document a été réalisée sous LaTeX2e au format pdf. Aussi, les références en tous genres sont “cliquables” : table des matières, mini-tables des matières, figures, sections et sous-sections (en rouge), bibliographie (en vert), sons (♫♪ en bleu) et vidéos, adresses url (en cyan)...
Bonne navigation !

Table des matières

Des outils de création qui évoluent : les effets audionumériques

Parmi les machines à musique futures, on peut penser qu'on trouvera non seulement des "instruments", mais de véritables "outils de création" susceptibles de confier à l'utilisateur des responsabilités musicales graduées, à divers niveaux : par exemple, reproduire simplement une musique préparée à l'avance – rôle d'auditeur –, ou bien interpréter une musique – rôle d'instrumentiste –, ou encore se contenter de régir le temps et de doser les équilibres – rôle d'orchestrateur, d'instrumentiste ou de luthier –, voire son texte même, ses articulations, ses détails – rôle de compositeur. De tels outils ne seraient donc pas seulement destinés à reproduire, ils permettraient de produire, de créer, de modeler, de sculpter et d'animer à sa guise le son musical.
Jean-Claude Risset [Risset, 1986]

Les effets audionumériques et la composition musicale

Les compositeurs se sont toujours attaché à représenter le son de manière à mieux le contrôler en lui imposant des conduites qui leur sont propres, via des règles d'écriture et de composition. Le premier outil a été la partition, permettant d'indiquer la hauteur des mélodies monophoniques. Ensuite, la polyphonie, les notions de durée et de rythme y ont été ajoutées, puis les nuances. Plus récemment, la composition du timbre a nécessité des écritures plus complexes, indiquant les modes de jeu, les modifications à apporter pour obtenir le son désiré par le compositeur. La composition de l'espace, donnée par les positions des instruments dans un ensemble et dans un lieu, ne date pas d'hier, mais ne dispose que depuis peu d'outils d'écriture.

La fabrication de sons de synthèse a fait son apparition au début du XX^{e} siècle, puis les techniques de studio ont permis d'appliquer des traitements sonores. La synthèse pure et les traitements sonores ont été employés dans la composition comme prolongement logique des outils mis à la disposition du compositeur [Bernardini and Rudi, 2002]. Depuis les années 1960, la synthèse numérique sur ordinateur, initiée par Max Mathews [Mathews, 1969], a pris place parmi les moyens de traiter le son. L'ordinateur est aujourd'hui devenu le principal outil de synthèse sonore alternatif aux instruments classiques. De nombreux systèmes et environnements de transformations sonores ont été développés partout dans le monde, notamment en France au GRM[1] [Geslin, 2002] puis à l'IRCAM[2]. Ces outils permettant de modifier en finesse des paramètres du son numérique, que ce soit

[1] GRM : Groupe de Recherches Musicales
[2] IRCAM : Institut de Recherche et de Coordination Acoustique–Musique

sa hauteur, sa durée, sa dynamique, son timbre, sa spatialisation [Risset, 2002]. Cependant, les techniques évoluent, de même que les courants musicaux et leurs besoins : il reste encore beaucoup de place à l'inventivité.

Le son et sa perception

Les portes de la perception... Lorsqu'on entend des sons, différents processus s'enchaînent pour qu'à partir d'un processus physique, une image mentale se forme. Une onde de pression acoustique parcourt une certaine distance, dans un certain milieu (l'air, l'eau) puis parvient à l'oreille : elle est analysée par cet organe perceptif en différentes étapes, grâce à différents étages d'analyse (cochlée, intégration dans le système nerveux, cortex). On peut alors décrire le son selon différents attributs de cette perception que l'on a, d'un point de vue musical : dynamique, hauteur, durée, timbre, spatialisation ; mais aussi d'un point de vue cognitif : reconnaissance de formes musicales, de voix ou d'airs connus, de l'expressivité, d'allusions, du contexte, etc.

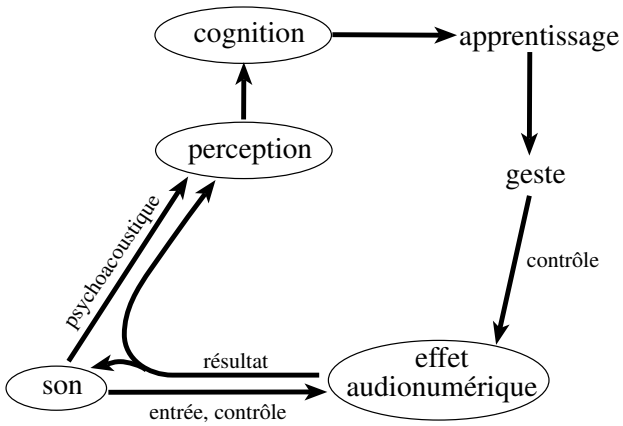


FIG. 1 – *Relations entre son, traitement sonore et perception du son.*

Il existe un lien étroit entre les traitements sonores, la perception et la cognition. En effet, les traitements modifient le son selon une ou plusieurs dimensions de la perception auditive. Ensuite, les processus physiques de modification simulés par un traitement peuvent être reconnus en tant que processus. On reconnaît un écho, une réverbération et en même temps le phénomène physique qui en est la cause, en l'occurrence l'effet de salle dans cet exemple. D'ailleurs, on utilise les effets audionumériques en musique populaire en fonction de connotations souvent inconscientes [Lacasse, 2000]. Il existe des liens entre traitement, perception et cognition, *cf. fig.* 1, qui induisent la possibilité d'un apprentissage du contrôle de l'effet (cognition) en vue d'obtenir un résultat sonore précis (geste et son), et ce par une boucle de rétroaction (*feedback*).

Effets et synthèse : des contraintes de flexibilité et de réalisme

Les effets et la synthèse sonore peuvent se qualifier en termes de flexibilité et de réalisme [Wyse, 1997]. Ces deux qualités ne doivent pas être vues de manière statique, mais plutôt évolutive dans le temps, *cf. fig.* 2. À ses débuts, l'informatique musicale permettait de réaliser de la synthèse sonore de manière flexible, mais les sons obtenus pouvaient manquer de réalisme. D'autre part, les effets et traitements sonores donnaient des sons souvent réalistes, puisque basés sur la modification de sons eux-mêmes réels (enregistrement du champ de pression acoustique dû à la vibration de structures). La richesse du son n'impliquait pas pour autant qu'il soit toujours perçu comme naturel. Le revers de ce réalisme était alors une faible flexibilité, puisqu'à chaque méthode de traitement correspondait un petit nombre d'effets, et beaucoup de traitements n'étaient pas réalisables : ils restaient à découvrir ou à perfectionner. Petit à petit, les modèles de synthèse se sont affinés et

ont pris en compte une connaissance toujours accrue du signal sonore qu'ils modélisent (synthèse additive, synthèse soustractive, modèle physique, modèle hybride physique-signal) ou des systèmes complexes de synthèse (synthèse par modulation de fréquence, par distorsion non linéaire). La flexibilité des modèles utilisés dans la synthèse en temps réel avec contrôle gestuel a été conservée au prix d'une grande réflexion sur la manière de mettre en correspondance le geste musical et les paramètres de synthèse, et dans le même temps, le réalisme s'est accru.

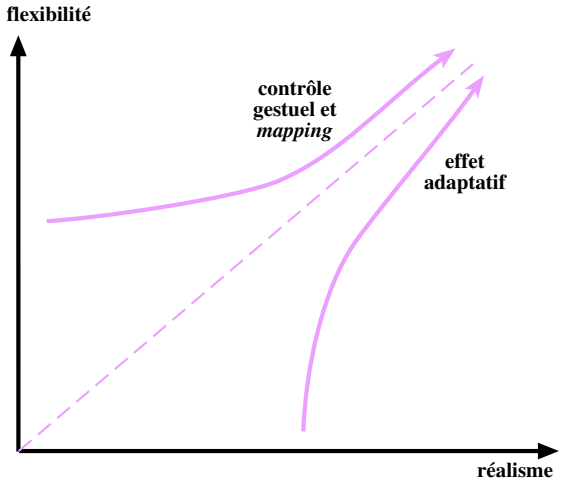


FIG. 2 – *Diagramme qualitative flexibilité/réalisme des effets et de la synthèse sonore.*

Les effets furent au départ basés pour la plupart sur du traitement en temps-réel, analogiques (mécanique, électromagnétique et électronique) et restaient simples à mettre en œuvre à partir d'un petit nombre d'opérateurs de base. Ainsi, le réalisme du son traité dépendait de la qualité de la méthode utilisée, du son à traiter et de l'adéquation entre le son et les hypothèses faites sur le son lors de la conception de la méthode. La flexibilité des effets était quant à elle très réduite : une méthode de traitement ne servait qu'à réaliser un petit nombre d'effets, et il fallait beaucoup de méthodes différentes pour couvrir les différents effets. Cela s'est structuré par la mise en œuvre de différents algorithmes dans une seule et même unité physique ou logicielle, appelée multi-effet et disponible notamment sous forme de *rack*, de pédalier (pour les guitaristes), d'ensemble logiciel. Aujourd'hui, le nombre de mécanismes a encore augmenté : plusieurs méthodes similaires, concurrentes ou complémentaires peuvent mettre en œuvre le même traitement, avec des résultats sonores, des artefacts et des limitations propres à chacune. La recherche dans ce domaine est toujours active, et les modèles d'Analyse–Transformation–Synthèse, en prenant en compte des informations sur le signal qu'ils modélisent, ont grandement ouvert les possibilités de traitement. Ceci a de plus augmenté la flexibilité des effets, en permettant d'utiliser des modèles communs à la synthèse sonore et aux effets.

Le contrôle gestuel de la synthèse sonore et des effets

Le contrôle gestuel des systèmes de synthèse en temps-réel a dès l'origine été pris en compte par les chercheurs (cf. le système temps-réel *Groove* décrit en 1970 par Mathews et Moore [Mathews and Moore, 1970]), et ceci afin de restituer aux instruments et modèles de synthèse un accès par le geste. C'est la condition *sine qua non* pour proposer un apprentissage de ces instruments, et par la suite une expressivité allant jusqu'à la virtuosité. Une réflexion soutenue a commencé récemment afin de proposer de bonnes manières de mettre en correspondance les paramètres décrivant le geste physique et instrumental avec les paramètres de contrôle de la synthèse : c'est la notion de *mapping* [Wanderley and Depalle, 1999]. Cette réflexion est nécessaire, du fait que les paramètres de synthèse sont de plus en plus nombreux au fur et à mesure que les modèles s'affinent et se complexifient.

Ce souci est celui de conserver la flexibilité tout en pouvant contrôler des modèles de plus en plus complexes et réalistes.

Le contrôle des effets s'est quant à lui principalement limité à la manipulation directe de ses paramètres de contrôle, rendant leur capacité d'évolution dépendante des capacités gestuelles humaines, en termes de rapidité de variation et de type d'évolution. Des efforts ont été faits pour combiner les paramètres de contrôle, afin de simplifier l'utilisation des effets. Ainsi, des interfaces graphiques et des interfaces matérielles donnent accès à des paramètres perceptifs, avec plus ou moins de finesse. On peut contrôler la plupart des effets avec un paramètre "sans/avec" (*dry/wet*), et une catégorisation psychoacoustique des paramètres de l'effet est proposée dans certains systèmes tel que le spatialisateur de l'IRCAM [Jullien *et al.*, 1993], la mise en correspondance entre ces critères perceptifs et les contrôles de l'effet étant partie intégrante du logiciel.

L'adaptation comme nouveau contrôle des effets

L'étude que nous proposons vise à augmenter la flexibilité des effets en proposant un nouveau contrôle, dit adaptatif, à la fois automatique (par le son) et gestuel. En tirant parti de connaissances sur le signal par le biais de descripteurs et de notions de psychoacoustique, nous allons appliquer un contrôle automatique au son, qui dépend donc du son lui-même. En utilisant des méthodes de *mapping* utilisées pour le contrôle gestuel de la synthèse sonore, nous allons généraliser le contrôle gestuel des effets audionumériques et ouvrir de nouvelles possibilités de transformations. Nous pourrons ainsi participer à l'évolution des effets dans leur trajectoire sur la figure 2, en les rapprochant des méthodes de synthèse en terme de flexibilité.

Le document s'articule donc en deux parties : une première partie bibliographique sur les effets audionumériques et les descripteurs sonores, et une seconde partie sur les effets audionumériques adaptatifs et leur contrôle gestuel. La première partie sur les effets audionumériques usuels s'articule en quatre chapitres introduits par une mise en contexte : des rappels sur les signaux sonores, leur mesure et leur perception, ainsi que des notions de vocabulaire et de taxonomie sur les effets (*cf.* chap. 1), une revue des méthodes de mise en œuvre (*cf.* chap. 2), une revue des effets existant, de leur contrôle et de leur mise en œuvre en temps réel et hors temps réel (*cf.* chap. 3), et la présentation de tout un ensemble de descripteurs sonores (*cf.* chap. 4). La seconde partie comporte trois chapitres, dans lesquels nous développons notre recherche sur les effets adaptatifs, introduits par une présentation des techniques et des outils à mettre en œuvre. Les effets adaptatifs ainsi que les techniques d'implémentation spécifiques que nous avons réalisés seront présentés et illustrés par des exemples sonores (*cf.* chap. 5). La structure de double contrôle, par le son et par le geste, sera décrite et justifiée, puis nous présenterons les interfaces graphiques réalisées (*cf.* chap. 6). Nous terminerons en décrivant les spécificités de mise en œuvre, en insistant sur l'intérêt que présente le contrôle adaptatif et en présentant des applications musicales réalisées (*cf.* chap. 7).

Première partie

Effets audionumériques usuels et description du signal sonore

Les effets audionumériques : des acteurs omniprésents dans la chaîne sonore

Je suis partagé entre mon goût pour les faits et mon goût pour l'effet.
Louis Scutenaire [Scutenaire, 1984]

Les effets audionumériques sont des traitements du signal numérique dédiés et appliqués à des signaux sonores. Leurs ancêtres analogiques sont utilisés depuis le début du XX^{e} siècle, dans les studios de radiophonie et d'enregistrement ainsi que dans dans les centres de création musicale. L'avènement des effets audionumériques et l'utilisation de l'ordinateur a permis de développer des matériels puis des logiciels que tout un chacun peut désormais utiliser dans des installations personnelles, dites *Home Studio*, depuis les années 1980.

On considère trois principaux contextes d'utilisation des effets audionumériques : la chaîne d'acquisition, la chaîne de diffusion et le studio.

Le premier contexte d'utilisation est dans la chaîne d'acquisition, lorsque le signal numérisé est modifié pour répondre à différents critères avant son stockage ou sa diffusion. Ces effets peuvent être une amplification (avant ou après la numérisation, d'ailleurs), une normalisation, une compression, etc. (*fig.* 3).

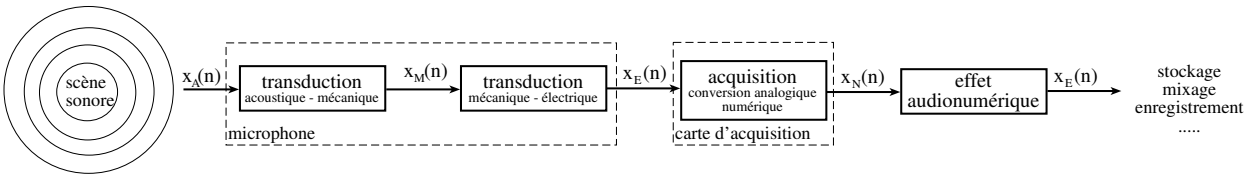


FIG. 3 – *Chaîne de traitement du signal acoustique jusqu'au stockage sous format numérique.*

Le deuxième contexte d'utilisation est la diffusion. L'utilisateur peut vouloir modifier le signal sonore pour répondre à des critères perceptifs et esthétiques : modifier l'équalisation spectrale se fait aisément sur la plupart des chaînes *Hi-Fi*, mais on peut aussi ajouter de l'écho ou de la réverbération, utiliser un filtre pour réduire le bruit de fond, modifier la panoramisation du signal, etc. (*fig.* 4).

Le troisième contexte d'utilisation des effets est le studio (*fig.* 5). Le studio permet d'utiliser les effets pour modeler le son, une fois encore selon des critères perceptifs et esthétiques. Plusieurs

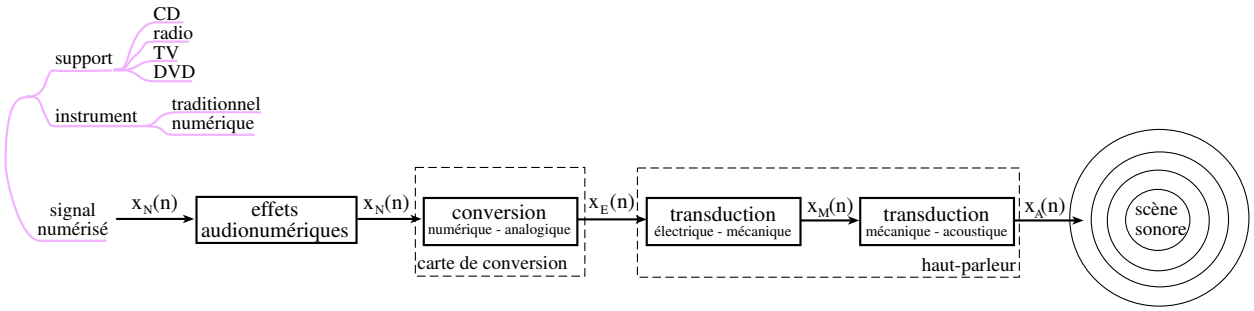


FIG. 4 – *Chaîne de diffusion du signal numérique au signal acoustique.*

démarches existent, allant du travail de production et de ***mastering*** visant à finaliser un enregistrement studio pour le rendre commercialisable (en terme de qualité sonore), jusqu'au travail de composition du son. Plusieurs styles de musique utilisent les effets audionumériques pour la composition, autant en musique populaire (musiques improvisées, musiques expérimentales, techno, électro) qu'en musique contemporaine (musique électroacoustique, musiques mixtes). C'est dans ce contexte que nous plaçons notre recherche, à savoir proposer de nouveaux effets et de nouveaux moyens de contrôles des effets dits "traditionnels" (même si la tradition n'est pas si vieille).

Dans ce contexte de travail en studio, de nombreuses solutions se sont développées : des solutions matérielles et logicielles coexistent ; plusieurs méthodes permettent d'obtenir des résultats similaires (mais pas forcément identiques) ; des logiciels commerciaux et des logiciels universitaires ont été développés et se destinent à des publics différents. Les centres de création musicale possèdent leurs propres systèmes, souvent développés par leurs soins. On le voit, la créativité ne s'arrête pas à la musique et à l'art, mais aussi aux outils que l'on développe !

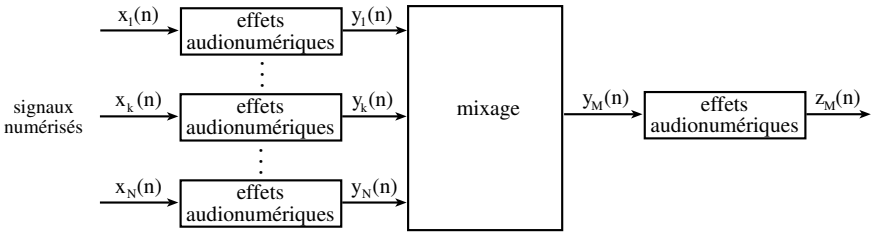


FIG. 5 – *Chaîne de traitement du signal numérique lors du mixage.*

Dans cette première partie, nous allons présenter les effets audionumériques usuels. La compréhension des effets audionumériques usuels nécessite une bonne connaissance d'un grand nombre de notions de traitement du signal numérique, mais aussi de perception auditive. C'est pour cela que nous allons d'abord effectuer des rappels sur les signaux sonores, leur mesure et leur perception. Nous préciserons alors le vocabulaire utilisé par la suite concernant les effets, et indiquerons différentes taxonomies permettant d'aborder les effets (*cf.* chap. 1). Nous poursuivrons par une revue détaillée des méthodes de mise en œuvre de traitement sonores, que ce soit dans le domaine temporelle, fréquentiel ou temps-fréquence (*cf.* chap. 2). Une fois toutes ces bases établies, nous effectuerons une revue des effets existant, en les ordonnant selon la taxonomie perceptive. Nous aborderons alors le contrôle de ces effets et les mises en œuvre spécifiques au temps-réel et au temps-différé (*cf.* chap. 3). Nous terminerons en présentant tout un ensemble de descripteurs sonores étudiés et utilisés pour automatiser les effets (*cf.* chap. 4) : ils proviennent de différentes disciplines, allant de la psychoacoustique et des modèles d'audition à la transcription automatique de partition en passant par la classification et la recherche de données ainsi que l'analyse-synthèse.

Chapitre 1

Rappels sur les signaux sonores, leur mesure, leur perception, les traitements appliqués

Se demander comment créer de l'inouï revient donc à se poser la question "comment accrocher la perception de l'auditeur", c'est-à-dire comment inventer des objets musicaux avec suffisamment de formes, au sens de gestalt, qui aient une grande force perceptive.
Jean-Luc Hervé [Hervé, 1999]

Sommaire

Nous proposons tout d'abord un chapitre donnant quelques bases et notions sur les différentes manières d'aborder un signal : par la mesure physique à l'aide de méthodes de traitement numérique de signal, et par la perception auditive (psychoacoustique). Ensuite, nous détaillerons le vocabulaire concernant les effets et la notion de temps-réel, puis terminerons par les taxonomies permettant de classifier les effets.

1.1 Représentations physiques et mesure d'un signal

Le signal que l'on reçoit dans une chaîne de traitement ou de diffusion peut se représenter dans le domaine temporel, dans le domaine fréquentiel ou dans le domaine temps-fréquence. De ces

différentes représentations, on peut extraire de l'information sous forme de descripteurs (*cf.* chap. 4). Cette information peut se limiter dans un premier temps à des mesures en lien avec la perception auditive, même si ce n'en sont pas des représentations exactes : l'énergie du signal, sa fréquence fondamentale si elle existe, ses formants s'ils existent, sa durée, et éventuellement sa répartition sur plusieurs canaux.

1.1.1 Représentations physiques d'un signal

Trois principales représentations du signal peuvent être considérées : la représentation temporelle, la représentation fréquentielle et la représentation temps-fréquence. Par simplicité, on considérera que les représentations temps-échelle, les représentations perceptives du cochléogramme ou des coefficients Mel-cepstraux sont des représentations temps-fréquences spéciales.

La *représentation temporelle* est la forme d'onde, courbe continue ou discrétisée du signal de pression acoustique ou du signal de tension électrique (responsable à terme d'un signal de pression acoustique en sortie d'un transducteur, tel un haut-parleur, un casque d'écoute). Un exemple est donné avec la forme d'onde de la voix de Pierre Schaeffer[1] en *fig.* 1.1. On remarque d'ores et déjà sur la forme d'onde que certains passages sont plus forts que d'autres. Cette forme d'onde compte 156844 échantillons, ce qui pour un signal échantillonné à $44,1\ kHz$ lui confère une durée de $3.55\ s$.

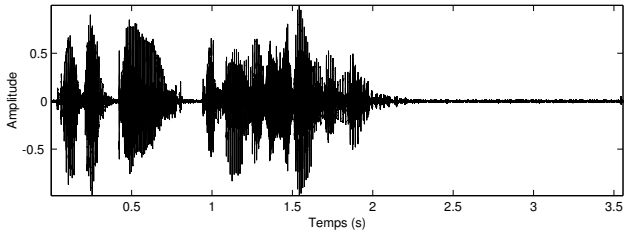


FIG. 1.1 – *Représentation temporelle d'un signal de parole, Piste n°16* ♫♪.

Ensuite, on peut représenter un signal périodique de durée infinie comme une somme de sinusoïdes : c'est la *représentation fréquentielle* par transformée de Fourier. Cette décomposition repose sur l'analyse de Fourier (1768-1830). Ainsi, le signal $s(t)$ peut se décomposer selon :

$$s(t) = \sum_{n=-\infty}^{\infty} c_n \exp(2i\pi nt/T) \tag{1.1}$$

avec T la période du signal et c_n les coefficients de Fourier définis par :

$$c_n = \frac{1}{T}\int_0^T s(t)\exp(-2i\pi nt/T)dt \tag{1.2}$$

On peut aussi calculer la transformée de Fourier discrète (approximation discrète de la formulation intégrale donnée en *eq.* (1.1)) sur un signal de durée finie, en supposant qu'il est infini périodique, de période égale à sa durée (*cf. fig.* 1.2).

Sur cette figure, toute l'information fréquentielle du signal est représentée, ce qui ne rend pas la lecture de la transformée de Fourier très facile, puisqu'on sait bien quelles fréquences sont présentes, mais pas quand elles apparaissent et disparaissent. Cette représentation n'est donc adaptée qu'à des sons stationnaires, ce qui est l'hypothèse sous-jacente de la transformée de Fourier. Pour y remédier, des *représentations temps-fréquence* ont été développées. Il s'agit de représenter le contenu

[1] Pierre Schaeffer est le père fondateur de la musique électroacoustique en France à la fin des années 40 au club d'essai de la radio française. Il a fondé dans les années 50 le Groupe de Recherches Musicales (GRM), qui continue aujourd'hui son activité au sein de l'INA, Institut National de l'Audiovisuel.

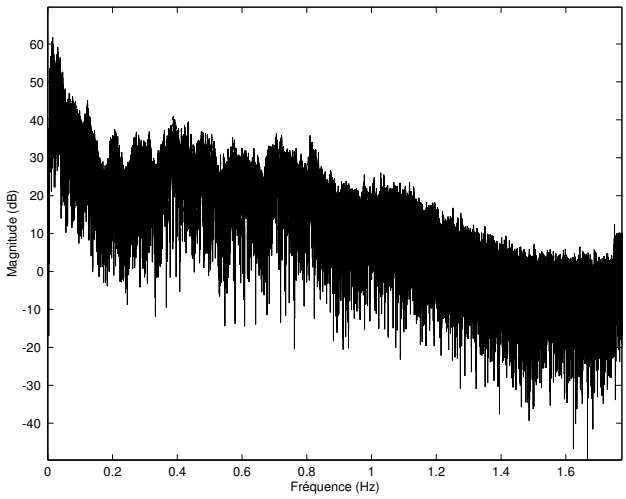


FIG. 1.2 – *Représentation fréquentielle d'un signal de parole, Piste n°16 ♫♪.*

en fréquence d'une petite portion de signal (appelée spectre à court-terme, ou TFCT pour transformée de Fourier à court-terme) considéré comme localement stationnaire, et de déplacer la fenêtre d'analyse tout au long du son. On obtient alors une image temps-fréquence (*cf. fig.* 1.3). Cette figure est bien plus lisible : on voit apparaître des structures du signal, que nous allons présenter en *sec.* 1.1.3. L'échelle d'amplitude utilisée est le décibel de puissance, noté dB et obtenu par la formule :

$$A_{dB} = 20 \log_{10} A_{lin} \tag{1.3}$$

La représentation temps-fréquence classique est appelée **sonagramme**, ou encore spectrogramme

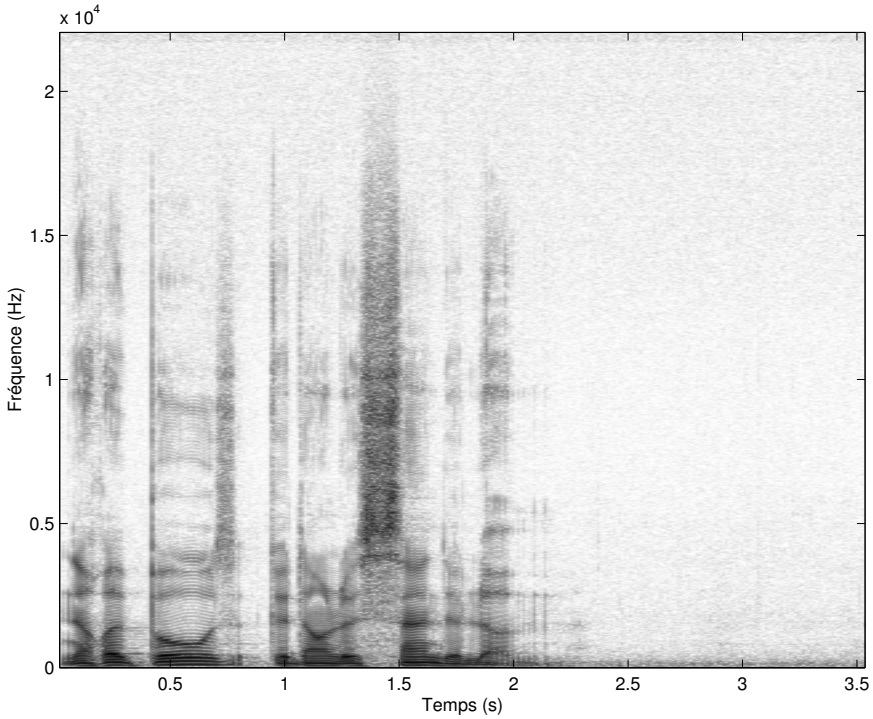


FIG. 1.3 – *Représentation temps-fréquence d'un signal de parole, Piste n°16 ♫♪.*

de magnitude, la magnitude correspondant à l'amplitude mesurée en décibels.

1.1.2 Caractérisations des sons et de leurs composantes

Sinusoïde, partiel, bruit

Une sinusoïde est une fonction périodique et invariable dans le temps. On appelle **partiel** une sinusoïde modulée en amplitude et/ou en fréquence dans le temps, que l'on reconnaît dans le spectre à court-terme d'un signal par une raie spectrale ou un pic. Les **bruits** peuvent être des signaux aléatoires filtrés ou des transitoires d'attaque, des discontinuités de l'onde, ou une combinaison de ces trois possibilités. Les pics qu'ils peuvent présenter dans des paniers de fréquences d'une TFCT ont ceci de différent avec les partiels que :

- leur amplitude est souvent plus petite ;
- les maxima locaux voisins sont plus nombreux et de plus grande amplitude ;
- d'une transformée à court-terme à la suivante (dans le temps), ils apparaissent et disparaissent, alors que les partiels demeurent présent et évoluent plus doucement (en fréquence et en amplitude).

Sons harmoniques

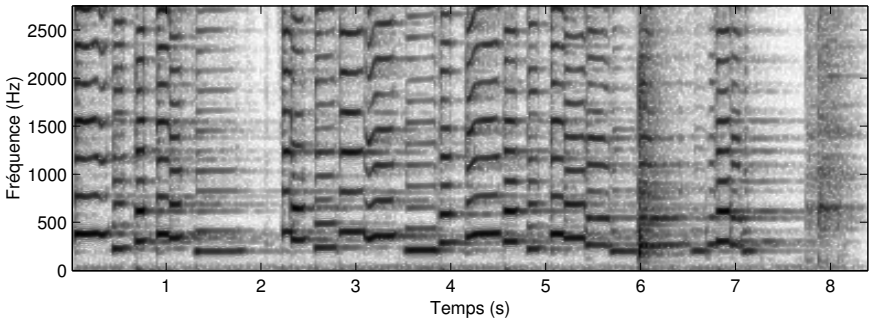


FIG. 1.4 – *Sonagramme d'un son instrumental (saxophone de Sylvain Beuf), Piste n°3* ♫♪.

Les sons harmoniques sont constitués d'un ensemble de partiels tous multiples d'une fréquence fondamentale (*cf. fig.* 1.4). C'est le cas de la plupart des sons instrumentaux (voix parlée ou chantée, instruments à vent ou à cordes, cuivres). Les sons bruités, quant à eux, correspondent à des sons sans hauteur prédominante, ou avec plusieurs hauteurs audibles. Leur spectre ne présente pas du tout la même allure (*cf. fig.* 1.5).

Entre ces deux extrêmes se situent :

- les sons quasi-harmoniques : les fréquences des partiels dévient petit à petit du multiple de la fréquence fondamentale, au fur et à mesure que l'on monte en fréquence et en numéro de partiel. C'est le cas de sons instrumentaux tels que ceux du piano ou certains instruments de percussion (ex : marimba, xylophone).
- les sons inharmoniques : les fréquences des partiels ne sont pas multiples d'une fréquence fondamentale ; cependant, on peut entendre une hauteur, voire plusieurs selon l'émergence de chaque partiel (ex : les sons de cloches, *cf. fig.* 1.6).
- les bruits filtrés présentant un phénomène d'émergence : il s'agit de bruits filtrés passe-bande étroit. Selon la structure du ou des filtres, une hauteur (obtenue par exemple avec un filtre résonant) peut apparaître à leur écoute .

On remarque que les sons harmoniques, quasi-harmoniques et inharmoniques sont constitués de partiels mais aussi de bruit, composé d'un éventuel transitoire d'attaque et d'un bruit résiduel.

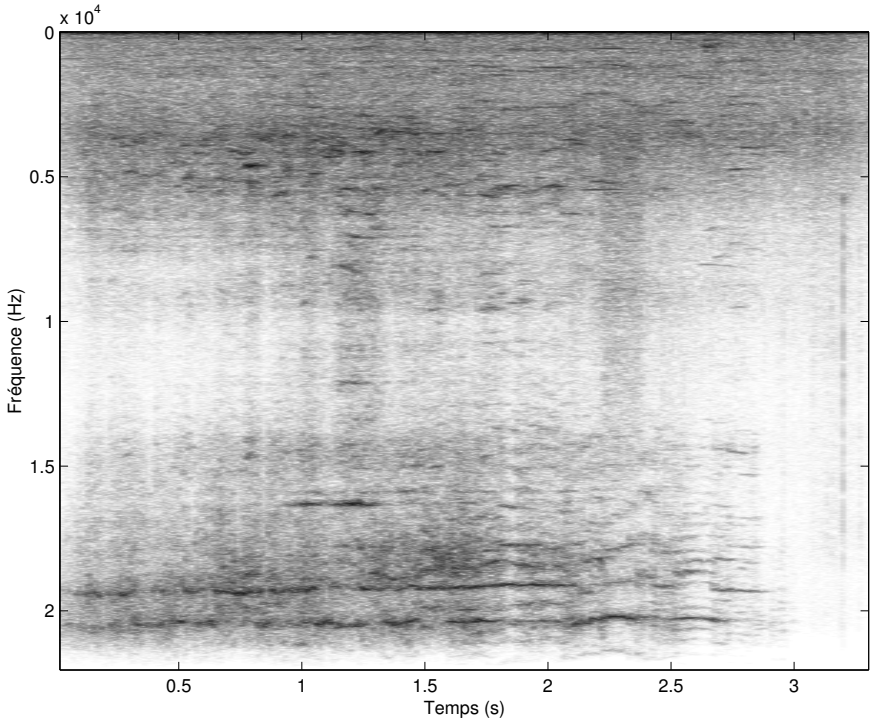


FIG. 1.5 – *Sonagramme d'un bruit de freins de bus, Piste n°56-CD2* ♫♪.

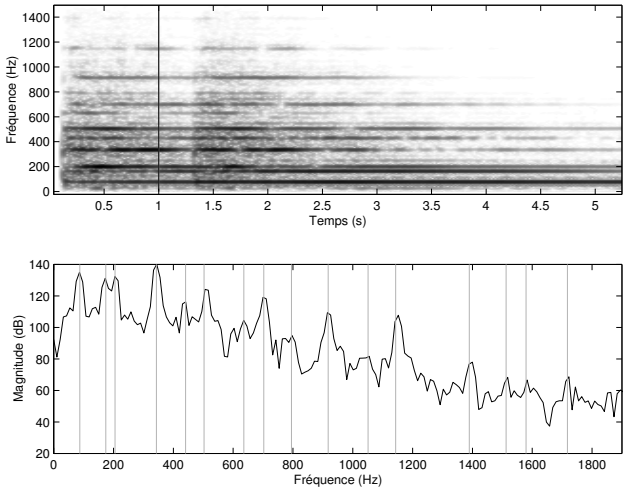


FIG. 1.6 – *Spectrogramme (*fig. *bas) d'un son de cloche, Piste n°2* ♫♪ *et sa TCFT (*fig. *bas) prélevée au niveau de la ligne verticale noire, à* 1 s*. Les traits gris verticaux sur la TFCT indiquent la fréquence de chaque partiel.*

Sons purs, monophoniques, polyphoniques

Un son pur est composé d'une seule sinusoïde. Musicalement, un tel son est très pauvre. Dès qu'on lui ajoute des harmoniques, il sonne beaucoup mieux, du fait de la plus grande richesse du spectre. Ceci dit, ce son est encore monophonique. Les sons musicaux occidentaux sont quantifiés sur une échelle dite tempérée. Une succession de sons monophoniques sur cette échelle constitue une mélodie, parfois une gamme. La gamme est un sous-ensemble de notes, une structure harmonique (au sens de l'harmonie en musique) constituant le passage musical. Si on joue simultanément

plusieurs sons monophoniques, on obtient des accords (sons polyphoniques), dont les noms sont donnés par les écarts respectifs entre les fréquences fondamentales des sons monophoniques le constituant. Il existe un lien fort entre un accord et une gamme, les deux respectant l'harmonie musicale. Bien qu'il faille un cours de solfège et d'harmonie pour maîtriser parfaitement ces notions, nous donnons quelques exemples de noms de notes et d'intervalles dans le tableau *Tab.* 1.1. Pour plus d'informations, tout traité de solfège y répond ; nous conseillons particulièrement l'ouvrage classique [Abromont and de Montalembert, 2001] ainsi que l'ouvrage [Siron, 1992] dont l'optique jazz et musiques improvisé donne une approche de l'harmonie très agréable, complète et sous plusieurs axes.

Nom	**rapport γ**	**Intervalle**	**Notation**
Do (C)	1	unisson	unisson
Do ♯ ou Ré b	$2^{1/12}$	seconde mineure	$2^{\text{nde}}m$
Ré (D)	$2^{2/12}$	seconde majeur	$2^{\text{nde}}M$
Ré ♯ ou Mi b	$2^{3/12}$	tierce mineure	$3^{\text{ce}}m$
Mi (E)	$2^{4/12}$	tierce majeure	$3^{\text{ce}}M$
Fa (F)	$2^{5/12}$	quarte juste	$4^{\text{te}}J$
Fa ♯ ou Sol b	$2^{6/12}$	quarte augmentée / quinte diminuée	$4^{\text{te}}A$ ou 5^{te} d
Sol (G)	$2^{7/12}$	quinte juste	$5^{\text{te}}J$
Sol ♯ ou La b	$2^{8/12}$	sixième mineure	$6^{\text{xte}}m$
La (A)	$2^{9/12}$	sixième majeure	$6^{\text{xte}}M$
La ♯ ou Si b	$2^{10/12}$	septième mineure	$7^{\text{ème}}m$
Si (B)	$2^{11/12}$	septième majeure	$7^{\text{ème}}M$
Do (C)	$2^{12/12} = 2$	octave	8^{ve}

TAB. 1.1 – *Noms des notes et rapport des fréquences γ, noms et notations des intervalles.*

1.1.3 Analyse physique des sons

Amplitude du signal

L'amplitude d'un signal peut se mesurer en terme d'énergie ou de puissance. Ces deux mesures peuvent se mesurer avec la TFCT, du fait de l'égalité de l'énergie d'un signal à court-terme en temps et en fréquence (identité de Parseval). L'amplitude en terme de puissance peut se calculer échantillons par échantillons sur la représentation temporelle.

Durée du signal

La durée τ d'un signal peut se mesurer en nombre d'échantillons N_e, ou en secondes, par la formule :

$$\tau = \frac{N_e}{F_e} \tag{1.4}$$

avec F_e la fréquence d'échantillonnage du signal, c'est-à-dire le nombre d'échantillons par seconde.

Spectre fréquentiel du signal

Comme nous l'avons vu grâce à la présentation des différents sons, le spectre fréquentiel d'un signal présente deux niveaux caractéristiques : la présence de partiels ou d'harmoniques, et la forme de l'enveloppe spectrale.

La fréquence fondamentale

Lorsqu'un son est parfaitement périodique, il peut se décomposer parfaitement comme la somme de sinusoïdes de fréquences $f_n = nf_0$, avec $f_0 = f_1$ la notation de la fréquence dont les harmoniques sont des multiples. Cette fréquence f_0 est appelée fréquence fondamentale, n est le numéro d'harmonique. Pour un son quasi-harmonique, on peut chercher une pseudo-période fondamentale comme étant la fréquence dont les rapports $\frac{f_n}{n}$ (fréquence du partiel sur son numéro d'harmonique) sont les plus proches. Elle est proche de la première harmonique, et lui est confondue avec pour les sons parfaitement harmoniques. Par contre, pour des bruits, des transitoires, cette notion n'a plus vraiment de sens.

Enveloppe spectrale et formants

Si l'on continue d'observer le spectre d'amplitude de signaux musicaux ou vocaux, cette fois-ci en regardant leur évolution dans le temps, nous remarquons que des formes plus générales que les partiels apparaissent. Il s'agit de l'enveloppe spectrale, définie comme l'enveloppe supérieure (une forme reliant les pics ou partiels), souvent approchée par un lissage du spectre (*cf.* sec. 2.4.4). Cette enveloppe est formée de bosses appelées formants, et de creux. Ces formants proviennent du mode d'émission du signal : d'après le modèle source-filtre (analyse soustractive) utilisé pour les instruments à vent par exemple, le signal est une source riche en harmoniques et en bruit, qui passe au travers d'un système résonant (la conduit vocal, par exemple), dans lequel certaines fréquences sont atténuées et d'autres amplifiées. Les fréquences sont amplifiées autour des fréquences de résonance du système que le signal traverse. C'est notamment grâce aux formants qu'on peut reconnaître les voyelles entre elles. La figure 1.7 montre bien la coexistence de ces deux niveaux de structure que sont les formants (*fig.* du haut) et les partiels (*fig.* du bas).

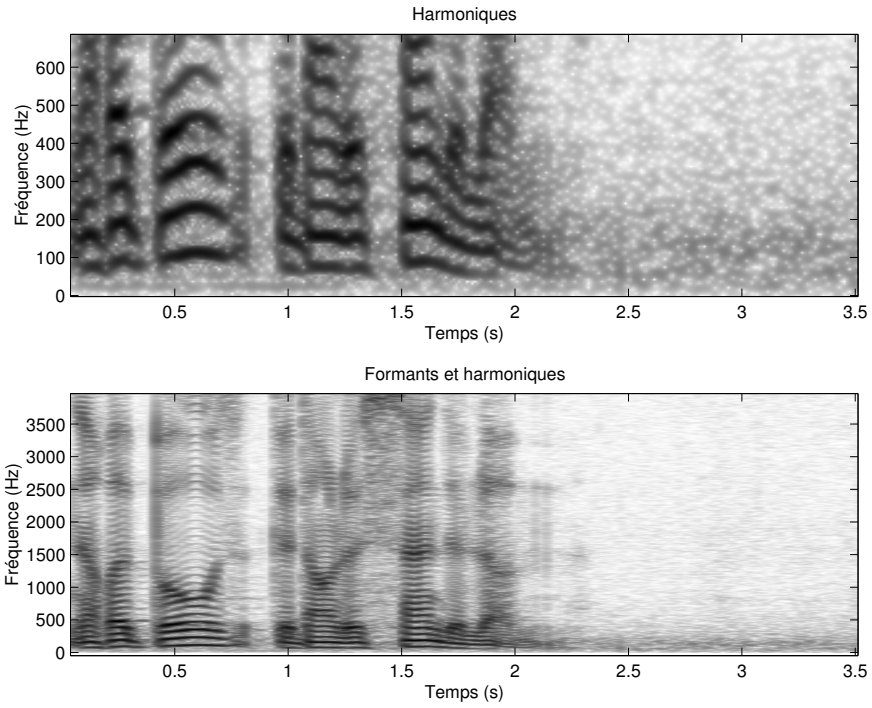


FIG. 1.7 – *Zoom sur le sonagramme d'une voix parlée : fréquences de 0 à* 700 *Hz pour visualiser les harmoniques, et de 0 à* 4000 *Hz pour visualiser les formants , pour le son Piste n°16* ♫♪.

Signal multi-canal

Un signal multicanal peut transporter une information supplémentaire au signal mono-canal. En effet, à partir de deux canaux de diffusion, on peut donner l'illusion de sources sonores de provenances différentes, selon comment les sons ont été mixés et diffusés dans le système. Les différences de niveau entre les canaux ainsi que les retards entre les canaux permettent d'indiquer assez bien des positions différentes, en dehors des haut-parleurs de diffusion. Nous y reviendrons dans la présentation des effets de spatialisation (*cf.* sec. 3.5).

1.2 Perception auditive

Maintenant que nous avons rappelé ces bases sur les représentations de signaux sonores ainsi que sur les constituants de ces signaux, nous allons parler des attributs de la perception auditive. Ceci est très important dans notre approche, puisque l'on veut se placer du côté du musicien qui utilisera les effets adaptatifs. Il nous faut donc connaître un minimum de notions sur la perception auditive, à la fois afin de connaître les paramètres qui la décrivent et leurs moyens de calculs (*cf.* chap. 4), et afin de classer les effets audionumériques selon des critères perceptifs ou musicaux (*cf.* sec. 1.5).

1.2.1 Audibilité des sons

Pour qu'un signal soit sonore et donc audible, il faut qu'il respecte plusieurs critères. Le premier est la composition fréquentielle de ce signal : seules les fréquences entre 20 Hz et 20 000 Hz peuvent être entendues par l'oreille humaine. On parlera alors de zone de fréquences audio (les fréquences inférieures sont appelées sub-audio ou infra-sons, les fréquences supérieures sont les ultra-sons). Cet intervalle de fréquences est normatif, et varie légèrement d'un individu à l'autre. Le deuxième critère est la puissance du signal : si le signal est trop faible, il ne peut être entendu. L'oreille humaine a un seuil d'audibilité qui varie d'une fréquence à l'autre. Ce seuil en fonction de la fréquence varie d'un individu à l'autre. Il varie de -4 dB à 4 kHz à quelques dizaines de dB en très basses et en très hautes fréquences. Un traumatisme auditif ou une fatigue de l'oreille feront qu'un individu perd un peut d'audibilité pour certaines fréquences. D'autres traumatismes au niveau du cortex peuvent aussi gêner la bonne compréhension d'un signal pourtant détecté par l'oreille interne.

1.2.2 Hauteur tonale

La **hauteur tonale** ou hauteur perçue est le "caractère de la sensation auditive lié à la fréquence d'un son périodique, qui fait dire que le son est aigu ou grave selon que cette fréquence est plus ou moins élevée" [AFNOR, 1977]. Cette définition est incomplète, puisqu'elle ne tient pas compte de la proximité de deux sons musicaux séparés d'un intervalle d'une octave (de fréquences f_0 et $2f_0$) qui est plus grande que deux sons séparés d'un ton (de fréquences f_0 et $\sqrt[12]{2}f_0$). De plus, les sons non périodiques peuvent aussi avoir une hauteur tonale. Ce paramètre de la perception auditive est très important, car il permet de percevoir les mélodies, les harmonies.

Une échelle de représentation, l'échelle des $mels$, a été développée pour les sons purs . La référence arbitraire est de 1000 $mels$ pour un son pur de fréquence 1000 Hz. Un son deux fois plus haut aura comme hauteur tonale 2000 $mels$, etc. La hauteur tonale exprimée en $mels$ n'est ni proportionnelle aux fréquences, ni au logarithme des fréquences. L'échelle des $barks$ est par contre corrélée avec l'échelle des $mels$, puisque 1 $bark = 100$ mel. En pratique, l'échelle des $mels$ est trop éloignée de l'échelle musicale occidentale, si bien qu'elle n'est pas employée.

L'échelle musicale occidentale (dite "tempérée" présentée *Tab.* 1.1) est proportionnelle au logarithme des fréquences. On remarque la redondance des noms de notes à chaque nouvelle octave : les noms des notes correspondent aux **chromas**, et la distance entre les chromas n'est pas directement corrélée à la hauteur. Ainsi, un écart d'octave semble plus proche qu'une tierce, de même qu'un

écart de quinte, et ceci pas uniquement pour une oreille de musicien entraîné. Cette dimension chromatique est différente de la dimension grave-aigu.

Notons que la hauteur tonale d'une sinusoïde pure varie avec le niveau[2] : elle diminue lorsque le niveau sonore augmente pour les fréquences inférieures à 1000 Hz, et elle augmente lorsque le niveau augmente pour les fréquences supérieures à 2000 Hz. Les variations de hauteur tonale ne sont pas monotones pour les auditeurs : il existe des différences interindividuelles. De plus, la hauteur tonale n'est pas forcément identique selon qu'on écoute de l'oreille gauche ou droite (diplacousie binaurale). Enfin, le masquage partiel d'un son pur par un bruit augmente la hauteur tonale du son pur. Un effet de masquage temporel fait que la hauteur d'un son pur est modifiée de manière répulsive par la présence d'un autre son pur le précédant dans le temps : si le second son est plus haut que le premier, il est entendu légèrement plus haut qu'il n'est réellement. De même, si le second son est plus bas que le premier, il est entendu légèrement plus bas qu'il n'est.

Concernant les sons complexes harmoniques, la hauteur peut être définie à partir de la fréquence fondamentale (hauteur tonale) ou en considérant tout le spectre (hauteur brute) : elle a deux dimensions. La première dimension, qui se base sur la fréquence fondamentale, est parfois appelée hauteur tonale pour les sons harmoniques complexes (et *periodicity pitch, residue pitch, low pitch, complexe pitch, virtual pitch* en anglais). Concernant la deuxième dimension, la **hauteur brute**, elle correspond à la dimension grave-aigu. Elle découle du lien fort existant entre la composition du spectre et le timbre du son. En effet, un son de spectre plus riche sera perçu comme plus brillant, et plus aigu. Le résultat de cet effet dû à la composition du spectre est l'erreur d'octave : on peut facilement confondre deux sons séparés d'une octave sur l'échelle musicale.

L'écoute des sons multiphoniques peut se faire de deux manières : soit on se focalise sur un seul instrument (écoute analytique), soit on écoute l'ensemble des sons (écoute synthétique). La capacité de séparer les notes simultanées fait appel à l'organisation tonotopique du système auditif, qui attribue un emplacement de l'activité mécanique ou électrique dans la cochlée à chaque fréquence. On parle alors de **hauteur spectrale** pour parler de la hauteur tonale de chaque son pur d'un son complexe harmonique.

Notons enfin l'énigme de la **fondamentale absente** : un son dont la fréquence fondamentale est supérieure à 60 ou $70Hz$, ayant 2 ou 3 composantes spectrales de fréquences inférieures à 5000 Hz et dont les premières harmoniques sont de numéro n inférieure à 20, sera entendu à la fréquence fondamentale, même s'il manque plusieurs des premières harmoniques. Plus il possède d'harmoniques, plus la hauteur est saillante. Les tentatives d'explication de ce phénomène sont à l'origine des nombreux modèles de perception de la hauteur.

Pour terminer, il faut savoir que les pseudo-bruits périodiques ou les bruits filtrés possèdent une hauteur, soit lorsque la période est inférieure à 15 ms (65 Hz), soit lorsque le filtre possède un ou plusieurs formants.

1.2.3 Sonie ou intensité sonore

La **sonie** est la grandeur subjective de l'intensité sonore. Elle se mesure en *sone* : un *sone* est la sonie d'un son pur de 1 kHz à 40 dB, en incidence frontale, d'une durée de 1 s. Doubler la sonie correspond à doubler l'intensité sonore perçue. Le silence approche les 0 *sone*, tandis qu'un son pur 1 kHz à un niveau de pression SPL de 40 dB présenté dans le plan facial dans un champ libre a une sonie de 1 *sone*.

Elle correspond à un codage effectué par le système auditif, prenant en compte des informations temporelles (enveloppe temporelle, ie. la durée et le niveau acoustique) et fréquentielles (énergie dans des sous-bandes du spectre d'amplitude). Elle dépend aussi du champ acoustique : libre, ou plus ou moins diffus. Le système auditif effectue un filtrage assez complexe, comme on a déjà pu s'en rendre compte concernant la hauteur tonale. On utilise fréquemment une représentation du spectre par bandes critiques appelée échelle de Bark : des bandes de largeur constante de 100 Hz

[2] Pour s'en convaincre, on pourra écouter un signal test ou musical sur un casque : la hauteur varie lorsqu'on approche le casque des oreilles.

en dessous de 500 Hz, et des bandes de largeur proportionnelle à leur fréquence centrale au dessus de 500 Hz. Les fréquences sont atténuées ou amplifiées lors de la transmission de l'oreille externe à l'oreille moyenne : on appelle facteur de transmission a_0 cette transmission. Ceci signifie que pour bien modéliser la sonie, il faut tenir compte de ces courbes de transmission, différentes selon le champ acoustique de diffusion. Le masquage fréquentiel intervient alors, car pendant l'excitation, la sélectivité de la cochlée n'est pas rectangulaire, contrairement aux filtres. Il faut donc modéliser ce masquage, dont la courbe n'est pas symétrique et dépend du niveau.

Le calcul de la sonie passe par la détermination de l'excitation dans chaque bande de fréquence, et du bruit de bande critique (les effets du masquage), ce qui donne la sonie spécifique dans chacune des 24 sous-bandes constitué du maximum de niveau dans le filtre (sonie de cœur) et de sonies de flanc (effets du masquage). La sonie totale se calcule par intégration de la sonie spécifique sur l'échelle des *barks*.

Le niveau de sonie, ou isosonie, correspond au niveau de pression SPL d'un son pur à 1 kHz ayant la même sonie que le son mesuré. L'isosonie L_N est mesurée en *phone*. Il se calcule à partir de la sonie totale (ceci sera développé dans le chapitre 4 sur les descripteurs), et la sonie globale d'un son variant dans le temps est la valeur dépassée 5 à 10 % du temps. Dans certains modèles, l'ERB (*Equivalent Rectangular Bandwidth*) remplace l'échelle des *barks* [Moore and Glasberg, 1996].

1.2.4 Spatialisation

La dimension spatiale d'un son a plusieurs caractéristiques : le positionnement de la source (*cf. fig.* 1.8), son rayonnement et les effets de salle (écho et réverbération).

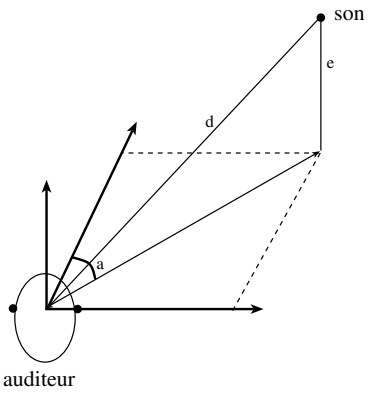


FIG. 1.8 – *Localisation 3D d'un son : azimut (a), distance (d) et élévation (e).*

Localisation

La **localisation** d'un son dans l'espace se fait en azimut, en élévation et en distance. La localisation en azimut se fait principalement à partir des différences interaurales : différence de temps d'arrivée du signal acoustique aux deux oreilles et des différences de niveau de ces signaux, tandis que la localisation en élévation se fait à partir du filtrage procédé par le torse, le cou et la tête :

- la différence d'intensité interaurale (*IID, Interaural Intensity Difference*), utilisée par le système auditif pour les fréquences supérieures à 1500 Hz ;
- la différence de temps d'arrivée interaurale (*ITD, Interaural Time Difference*), pour les fréquences inférieures à 1500 Hz ;
- Les fonctions de transfert relatives à la tête (*HRTF*), qui permettent de lever l'ambiguïté du cône de confusion entre l'avant et l'arrière, lorsque les ITD et IID sont identiques.

La distance quant à elle est donnée en espace clos par la réverbération (son niveau par rapport au niveau du son direct), et en espace ouvert par le niveau du signal et son filtrage effectué par l'air).

La distance d'une source sonore est évaluée par l'auditeur à l'aide de plusieurs mécanismes : tout d'abord le niveau du signal émis, ensuite le filtrage passe-bas du son (effectué par l'air), puis l'écho et la réverbération (le rapport entre la puissance du signal clair sur la puissance de la réverbération du signal), et enfin des indices cognitifs (c'est notamment grâce à eux que l'on reconnaît une voix chuchotée d'une voix parlée ou criée, enregistrées en chambre source et égalisées en sonie).

Réverbération, rayonnement

La réverbération d'un son est la somme des réflexions du son original sur les parois d'un espace clos. Il s'agit d'un filtrage, d'une coloration donnée par un lieu sur les sons. C'est un indice perceptif fort, qui nous permet de reconnaître le type de lieu (autant la taille que les types de surfaces, réfléchissantes, diffusantes ou absorbantes) qui recouvrent les parois. La réverbération se décompose en trois constituants : le son direct, les réflexions primaires, correspondant à des échos très courts, en nombre fini, et les réflexions secondaires, correspondant aux réflexions plus tardives, et de ce fait davantage filtrées par la salle.

Les attributs perceptifs d'une réverbération permettant de caractériser le qualité acoustique d'une salle se répartissent en 3 catégories [Jot and Warusfel, 1995] :

- ◇ la perception de la source (en lien avec le spectre et l'énergie relative du son direct et des premières réflexions) :
  - la présence (rapport entre énergie du son direct et celle des réflexions primaires) ;
  - la brillance (variations hautes fréquences des réflexions primaires) ;
  - la chaleur (variations basses fréquences des réflexions primaires) ;
- ◇ l'intéraction source – salle (en lien avec les énergies relatives du son direct, des réflexions primaires et secondaires, et avec les temps de décroissance primaires) :
  - l'enveloppement (énergie des réflexions primaires par rapport au son direct) ;
  - la présence de la salle (énergie du son réverbéré tardif) ;
  - réverbération traînante (temps de décroissance primaire de la réponse impulsionnelle de la salle) ;
- ◇ la perception de la salle (en lien avec les temps de décroissance secondaires et leurs variations en fréquence) :
  - la réverbération tardive (temps de décroissance secondaire de la réponse impulsionnelle de la salle) ;
  - la lourdeur ou *heaviness* (variations basses fréquences du temps de décroissance) ;
  - la vivacité ou *liveness* (variations hautes fréquences du temps de décroissance).

La directivité du rayonnement de sources sonores est propre à chaque type de source. Ainsi, une source ne distribue pas toutes ses fréquences de la même manière dans les différentes directions.

1.2.5 Timbre

Le timbre est défini par la norme ANSI [ANSI, 1960] comme : "*Timbre is that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar*". Ainsi, on différencie deux instruments de musique jouant la même note à la même intensité parce qu'ils ont un timbre différent. Il manque cependant à cette définition la dimension temporelle, qui joue un rôle important. Le timbre est multidimensionnel, et selon si l'on se place du point de vue acoustique, perceptif ou sémantique, on peut trouver différentes dimensions.

Le spectre est l'accès le plus immédiat au timbre : son enveloppe et ses formants sont spécifique à l'instrument, au locuteur. Ensuite, la dimension temporelle et l'amplitude du signal dans cette dimension joue un rôle important : c'est ainsi que les attaques des notes instrumentales permettent elles aussi l'identification du timbre. Cette dimension temporelle se justifie encore plus quand on remarque qu'un son modifié par un filtrage (la réponse du lieu de diffusion ou le filtrage lors de la transmission téléphonique, par exemple) reste très reconnaissable. Ce qui est vrai pour l'attaque

l'est aussi pour la décroissance. Enfin, les propriétés spectrales fines (les micro-variations des harmoniques, par exemple) jouent un rôle et sont en quelque sorte la troisième dimension identifiable du timbre. Pour résumer, le **timbre** est codé principalement (pour les sons instrumentaux) par les trois dimensions suivantes :

1. la brillance, caractéristique de la distribution de l'énergie spectrale, et corrélée au centroïde (ou centre de gravité spectrale) [Grey, 1975; McAdams and Cunibile, 1992; Iverson, 1995] ;
2. les propriétés spectrales fines, à savoir la cohérence des micro-variations des composantes spectrales, mesurées par le flux spectral et le synchronisme des attaques et décroissances des harmoniques [McAdams and Cunibile, 1992] (la synchronie en question est statique pour les instruments à vents et dynamique pour les cuivres) ;
3. le transitoire d'attaque qui confère le caractère d'explosivité ou non de l'attaque d'un son ; il est décrit par la synchronie et l'harmonicité des partiels dans différentes parties du spectre.

Sur les partitions, le timbre correspond aux instruments (précisés en début de portée) et à certains modes de jeux (indiqués au fur et à mesure de leur apparition dans la partition). Le vibrato et la consonance peuvent aussi être considérés comme des attributs du timbre des sons harmoniques (vibrato, consonance) ou non (dissonance).

1.2.6 Rugosité, consonance-dissonance

La rugosité est cet attribut des sons contenant des battements si rapides qu'ils ne sont plus perçus comme des battements mais comme une granulation rapide du son. La rugosité est soit due à des fréquences proches et dont la somme s'entend comme un produit, soit due à une modulation d'amplitude appliquée à un son harmonique (voir la modulation d'amplitude, *cf.* sec. 2.3.1). Ces deux représentations sont duales, l'une dans le domaine fréquentiel, l'autre dans le domaine temporel. On appelle encore dissonance la rugosité pour des sons simples [Pressnitzer, 1998] ; l'absence de rugosité est appelée consonance en musique. La rugosité peut être considérée comme une dimension du timbre.

1.2.7 Vibrato, trémolo

Le vibrato correspond à une modulation de fréquence d'un son harmonique, perçu quasi indépendamment de la hauteur tonale, comme un attribut à part entière. Cette modulation est au maximum de l'ordre d'un demi-ton (75% d'un demi-ton, quelques % pour les instruments traditionnels), et est produite à une fréquence variant entre 4 et 8 Hz. Elles est produite notamment par les chanteurs et chanteuses lyriques en modulant le larynx, par les instruments à vents en modulant la pression dans l'embouchure, par les instruments à cordes (tels que la guitare et le violon, mais pas le piano !) en modulant la pression et la position du doigt sur la corde. Il fait partie des modes de jeu de l'instrumentiste. Le fait que la hauteur varie peu fait que les harmoniques balayent le spectre mais en conservant la forme de l'enveloppe spectrale : l'enveloppe est plus clairement dessinée, et la qualité du son rendu plus appréciable. En effet, l'oreille fonctionne de façon différentielle, et il a été montré que des sons de synthèse dont les fréquences des harmoniques ne varient absolument pas sonne très terne, comparés aux sons dont les harmoniques subissent des micro-modulations de fréquence [Risset and Wessel, 1999].

Le trémolo quant à lui est le dual du vibrato dans le domaine temporel : il s'agit d'une modulation d'amplitude de fréquences de quelques Hertz, accompagnant souvent le vibrato (en effet, il est par exemple impossible de moduler la pression dans l'embouchure d'un instrument à vent pour faire un vibrato sans modifier l'intensité du signal sonore).

1.2.8 Relations temps-fréquence

Une onde sinusoïdale correspond à un phénomène temps et fréquence par essence. La modification de l'amplitude d'un signal par modulation d'amplitude (multiplication d'un signal par une sinusoïdale de fréquence f_{mod}) produit un changement de dynamique (trémolo) pour les fréquences sub-audio. Les fréquences sub-audio sont en effet du domaine temporel, du domaine du rythme. Si l'on ajoute à un signal une version retardée du même signal, la modification de la perception se fera sur un paramètre différent selon le temps de délai. Ainsi, un délai très court (quelques ms) correspond à un effet de filtrage, sur le timbre. Un délai plus long (200 ms) est entendu comme une modification de l'attaque et donc du timbre, et un délai encore plus long est entendu comme un écho (paramètre décrivant l'espace sonore). On voit donc par ces deux exemples simples que les relations entre temps et fréquences sont complexes.

1.2.9 Illusions sonores

Nous avons expliqué que certains descripteurs perceptif du son ont plusieurs dimensions : c'est le cas notamment de la hauteur et du timbre. [Risset, 1971] s'est servi des deux dimensions de la hauteur (hauteur tonale et hauteur spectrale) pour créer des illusions de son montant ou descendant indéfiniment, en modifiant numériquement et séparément les deux hauteurs : la hauteur spectrale reste constante (l'enveloppe spectrale est fixe), et seul le peigne harmonique se déplace en s'étirant pour respecter les relations harmoniques. Ainsi, une fois parvenu à l'octave, le son voit ses plus hautes harmoniques disparues, tandis que de plus basses réapparaissent régulièrement. L'oreille entend un son montant continûment. D'autres illusions ont été obtenues par la suite : l'illusion du rythme ralentissant lorsqu'on l'accélère, l'illusion de son descendant quand on le transpose vers le haut.

1.3 Représentation musicale - Prosodie

1.3.1 Représentation musicale

Dans la notation musicale occidentale, l'intensité sonore est appréhendée sur des échelles de temps de l'ordre de la seconde, sous une forme appelée **nuance**. Sa représentation évolue par pallier de 3 dB selon l'échelle *ppp, pp, p, mp, mf, f, ff, fff*. Les hauteurs sont appréhendées sous deux formes différentes : la mélodie et l'harmonie. La mélodie est constituée de suites dans le temps de sons de hauteurs définies. L'harmonie est constituée de hauteurs simultanées, et peuvent sous-tendre l'organisation des lignes mélodiques. La **durée** est représentée dans la notation occidentale par la forme reliée à la hauteur : ronde, croche. Les suites de durées sous-tendent un rythme. Le timbre correspond à un instrument, mais aussi à sa qualité sonore, pouvant évoluer selon les modes de jeu.

On le voit, l'échelle de temps de la musique n'est pas la même que celle de l'analyse ou de la perception auditive. L'écoute de la musique ne fait pas appel à la perception seule : l'étape de cognition est nécessaire pour appréhender l'organisation de ce qui est perçu par l'oreille.

1.3.2 Voyelles

Nous avons illustré le fait que pour une voix parlée ou chantée, les voyelles correspondent à une modification des résonances du système phonatoire, et permettant d'identifier plusieurs sons de même hauteur (*cf.* sec. 1.1.3). C'est à la base même de toutes les langues, même s'il a été montré que les consonnes ont encore plus d'importance. Lors du chant lyrique, l'articulation est modifié pour permettre de produire un signal sonore bien plus puissant, et les consonnes comme les voyelles en sont modifiées.

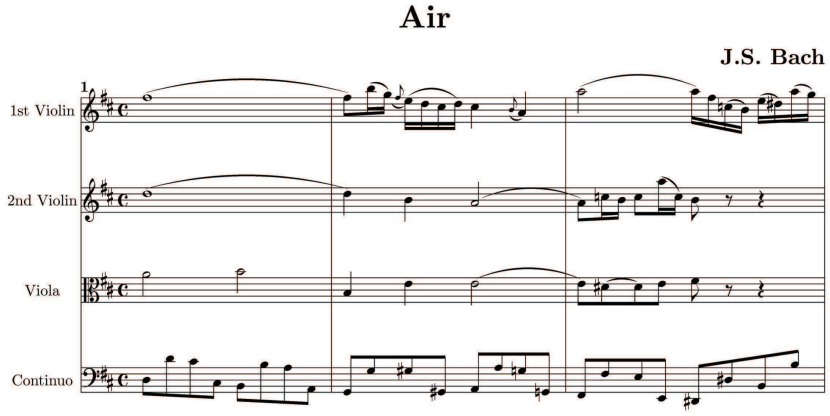


FIG. 1.9 – *Exemple de partition.*

1.3.3 Prosodie, intonation

La **prosodie** est le domaine particulier de la phonétique qui s'occupe de décrire les sons du langage au niveau de l'énoncé, l'énoncé pouvant être un mot, un groupe de mots ou une phrase. La prosodie s'attarde plus précisément à l'impression musicale que fournit l'énoncé. On y observe des phénomènes prosodiques tels que :

- l'**intonation** (fréquence) ;
- l'accentuation (durée, hauteur, intensité) ;
- le rythme (durée, intensité) ;
- le débit (durée) ;
- les pauses (durée).

Les paramètres principaux sont la variation de la fréquence fondamentale, la durée des portions de son (phonèmes) et les variations d'intensité d'une voix parlée.

Il existe quatre niveaux de la variation de la fondamentale d'une voix parlée (intonation, *cf. fig.* 1.10) :

- l'allure globale, obtenue par approximation de F_0 par un polynôme de degré 3 ;
- la structure constituante, obtenue ici après segmentation par approximation de F_0 par des polynômes de degré 2 sur chaque segment ;
- la mise en relief, obtenue ici après segmentation par approximation de F_0 par des polynômes de degré 5 ou 6 (6 dans l'exemple) sur chaque segment ;
- la micro-prosodie (F_0 sur chaque segment) ;

Les variations de F_0 sont liées à [Cristo, 1982] :

- la modalité de la phrase (qui implique un modelé intonatif) ;
- la structure constituante de la phrase ;
- l'état d'esprit et l'attitude du locuteur ;
- la quantité relative d'information apportée par les différentes unités sémantiques de la phrase (plus ou moins de détachement sémantique) ;
- la présence d'accents lexicaux et morphologiques ;
- la nature des unités segmentales (voyelles/consonnes) qui constituent le signal de parole, ainsi qu'à la concaténation de ces unités (co-articulation).

Chacune de ces unités implique des variations micro-prosodiques. On remarque deux catégories principales de variations pour la fréquence fondamentale :

- les variations dues à des instructions linguistiques (les formants linguistiques sont des unités discrètes et organisées hiérarchiquement en fonction de chaque langue) ;
- les variations dues à des contraintes inhérentes aux mécanismes (physiques) de la parole.

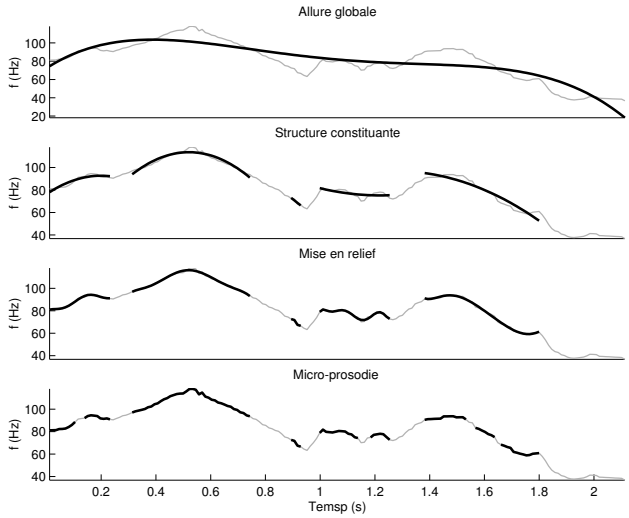


FIG. 1.10 – *Quatre niveaux de variation de la voix parlée : de l'allure globale à la micro-prosodie.*

Ces contraintes sont plus enclines à des variations libres ne dépendant pas d'une langue en particulier.

Enfin, on dénote deux fonctions de l'intonation :

- la fonction syntaxique qui permet essentiellement de différencier les types de phrase : déclarative, impérative, interrogative et exclamative.
- la fonction expressive qui permet de traduire une émotion, une opinion, un sentiment, etc.

Concernant la langue française, deux tendances phonétiques sous-tendent la prosodie : une tension généralisée et un caractère croissant. Ceci implique l'absence de diphtongaisons, de neutralisation, d'aspiration. Les paramètres clefs de la description de l'intonation sont la fréquence fondamentale intrinsèque des voyelles et des consonnes, l'intensité spécifique des consonnes et l'influence des consonnes sur la fréquence fondamentale des voyelles adjacentes (mécanismes de la co-articulation).

1.4 Définitions : traitement, effet, transformation et temps-réel

Nous allons maintenant définir les termes que nous utilisons pour parler de modifications d'un signal numérique. En effet, plusieurs termes existent, avec des nuances, et il est nécessaire de les connaître avant de les utiliser. Nous donnerons aussi les différents sens de l'expression "temps-réel" et préciserons laquelle nous utilisons.

1.4.1 Traitement

Un **traitement** correspond à un calcul appliqué à un signal, dans notre cas un signal numérique sonore. Les traitements sonores sont des calculs mathématiques, basées sur les opérations mathématiques algébriques : additions, multiplications. Ce terme comprend à la fois les effets sonores, les transformations sonores, et aussi d'autres traitement qui ne sont pas appliqués dans un but musical à proprement parler : débruitage et dé-cliquage d'enregistrement anciens, normalisation d'un signal avant de le stocker en mémoire informatique dans un fichier, codage–décodage compressif en vue du stockage, etc.

1.4.2 Effet

Il existe deux définitions du terme "**effet**", selon si l'on se place du côté de la perception ou de la modification d'un signal :

i) perception d'une transformation caractéristique du son (ce que l'on perçoit et uniquement cela) ;
ii) méthode de traitement et de modification d'un signal (changement en surface, par ajout ou soustraction).

Un effet, dans son acceptation "traitement de signal", en tant que traitement du son, est considéré comme un traitement de bas niveau ou de surface. De plus, il est employé principalement pour les traitements en temps-réel, qui au départ ne concernaient que les traitements faciles à mettre en œuvre (simplicité algorithmique) et rapides en terme de coût de calcul (faible coût de calcul), ce qui impliquait des traitements assez simples, d'où cet aspect de traitement en surface.

1.4.3 Transformation

Dans le cas du son, une **transformation** consiste à appliquer un traitement qui modifie le son en profondeur (une connaissance de la structure du signal est sous-jacente). Une transformation est considérée comme un traitement de haut niveau. De plus, le terme transformation a été employé principalement pour des traitements ne pouvant se réaliser en temps-réel. En effet, les méthodes d'analyse–transformation–synthèse ne sont pas encore toutes (et le seront-elles un jour ?) disponibles en temps-réel, car la méthode d'analyse elle-même peut nécessiter un traitement anti-causal. Cependant, la plupart des méthodes sont maintenant disponibles en temps-réel (PSOLA, le vocodeur de phase, le modèle additif avec un pistage de partiels causal).

Les musiciens se posent la question de ce qu'est une transformation de manière bien différente. Par exemple, Leigh Landy [Landy, 1991] considère qu'une transformation est une métamorphose timbrale. Il propose plusieurs catégories de transformations :

- sons comparables – sons incomparables ;
- sons discrets – sons continus ;
- sons longs – courts ;
- sons représentatifs – sons abstraits ;
- sons identiques – contextes différents.

Plusieurs types de métamorphoses sont possibles, selon le contrôle que l'on applique entre le son de départ et la cible : un contrôle linéaire, un contrôle concave ou convexe (allant à des vitesses de transformation différentes au début et à la fin), ou un contrôle imposant des étapes intermédiaires.

1.4.4 Quel terme utiliser ?

A la suite de ces définitions, il apparaît clair que si l'on veut parler aussi bien des effets que des transformations sonores, l'expression "traitements sonores" est la plus appropriée, ceci afin de se placer du côté de la personne qui conçoit et applique les traitements. Ainsi, on ne considère que les traitements ont un sens musical, ce qui revient à borner l'ensemble des traitements à l'union des effets et des transformations sonores. De plus, la distinction entre effet et transformation sur le plan de la mise en œuvre hors temps-réel ou en temps-réel n'a plus vraiment de raison d'être, étant donné la puissance de calcul et la rapidité des ordinateurs d'un part, et les méthodes existantes d'autre part.

La seule raison d'une éventuelle différentiation entre "effet" et "transformation" résiderait alors dans l'interprétation en terme de "traitement de surface" et "traitement en profondeur". Etant donnée la richesse de traitement offerts par les effets dits traditionnels, et la possibilité de réaliser des traitements simples à partir des modèles d'analyse–transformation–synthèse, nous pensons que la distinction surface/profondeur n'a de sens qu'en terme historique (l'origine des méthodes de traitement) et non en terme de capacité. A quelques exceptions près, les deux grandes classes de

traitements que sont “effets” et “transformation” peuvent traiter le son de manière similaire (mais pas forcément identique). Elles sont même complémentaires. Aussi, nous utiliserons par la suite (et même depuis la première page de ce document, avec le titre du sujet de recherche) le terme “effet” au sens de “traitement”, c'est-à-dire autant comme “effet” que “transformation” sonore.

1.4.5 Temps-réel

Il existe deux définitions pour l'expression “temps-réel”. La première définition d'un processus temps-réel est que son temps de calcul est inférieur au temps d'écoute du son produit, l'écoute pouvant être différée. La seconde définition d'un processus temps-réel est subordonnée à celle-ci, puisqu'il s'agit de processus dont l'exécution se fait en direct, dont le traitement d'un bloc d'échantillons est plus rapide que la lecture de ce bloc, et dont le temps de latence (la taille du bloc, en secondes) est suffisamment faible pour que l'oreille ne perçoive pas le décalage entre l'événement sonore original et l'événement sonore produit par le processus. Cette latence doit être de quelques millisecondes, inférieure à 20 ms. Lorsque nous parlons d'effets en temps-réel dans ce document, nous utilisons cette seconde définition, car tout l'enjeu d'ajouter un contrôle gestuel aux effets est de pouvoir jouer en direct sur l'effet et sur le son.

1.5 Taxonomies

Nous pouvons maintenant rechercher une taxonomie (ou classification) des méthodes et techniques de traitements du signal utilisées pour écrire des algorithmes d'effets audionumériques. Plusieurs approches permettent de présenter les traitements selon différentes classes. Nous en présentons six faisant sens : la taxonomie technologique, la taxonomie basée sur la typologie surface/profondeur, la taxonomie basée sur la typologie simple/complexe, la taxonomie méthodologique, la taxonomie cognitive et la taxonomie perceptive.

1.5.1 Taxonomie technologique ou historique

La taxonomie que nous appelons **technologique** est basée sur la technologie utilisée, elle-même dépendant de l'époque et des connaissances scientifiques et techniques, d'où l'appellation **historique**. Les effets sont alors classifiés selon qu'ils utilisent :

- ◇ du matériel analogique :
  - mécanique : réverbération naturelle (salle de concert), écho naturel (vallée, montagne ou très grande salle), filtrage naturel (structure architecturale) ;
  - mécanique et électromagnétique (magnétophones à bandes) : rééchantillonnage avec un disque en vinyl ou un rouleau de cire, transposition avec respect de l'échelle temporelle, et réciproquement, avec un magnétophone à bande et à têtes circulaires ;
  - électronique : modulation, filtrage ;
- ◇ du matériel numérique :
  - lignes à retard (simple, modulée) ;
  - filtres numériques ;
  - amplificateur et modification de gain ;
  - calculs complexes : modèle physique, système d'analyse–transformation–synthèse.

Cette taxonomie n'est pas satisfaisante, parce que plusieurs effets peuvent être réalisés en utilisant du matériel analogique ou numérique, même si la qualité sonore ne sera pas forcément la même.

1.5.2 Taxonomie basée sur la typologie : surface/profondeur

Nous avons vu que les traitements proviennent de deux classes pour des raisons qui nous semblent aujourd'hui plus méthodologiques et épistémologiques que perceptives (*cf.* sec. 1.4). Ces deux classes sont :

– les traitements en surface (originellement, les effets) ;
– les traitements en profondeur (originellement, les transformations).

Comme nous l'avons expliqué, cette classification nous semble peu satisfaisante ; on retrouve dans celle-ci une l'histoire des méthodes de traitement, et de la formalisation et classification des effets qui a été faite au fur et à mesure. Cette classification est faite par les ingénieurs et chercheurs en traitement numérique et analogique du signal, plutôt que par les musiciens, lesquels sont pourtant aussi des utilisateurs de ces techniques. Il manque de plus des distinctions plus précises quand aux types de traitements.

1.5.3 Taxonomie basée sur la typologie : simple/complexe

Une autre taxonomie basée elle aussi sur la typologie donne plus de détails quand aux effets en surface, en allant des plus simples aux plus complexes [Orfanidis, 1996]. Ainsi sont différenciés les effets basés sur :

– un délai simple : écho, filtrage en peigne ;
– un délai variable : chorus, *flanging*, *phasing*, vibrato ;
– plusieurs délais : échos multiples ;
– plusieurs délais et filtrages : réverbération ;
– un filtrage par banc de filtres : équalisation ;
– une modification de la dynamique : compresseur, expandeur, limiteur, *gate*.

Une fois encore, cette classification nous semble peu satisfaisante pour deux raisons. La première est qu'elle est incomplète : des transformations de haut-niveau ne sont pas présentes. Ceci peut se pallier en réalisant un inventaire des traitement manquants et en trouvant comment les regrouper sous les catégories présentes, quitte à en ajouter. La seconde raison est que pour le musicien, la même présentation détaillée techniquement et écrite en chinois n'en serait pas moins compréhensible. Cette taxonomie est une précision intéressante apportée à la taxonomie basée sur la typologie surface/profondeur sous forme d'une sous classification, mais elle nécessite de poursuivre nos investigations.

1.5.4 Taxonomie basée sur la méthodologie de traitement numérique

Nous nous concentrons maintenant sur les traitements numériques, puisque ce sont les traitements dont nous faisons l'étude. On peut présenter ces traitements en fonction des méthodes de mise en œuvre ; c'est le choix qui a été fait pour la rédaction du livre "*DAFx : Digital Audio Effects*" [Zoelzer, 2002]. Les méthodes ainsi mises en avant sont :

– filtrages ;
– lignes à retard ;
– modulateurs et démodulateurs ;
– traitements non-linéaires ;
– effets spatiaux ;
– traitements de segment-temporels ;
– traitements temps-fréquence (vocodeur de phase) ;
– traitements source-filtre (cepstre, LPC, vocodeur à canaux) ;
– traitements spectraux (basés sur une analyse additive) ;
– conformation temporelle et fréquentielle.

Cette classification est dite **méthodologique** et a pour grand mérite de permettre un panorama exhaustif des traitements, tout en gardant à l'esprit les classes de traitements en terme de méthode de mise en œuvre. Ainsi, si on dispose d'un système ne mettant en œuvre qu'une méthode de traitement, par exemple le vocodeur de phase, on sait tout de suite quels effets il permet de réaliser. Cependant, la limitation que nous y voyons se trouve toujours du côté de l'utilisateur musicien. S'il

veut réaliser une transposition, pourquoi ne pas lui proposer d'accéder directement à la transposition, quitte à lui proposer plusieurs méthodes pour le faire ? Avec la taxonomie méthodologique, il lui faudra chercher dans chaque méthode si elle permet la transposition, ce qui est trop laborieux et peu pratique à nos yeux.

1.5.5 Taxonomie basée sur la cognition

Une autre classification possible est donnée dans [Augoyard and Torgue, 1995], et permet plusieurs angles de lectures. Ce répertoire des effets sonores est le fruit de collaborations et recherches entre ingénieurs, chercheurs, architectes, urbanistes, sociologues, philosophes, géographes, musicologues et chercheurs du CRESSON[3], et porte sur différents niveaux d'interprétation et de compréhension des effets, d'où le nom de taxonomie **cognitive**. Les catégories proposées sont :

- **effets élémentaires** (ex : filtrage, distorsion, résonance, réverbération) : ils concernent la matière sonore (selon la taxonomie perceptive) ou la modalité de propagation. Ils sont enracinés dans la connaissance acoustique contemporaine et sont tous quantifiables ;
- **effets de composition** (ex : masquage, traînage, coupure, bourdon, effet téléphone) : "ils concernent des agencements sonores complexes et se sont définis par des caractères remarquables touchant soit à la dimension synchronique, soit à la dimension diachronique du contexte. Tous ces effets dépendants du dispositif spatio-temporel de la propagation sont sujets à évaluation physique pour une part au moins de leurs composantes" ;
- **effets liés à l'organisation perceptive** (ex : gommage, synecdoque, rémanence, anticipation, métabole) : "ils sont dus en priorité à l'organisation perceptive et mnémique des individus en situation concrète. On les repère toujours à partir d'une expression ou d'une aperception de la part des entendants. Par ailleurs, les caractères propres à la culture et la sociabilité de références sont partie prenante dans les particularités et la force de l'effet" ;
- **effets psychomoteurs** (ex : enchaînement, créneau, attraction, effet phonotonique) : "ces effets impliquent l'existence d'une action sonore de l'entendant ou tout au moins d'une esquisse motrice ou d'un schème faisant interagir perception et motricité" ;
- **effets sémantiques** (ex : décalage, imitation) : "ces effets sonores jouent sur l'écart de sens entre le contexte donné et la signification émergente. Il y a toujours décontextualisation, que ce soit sous la forme de l'imprévu anxiogène, de l'humour, du jeu conscient, ou d'une valeur esthétique ajoutée".

Cette classification présente certains avantages. Tout d'abord, les effets sont pris en compte dans les deux sens du terme, à la fois en tant que traitements sonores et en tant qu'effet perçu lors de l'audition de situations sonores, ce qui n'était pas encore le cas avec les taxonomies typologiques et méthodologiques. Ensuite, elle ne fait pas de distinction à priori selon la méthode mise en œuvre ni le matériel utilisée. De plus, la classification cognitive a été réalisée par une équipe de personnes d'horizons divers, ce qui lui confère une généralité et une lisibilité bien plus grande que les autres classifications, quant à elle réalisées par et pour les professionnels du traitement audionumérique. L'inconvénient que nous trouvons à cette classification vient de ce qu'elle échappe au cadre des traitements sonores pour le musicien et d'un point de vue technique pour englober un bien plus grand nombre de notions passionnantes, mais qui élargirait ce travail de recherche d'une telle manière qu'il faudrait plusieurs thèses pour en faire le tour ! Ce n'est pas un inconvénient en soi, mais plutôt une révélation des limites que nous nous fixons et hors desquelles cette classification s'aventure.

1.5.6 Taxonomie basée sur la perception

On peut enfin présenter les traitements en fonction des paramètres perceptifs qui seront modifiés dans le son. Nous avons choisi cette dernière classification, dite **perceptive**, pour présenter le reste

[3] CRESSON : centre de recherche sur l'espace sonore et l'environnement urbain, Grenoble.

de l'exposé, du fait qu'elle permet de se placer du point de vue (d'écoute) de l'auditeur musicien qui manipule des traitements audionumériques à sa disposition. On rappelle que les différentes dimensions perceptives d'un signal sonore sont :

- l'intensité sonore (sonie) ;
- la hauteur perçue (tonie) ;
- l'espace : il s'agit de la spatialisation (localisation des sources et de leurs déplacements) et du rayonnement ;
- la durée des sons, en secondes ;
- le timbre : il est alors représenté par un espace de timbres, dont les dimensions ne sont par les mêmes selon que l'on décrit un son tonal ou non.

Bien évidemment, on pourra aussi trouver des effets agissant sur plusieurs axes à la fois. Nous prenons aussi le parti de ne nous intéresser qu'aux traitements numériques, et d'oublier les effets du domaine de la cognition musicale. Les effets sur le timbre peuvent se classer selon les dimensions d'un espace de timbres. Cependant, nous avons préféré utiliser la taxonomie méthodologique, avec trois catégories : ceux qui ne modifient que le contenu fréquentiel (les valeurs de fréquence des partiels) et agissent sur le spectre, ceux qui modifient l'enveloppe spectrale et les formants (et donc l'amplitude des partiels), et ceux qui modifient les deux en même temps. Comme on le voit, la connaissance des autres taxonomies permet d'ordonner, pour chaque paramètre perceptif, les différents effets qui le modifient, et offrent un deuxième niveau de hiérarchie. Un travail de synthèse encore plus achevé consisterait à ordonner les effets par sous-paramètres perceptifs en utilisant entre autres les espaces de timbre ; c'est une tâche à laquelle nous avions commencé à réfléchir, mais qui s'est vite révélée trop délicate pour être menée dans cette étude. Nous nous somme donc limité à utiliser la taxonomie perceptive au premier niveau de classification.

C'est cette taxonomie que nous utiliserons par la suite, et qui s'explique dès à présent par l'attachement que nous avons de nous placer du côté du musicien en tant qu'auditeur du traitement sonore. En excluant les effets du domaine de la cognition musicale, nous prenons le risque de nous séparer du processus de composition, qui fait appel à la cognition. Cependant, notre but est de fournir des outils de traitement sonore, ainsi que des outils de composition du son à un niveau micro-temporel (micro-son [Roads, 1999; Roads, 2002]) et macro-temporel. Comme nous le verrons par la suite, le contrôle automatique d'effets offre encore plus de possibilités.

Chapitre 2

Méthodes de mise en œuvre d'effets audionumériques

La musique électronique a été critiquée en raison de l'absence d'interprète. Bien au contraire ! Le studio de musique électronique est difficile à maîtriser précisément car les compositeurs doivent eux-mêmes devenir des virtuoses. Une nouvelle éducation, à la fois technique et musicale, attend ceux qui osent s'aventurer au-delà du niveau superficiel de la séquence et de l'échantillonnage. Mais comment acquérir la connaissance nécessaire à la maîtrise de ce nouveau moyen d'expression ?
Curtis Roads [Roads, 1998]

Sommaire

Nous présentons tout d'abord des bases sur les méthodes et techniques numériques utilisées pour les traitements audionumériques. Nous présenterons les filtrages, les lignes à retard, les modulateurs et démodulateurs, et enfin les méthodes d'analyse–transformation–synthèse. Nous terminerons par quelques remarques sur le repliement de spectre. Ces bases seront utilisées pour tout le chapitre 3 sur les effets audionumériques usuels ainsi que le chapitre 5 sur les effets audionumériques adaptatifs. La classification utilisée ici est méthodologique. En effet, les effets audionumériques traditionnels et adaptatifs seront présentés selon la taxonomie perceptive, mais les questions sur la méthodologie que pourra se poser le lecteur trouvent plus aisément une réponse lorsque l'information est donnée selon la classification la plus adaptée.

2.1 Filtres

Les filtres peuvent se classer selon le type de leur réponse impulsionnelle : finie ou infinie. Ils peuvent aussi se classer en fonction de la forme de leur réponse en fréquence : les filtres à étage (passe-haut, passe-bas et passe-bande), les filtres passe-tout et les filtres en pic. Nous allons passer en revue les filtres selon ces deux classifications.

2.1.1 Filtre à réponse impulsionnelle finie (RIF)

Un filtre à réponse impulsionnelle finie est un filtre dont la réponse à un signal impulsionnel est de longueur finie N. Un échantillon en sortie d'un tel filtre s'obtient par une somme pondérée des valeurs des N derniers échantillons entrants, *cf. eq.* (2.1).

$$y(n) = a_0\ x(n) + \sum_{i=1}^{N} a_i\ x(n-i) \tag{2.1}$$

2.1.2 Filtre à réponse impulsionnelle infinie (RII)

Un filtre à réponse impulsionnelle infinie est un filtre dont la réponse à un signal impulsionnel peut être de longueur infinie. Un échantillon en sortie d'un tel filtre s'obtient par une somme pondérée des valeurs de l'échantillon entrant et des N échantillons sortants, *cf. eq.* (2.2). Un bouclage a donc lieu entre la sortie et l'entrée du filtre RII.

$$y(n) = a_0\ x(n) + \sum_{i=1}^{N} b_i\ y(n-i) \tag{2.2}$$

2.1.3 Filtre à étage (passe-haut ou passe-bas)

Un filtre à étage est soit passe-bas (qui ne laisse passer que les fréquences en dessous de la fréquence de coupure f_c), soit passe-haut (qui ne laisse passer que les fréquences au dessus de la fréquence de coupure), comme représenté *fig.* 2.1, soit les deux à la fois.

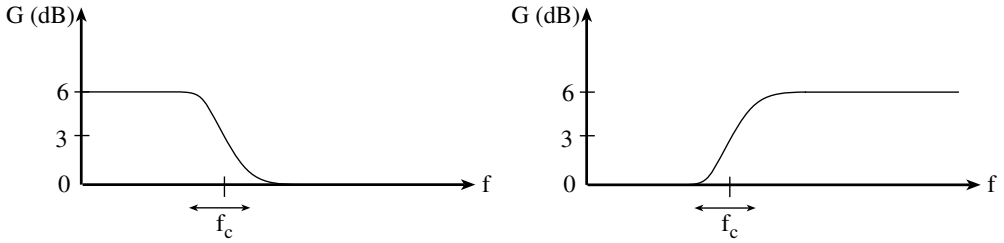


FIG. 2.1 – *Réponse en fréquence des filtres passe-bas (à gauche) et passe-haut (à droite).*

Un filtre passe-bas du premier ordre est décrit par sa fonction de transfert :

$$H(z) = 1 + \frac{H_0}{2}\ [1 + A(z)] \tag{2.3}$$

tandis qu'un filtre passe-haut du premier ordre est décrit par sa fonction de transfert :

$$H(z) = 1 + \frac{H_0}{2}\ [1 - A(z)] \tag{2.4}$$

avec $A(z)$ le filtre passe-tout du premier ordre :

$$A(z) = \frac{z^{-1} + a}{1 + az^{-1}} \tag{2.5}$$

et a la fréquence de coupure. Sa valeur diffère selon que l'on veut renforcer la bande de fréquence (a_B) ou l'atténuer (a_C), avec f_c la fréquence de coupure du filtre :

$$a_B = \frac{\tan(\pi f_c/F_e) - 1}{\tan(\pi f_c/F_e) + 1} \qquad \text{et} \qquad a_C = \frac{\tan(\pi f_c/F_e) - V_0}{\tan(\pi f_c/F_e) + V_0} \tag{2.6}$$

Le facteur gain G en décibels peut être ajusté à l'aide du paramètre : $H_0 = V_0 - 1$ avec $V_0 = 10^{G/20}$. La pente de ces filtres est de 6 dB par octave. Du fait que le filtre à étage du premier ordre possède un pôle et un zéro, sa réponse en fréquence possède un plateau (une tangente horizontale) au gain G désiré pour la fréquence nulle et un autre à 0 dB pour la fréquence maximale. La fréquence de coupure indique la zone où la réponse en fréquence chute vers 0. Les filtres à étage du second ordre sont quant à eux à la fois passe-haut et passe-pas, du fait qu'ils contiennent deux pôles et deux zéros : ils sont définis par deux fréquences de coupure. Ils permettent notamment d'obtenir des filtres de pente plus élevée que celle des filtres du premier ordre.

2.1.4 Filtre en pic (passe-bande)

Un filtre en pic (ou passe-bande) est donné par sa réponse en fréquence :

$$H(z) = 1 + \frac{H_0}{2}\left[1 - A_2(z)\right] \tag{2.7}$$

avec le filtre passe-tout d'ordre deux :

$$A_2(z) = \frac{-a + d(1-a)\ z^{-1} + z^{-2}}{1 + d(1-a)\ z^{-1} - a\ z^{-2}} \tag{2.8}$$

Le paramètre d relatif à la fréquence centrale du filtre et le coefficient H_0 sont donnés par : $d = -\cos(2\pi\ f_c/F_e)$, $V_0 = 10^{G/20}$ et $H_0 = V_0 - 1$

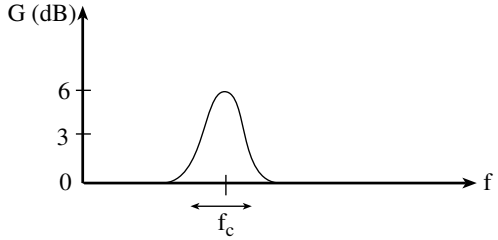


FIG. 2.2 – *Réponse en fréquence du filtre passe-bande.*

La largeur de bande est ajustée à l'aide des paramètres a_B pour augmenter le gain et a_C pour diminuer le gain :

$$a_B = \frac{\tan(\pi f_b/F_e) - 1}{\tan(\pi f_b/F_e) + 1} \qquad \text{et} \qquad a_C = \frac{\tan(\pi f_b/F_e) - V_0}{\tan(\pi f_b/F_e) + V_0} \tag{2.9}$$

2.2 Lignes à retard

Les lignes à retard sont utilisées dans un grand nombre d'effets, notamment pour des raisons historiques. Les effets d'espace tels l'écho, la réverbération ainsi que les effets de filtrage basés sur le déphasage (chorus, *flanger, phaser*) sont basés sur les lignes à retard. Une ligne à retard étant mise en œuvre sous forme d'une suite d'échantillons stockés, sa longueur est un discrétisée : c'est un nombre entier d'échantillons. Il est donc nécessaire d'utiliser des lignes à retard fractionnaires si l'on veut des temps de délais non entiers [Laakso *et al.*, 1996].

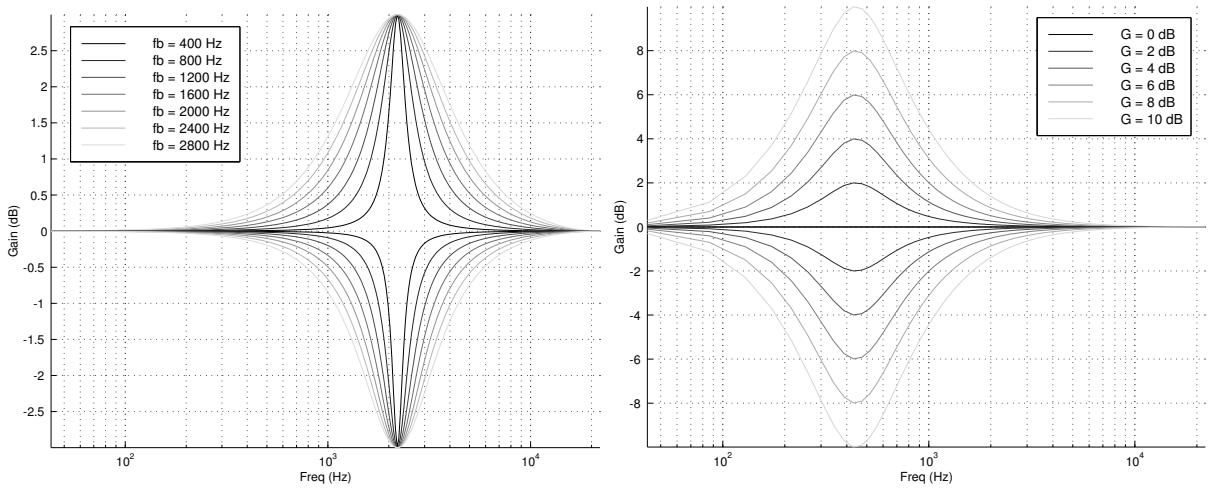


FIG. 2.3 – *Réponse en fréquence du filtre passe-bande pour différentes largeurs de bandes (différentes valeurs de Q) à gain constant (*fig. *gauche) et pour différents gains à Q constant (*fig. *droite).*

2.2.1 Principe

Une ligne à retard est une structure tampon dans laquelle les échantillons de sortie sont identiques aux échantillons qui y sont entrés, avec un temps de retard dépendant de la longueur L de la ligne à retard (en nombre d'échantillons).

2.2.2 Filtre en peigne (RIF) à l'aide de lignes à retard

Soit $x(n)$ un signal entrant dans une ligne à retard de longueur L. Si l'on ajoute le signal en entrée de la ligne à retard au signal en sortie, avec g_x le gain, on obtient le signal :

$$y(n) = x(n) + g_x \; x(n - L) \tag{2.10}$$

correspondant au filtrage en peigne à réponse impulsionnelle finie (RIF), de fonction de transfert :

$$H(z) = 1 + g_x \; z^{-L} \tag{2.11}$$

Le passage dans la ligne à retard induit un délai temporel (ou temps de retard) $\tau = \frac{L}{F_e}$.
Dans le cas où le gain g_x est positif, le filtre amplifie toutes les fréquences multiples de $\frac{1}{\tau} = \frac{F_e}{L}$ et atténue toutes les fréquences entre ces multiples. Le gain varie en amplitude entre $1 - g_x$ et $1 + g_x$, et en décibels entre $20 \log_{10}(1 - g_x)$ et $20 \log_{10}(1 + g_x)$ dB. La réponse en fréquence de ce filtre ressemble à un peigne, les trous dans le spectre sont de plus en plus prononcés lorsque le gain $|g_x|$ augmente en valeur absolue (*cf. fig.* 2.4). Selon le délai du filtre, le son produit sera perçu comme une coloration (transformation du timbre, pour des τ petits) ou comme un écho (τ grand). On peut aussi réaliser des filtres en peigne RII avec une ligne à retard (cf. 3.6.1).

2.2.3 Synthèse de sons (percussifs et entretenus)

La ligne à retard est l'équivalent numérique du guide d'onde physique parfait, ce qui explique son utilisation très fréquente pour la modélisation de sons instrumentaux. Les cordes vibrantes et les colonnes d'air sont des exemples de guides d'ondes. On retrouve ces idées de modélisation depuis les travaux de Kelly et Lochbaum [Kelly and Lochbaum, 1962] puis ceux de Karplus et Strong [Karplus and Strong, 1983] ; l'utilisation qui en est faite aujourd'hui dépasse les simples guides d'ondes tubulaires, puisque des équipes [Smith, 1987; Kronland-Martinet *et al.*, 1997; Avanzini and Rocchesso, 2000; Aramaki *et al.*, 2002] travaillent sur ces structures pour modéliser des plaques vibrantes (table d'harmonie de piano ou de guitare) par exemple.

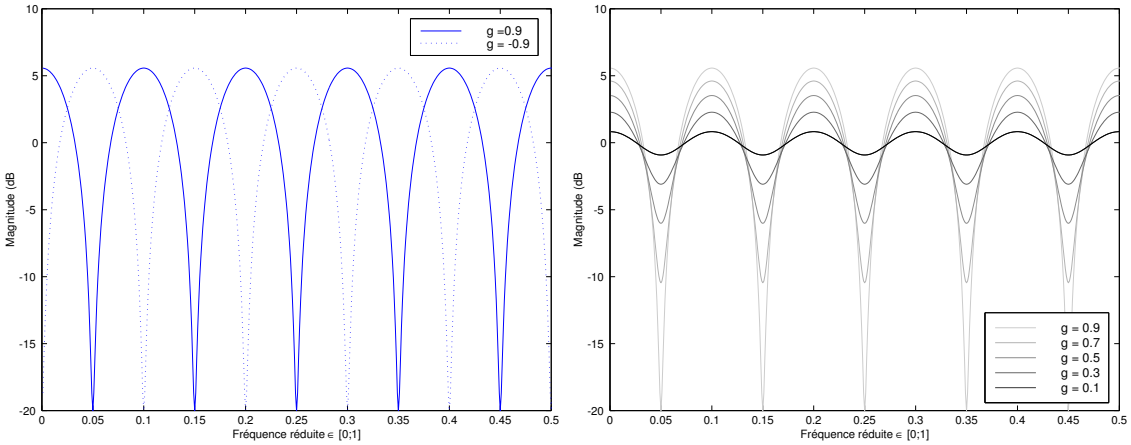


FIG. 2.4 – *Réponse en fréquence d'un filtre en peigne FIR : à gauche pour $g_x = 0.9$ et $g_x = -0.9$, à droite pour $g_x = 0.1$, $g_x = 0.3$, $g_x = 0.5$, $g_x = 0.7$, $g_x = 0.9$.*

2.2.4 Retard fractionnaire

Pour certains effets utilisant des lignes à retard, la précision temporelle donnée par l'échantillonnage n'est pas suffisante. Ainsi, pour des lignes à retard dont la longueur peut varier, avec $F_e = 44,1\ kHz$, un temps de retard de 10 ms sera modélisé par une ligne à retard de 44,1 échantillons, ce qui n'est pas réalisable avec des retards entiers. Dans ce cas, on utilise des lignes à retard plus grandes avant de procéder à une interpolation. Soit $d \in \mathbb{R}$ le délai en nombre d'échantillons. Il peut se décomposer en une partie entière $M \in \mathbb{N}$ et une partie réelle $\gamma \in]0;1[$. Le retard fractionnaire peut alors être modélisé comme la sortie d'une ligne à retard fractionnaire $y(n) = x(n-d) = x(n-(M+\gamma))$.

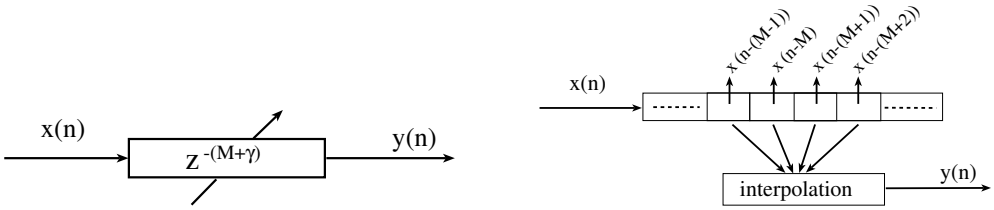


FIG. 2.5 – *Diagrammes de la ligne à retard fractionnaire : représentation "traitement du signal" (à gauche), détail de la méthodologie utilisée (à droite).*

Pour calculer la valeur de $x(n-d)$, on procède à une interpolation à partir des valeurs voisines (au moins deux). Différentes interpolations peuvent être utilisées [Laakso *et al.*, 1996; Zoelzer, 2002] :

- linéaire : $y(n) = \gamma.x(n-(M+1)) + (1-\gamma).x(n-M)$
- passe-tout : $y(n) = \gamma.x(n-(M+1)) + (1-\gamma).x(n-M) - (1-\gamma).y(n-1)$
- splines : $y(n) = \frac{\gamma^3}{6}.x(n-(M+1)) + \frac{(1+\gamma)^3-4\gamma^3}{6}.x(n-M) + \frac{(2-\gamma)^3-4(1-\gamma)^3}{6}.x(n-(M-1)) + \frac{(1-\gamma)^3}{6}.x(n-(M-2))$

2.3 Modulateurs

Pour tous les modulateurs, on considère un signal $x(n)$ appelé "porteuse", modulé en amplitude (ou en anneau, ou encore BLU), en fréquence, ou en phase par une modulante $m(n)$. En télécommunications, la porteuse est un signal composé d'une seule sinusoïde. En tant qu'effet musical, la modulation peut se faire avec un signal complexe comme porteuse.

2.3.1 Modulateur en amplitude, en anneau et à bande latérale unique

Modulation en anneau

Le modulateur en anneau (*ring modulator*) est un multiplicateur de deux signaux, échantillon par échantillon. Les modulateurs en anneau analogiques étaient réalisés à l'aide d'un circuit de 4 diodes disposées en anneau, d'où le nom. Le diagramme correspondant à la modulation en anneau est donné en *fig.* 2.6.

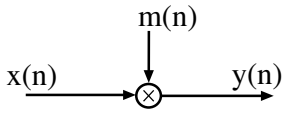


FIG. 2.6 – *Diagramme du modulateur en anneau.*

Soient deux signaux sinusoïdaux x et m :

$$x(n) = x_0 \sin(2\,\pi\,f_x\,n + \phi_x) \quad (2.12)$$
$$m(n) = m_0 \sin(2\,\pi\,f_m\,n + \phi_m) \quad (2.13)$$
$$y(n) = x(n).m(n) \quad (2.14)$$
$$= x_0\,m_0 \sin(2\,\pi\,f_x\,n + \phi_x)\,.\,\sin(2\,\pi\,f_m\,n + \phi_m) \quad (2.15)$$

En utilisant la formule trigonométrique usuelle :

$$\sin(a)\sin(b) = \frac{\cos(a-b) - \cos(a+b)}{2} \quad (2.16)$$

on obtient :

$$y(n) = \frac{x_0\,m_0}{2}\left(\cos\left[2\pi n(f_m - f_x) + (\phi_m - \phi_x)\right] - \cos\left[2\pi n(f_m + f_x) + (\phi_m + \phi_x)\right]\right) \quad (2.17)$$

c'est-à-dire la somme de deux signaux sinusoïdaux de fréquences $f_m + f_x$ et $f_m - f_x$ (*cf.* ex. *fig.* 2.7). Pour des fréquences de modulation en dessous de 20 Hz, la modulation sera entendue dans le domaine temporel (ex : trémolo), alors qu'au dessus de 20 Hz, elle sera entendue dans le domaine fréquentiel comme la somme de trois signaux : la porteuse $x(n)$, le signal composé des sommes et celui composé des différences de fréquences. Cette différence de perception provient du temps d'intégration de l'information de notre système auditif.
Dans le cas où le signal x est harmonique, la lecture du spectre devient moins aisée, comme en atteste la figure *fig.* 2.7. Prenons maintenant le cas où les signaux x et m sont harmoniques :

$$x(n) = \sum_{k=1}^{H_x} x_k \sin(2\,\pi\,k\,f_x\,n + \phi_{x,k})$$
$$m(n) = \sum_{l=1}^{H_m} m_l \sin(2\,\pi\,l\,f_m\,n + \phi_{m,l})$$
$$y(n) = \sum_{k=1}^{H_x}\sum_{l=1}^{H_m} x_k\,m_l \sin(2\,\pi\,k\,f_x\,n + \phi_{x,k})\,.\,\sin(2\,\pi\,l\,f_m\,n + \phi_{m,l})$$
$$= \sum_{k=1}^{H_x}\sum_{l=1}^{H_m} x_k\,m_l \left(\cos\left[2\pi n(l\,f_p - k\,f_x) + (\phi_{m,l} - \phi_{x,k})\right]\right.$$
$$\left. - \cos\left[2\pi n(l\,f_m + k\,f_x) + (\phi_{m,l} + \phi_{x,k})\right]\right)$$

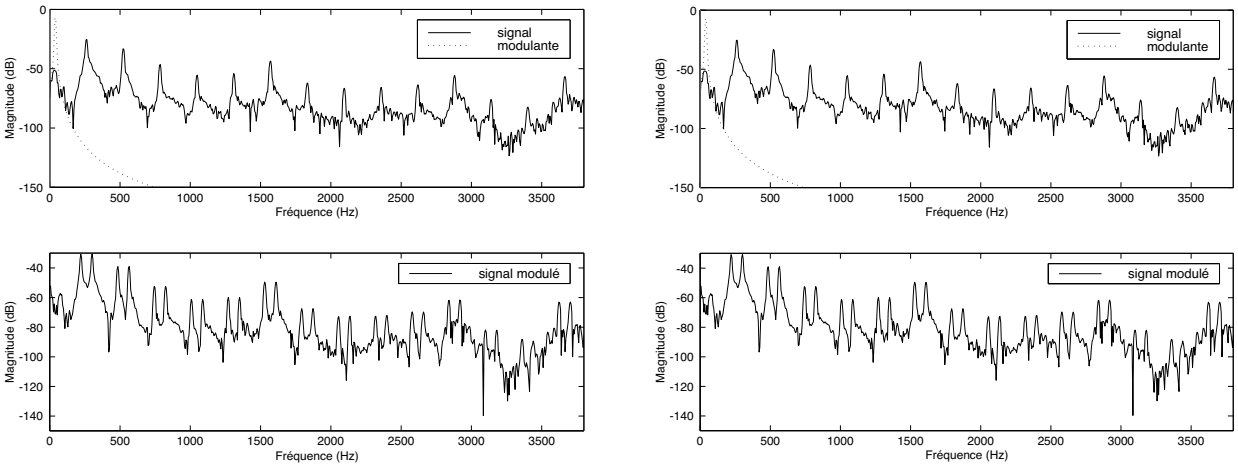


FIG. 2.7 – *Modulation en anneau.*
A gauche : avec une fréquence de 500 *Hz pour la modulante, appliquée à un signal contenant une sinusoïde pure à* 2000 *Hz.*
A droite : avec une fréquence de 40 *Hz pour la modulante, appliquée à un signal réel harmonique de fondamentale* 200 *Hz.*

Si le rapport entre les fréquences f_m et f_x est une fraction entière, par exemple $f_m = \frac{M}{N}\ f_x$, alors le signal résultant de la modulation en anneau est harmonique de fréquence fondamentale $\frac{f_x}{N}$:

$$\begin{aligned} y(n) \quad = \quad & \sum_{k=1}^{H_x} \sum_{l=1}^{H_m} x_k\ m_l\ \left(\cos\left[2\ \pi\ \left(l\frac{M}{N} - k \right) f_x\ n + (\phi_{m,l} - \phi_{x,k}) \right] \right. \\ & \left. - \cos\left[2\ \pi\ \left(l\frac{M}{N} + k \right)\ f_x\ n + (\phi_{m,l} + \phi_{x,k}) \right] \right) \end{aligned}$$

Les figures *fig.* 2.8 illustrent ceci pour deux signaux harmoniques composés de 10 sinusoïdes d'amplitudes $1/k^2$, k étant le numéro de l'harmonique. Le signal à traiter a pour fondamentale $f_x = 400\ Hz$, la fondamentale de la modulante prenant quant à elle des valeurs $f_m = \frac{P}{4}\ f_x$, multiples du quart de f_x.

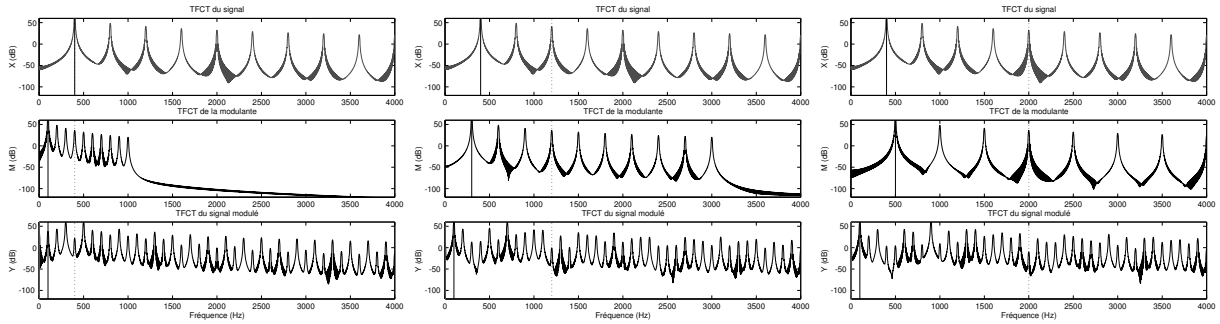


FIG. 2.8 – *Modulation en anneau de deux signaux harmoniques de fréquences fondamentales* 400 *Hz et* $\frac{M}{N}$100 *Hz pour la modulante, soit un rapport* $\frac{M}{N} = 1/4$ *à gauche,* $\frac{M}{N} = 3/4$ *au centre et* $\frac{M}{N} = 5/4$ *à droite.*

La fréquence fondamentale des trois signaux obtenus dans l'exemple donné *fig.* 2.8 est identique ; cependant, les spectres sont différents, du fait que les amplitudes des pics ne sont pas les mêmes. Il en résulte des timbres différents, pouvant aller jusqu'à des sons dont la fondamentale ne s'entend plus et dont un groupe d'harmoniques est prédominant, ce qui donne une hauteur perçue différente

de la fondamentale. Aussi, selon l'effet que l'on veut rendre (choisir la fondamentale comme hauteur perçue, ou non), on choisira avec parcimonie le ratio $\frac{M}{N}$.

Modulateur d'amplitude

La modulation d'amplitude (*AM : amplitude modulation*) consiste à appliquer une modulation en anneau de modulante à basse fréquence $m(n)$ puis à l'additionner au signal de départ :

$$y(n) = [1 + \alpha\, m(n)]\, .x(n) \tag{2.18}$$

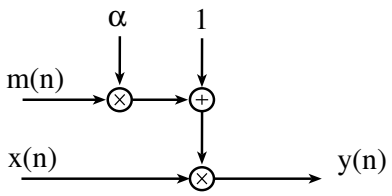


FIG. 2.9 – *Diagramme du modulateur d'amplitude.*

Le signal de modulation $m(n)$ a pour amplitude maximale 1 ; α est le coefficient déterminant l'amplitude de la modulation. La perception de cet effet est la même en dessous de 20 *Hz* ; par contre, au dessus, la transformation du son est plus complexe, puisque résultant de l'addition du son de modulation en anneau (décalage du spectre vers la gauche et recopie vers la droite) et du son original.

Modulateur BLU ou à bande latérale unique

Principe Dans le cas des signaux sonores (ie. réels), les fréquences positives et négatives du spectre transportent la même information, bien qu'organisée différemment. En effet, soit $x(n) \in \mathbb{R}$, avec $X(k)$ la transformée de Fourier à temps discret (définie en *eq.* (2.29) avec le vocodeur de phase) pour $k = 0, \ldots, N-1$. On a l'égalité $X(k) = \overline{X}(N-k)$ (avec $\overline{X}$ le conjugué). L'idée utilisée pour le modulateur BLU (ou *(SSB, Single Side Band)*) était de ne transmettre que la moitié pertinente de l'information pour la transmission.

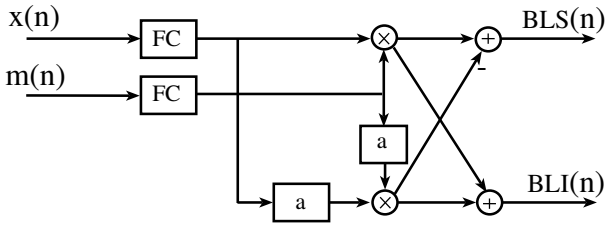


FIG. 2.10 – *Diagramme du modulateur BLU.*

Mise en œuvre Tout d'abord, on déphase de 90° les phases du signal. Ceci se fait à l'aide de la Transformée de Hilbert à temps discret. Elle se modélise par un filtre RIF, de réponse impulsionnelle :

$$a(n) = \frac{1-\cos(\pi\, n)}{2} = \begin{cases} \frac{2}{\pi\, n} \text{ si n est impair} \\ 0 \text{ si n est pair} \end{cases} \tag{2.19}$$

$$\overline{a}(n) = a(n).w(n) \tag{2.20}$$

où w est une fenêtre de taille N, par exemple une fenêtre de Hamming. Si l'on note $\hat{x} = x * \overline{a}$ et $\hat{m} = m * \overline{a}$ les signaux dont la phase est décalée de 90°, on obtient l'expression analytique des signaux BLI (bande latérale inférieure, ou *LSB, Lower Side Band*) et BLS (bande latérale supérieure, ou *USB, Upper Side Band*) :

$$BLS(n) = x(n)\, m(n) - \hat{x}(n)\, \hat{m}(n) \tag{2.21}$$

$$BLI(n) = x(n)\, m(n) + \hat{x}(n)\, \hat{m}(n) \tag{2.22}$$

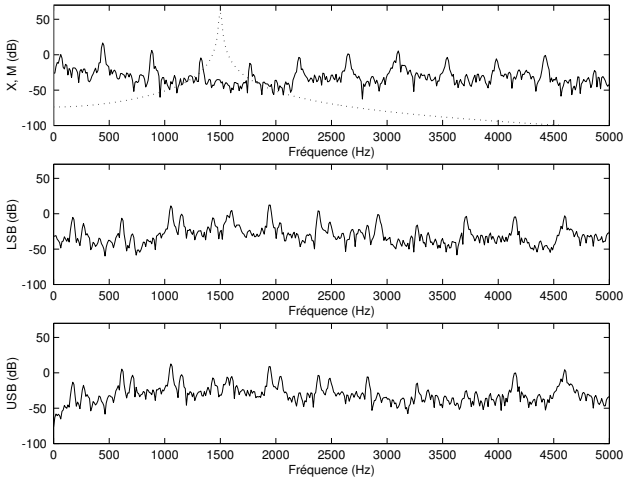


FIG. 2.11 – *Modulation BLU d'un signal harmonique complexe.*

A partir de $N = 60$, les résultats sonores sont de qualité acceptable.

2.3.2 Modulateur de fréquence et de phase

La modulation de fréquence (MF ou *FM : Frequency Modulation*) a connu de nombreuses applications dans la diffusion (télé-diffusion, radio-diffusion) ainsi que dans la synthèse sonore, grâce aux travaux de John Chowning [Chowning, 1971b] dans les années 70 (et son brevet avec Yamaha sur le DX7, premier synthétiseur FM, au début des années 80).

Principe : le signal modulé est une sinusoïde de fréquence porteuse f_p et de phase $\Phi(t)$ fonction d'une modulante $m(t)$. La formulation analytique est :

$$x_{MF/MP}(t) = A_p \cos\left(2\pi f_p t + \Phi(t)\right) \tag{2.23}$$

avec $\Phi(t) = k_{MP}\ m(t)$ pour la modulation de phase (la phase est directement proportionnelle à la modulante) et $\Phi(t) = 2\pi\ k_{MF} \int_{-\infty}^{t} m(\tau) d\tau$ pour la modulation de fréquence (la phase est proportionnelle à l'intégrale de la modulante).
Dans le cas des effets sonores, la modulation d'angle n'est plus tout à fait la même chose. L'idée est toujours d'utiliser une modulante $m(n)$ pour modifier la phase $\Phi(t)$ d'une sinusoïde porteuse, de fréquence f_p. Cette fois-ci, la représentation utilisée est un filtre à réponse impulsionnelle $h(n)$ variant dans le temps, définie par :

$$h(n) = \delta\left(n - m(n)\right) \tag{2.24}$$

Le signal résultant de ce filtrage est un signal à modulation de phase :

$$y(n) = x(n) * h(n) = x(n - m(n)) = x_{MP}(n) \tag{2.25}$$

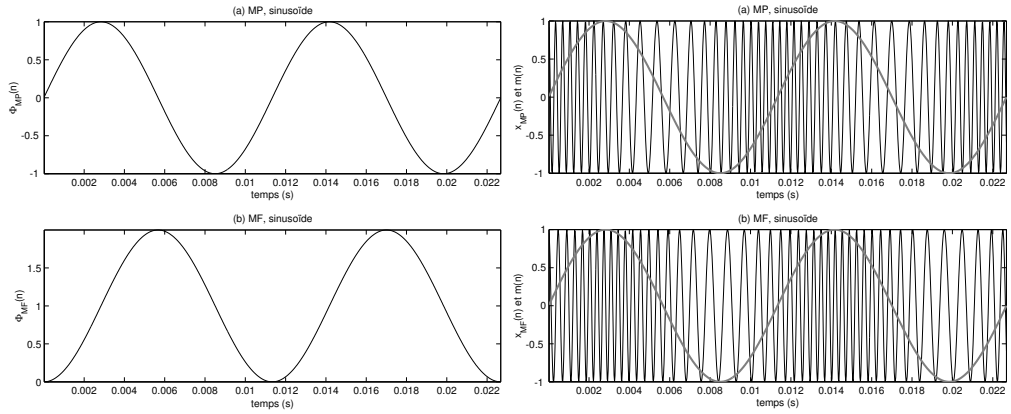


FIG. 2.12 – *Modulation en fréquence et en phase pour une modulante sinusoïdale.*

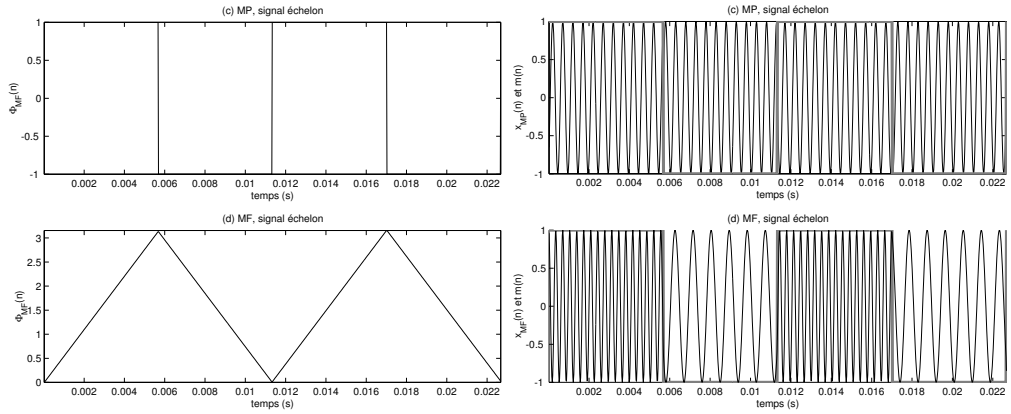


FIG. 2.13 – *Modulation en fréquence et en phase pour une modulante "échelon".*

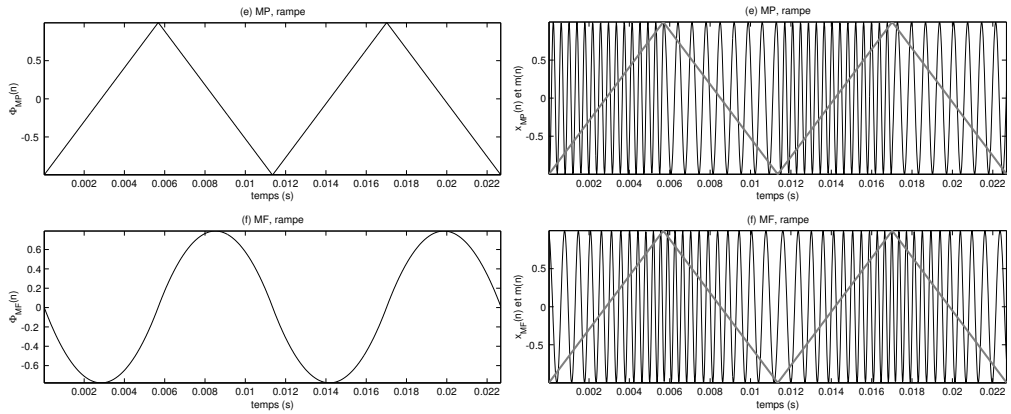


FIG. 2.14 – *Modulation en fréquence et en phase pour une modulante "rampe".*

La transformée de Fourier de ce signal est :

$$Y\left(e^{j\Omega}\right) = X_{MP}\left(e^{j\Omega}\right) = X\left(e^{j\Omega}\right)\ e^{-j\Omega m(n)} \tag{2.26}$$

On voit bien apparaître dans le dernier terme la modulation sur la phase. La variable $m(n)$ est continue, et peut donc se décomposer en une partie entière M et une partie fractionnaire γ (*cf.* les délais fractionnaires 2.2.4). Le fait que le temps de retard varie implique que la partie de signal dupliquée est lue à une distance plus où moins grande, donc à un taux d'échantillonnage différent

de celui d'écriture ; il en résulte une différence de hauteur perçue.

Modulante sinusoïdale : le signal de modulation s'écrit : $m(n) = M + d\ \sin(\omega_M nT)$ avec d la profondeur de la modulation, ω_M la fréquence de la modulation (en radians). Pour un signal d'entrée sinusoïdal, le facteur de ré-échantillonnage s'obtient ainsi : $\alpha(n) = \frac{\omega_I}{\omega} = 1 - \gamma\ \omega_M\ T\ \cos(\omega_M\ n\ T)$ avec ω_I la fréquence instantanée du signal sortant (en radians), ω la fréquence de la sinusoïde entrant (en radians). La valeur moyenne de $\alpha(n)$ est $< \alpha(n) >= 1$, ce qui signifie qu'en moyenne, le signal de sortie sera de même longueur, avec une variation de hauteur autour de la fréquence de la sinusoïde entrante (vibrato).

Modulante rampe : le signal de modulation s'écrit $m(n) = M \pm p\ n$ avec p la pente. Le facteur de ré-échantillonnage se dérive ainsi : $\alpha(n) = 1 \mp p$, d'où un signal en sortie de longueur $L_{sortie} = \frac{L_{entree}}{\alpha}$, correspondant à la **transposition** d'un facteur α (ie. $F_{sortie} = \alpha F_{entree}$).

2.3.3 Démodulateurs

A chaque méthode de modulation est associée une méthode de démodulation. Un démodulateur est constitué d'un détecteur, d'un moyenneur et d'un élément de mise à l'échelle (*scaler*), *cf. fig.* 2.15. Ce dernier est très souvent partie intégrante du matériel recevant le signal démodulé.

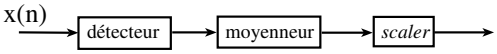


FIG. 2.15 – *Diagramme du démodulateur.*

Les détecteurs et moyenneurs sont quant à eux utilisés pour réaliser le VU-mètre, le *Peak-meter*, le RMS, pour le suivi d'enveloppe, mais aussi dans différents effets ou méthodes d'analyse-synthèse, aussi nous allons détailler un peu plus ce dont il s'agit.

Détecteurs

Soit $x(n)$ un signal et $\hat{x}(n)$ sa Transformée de Hilbert à temps discret. Il y en a quatre principaux. Le détecteur $d_h(n) = max\,(0, x(n))$ est appelé rectificateur demi-forme d'onde (*half-wave rectifier*), $d_f(n) = |x(n)|$ est le rectificateur (*full-wave rectifier*). Le détecteur $d_r(n) = x^2(n)$ correspond au carré (*squarer*), et est notamment utilisé pour le calcul du RMS. Enfin, $d_i^2(n) = x^2(n) + \hat{x}^2(n)$ est le détecteur d'enveloppe instantanée.

Moyenneurs

Un moyenneur est un filtre du premier ordre. A l'origine, il s'agissait d'un réseau Résistance-Capacité en circuit analogique. Soit $g = \exp(-1/F_e\tau)$ le gain, τ le temps de montée du moyenneur et $d(n)$ la sortie d'un détecteur. Le moyenneur le plus simple est donné par :

$$y(n) = (1 - g)\ d(n) + g\ y(n - 1) \tag{2.27}$$

Dans le cas général, ce moyenneur n'est pas suffisant, du fait qu'aucune distinction ne soit faite entre la montée et la descente en énergie du signal. Aussi, un second moyenneur, appelé moyenneur AR (auto-régressif) à deux constantes de temps a été développé. Pour la montée ou l'attaque, on utilise $g = \exp(-1/(f_e\tau_a))$, avec $\tau_a \approx 5\ ms$. Pour la descente ou relaxation, on utilise $g = \exp(-1/(f_e\tau_r))$, avec $\tau_r >> \tau_a$.

$$\begin{aligned} y_{ar}(n) &= (1 - g)\ d(n) + g\ y_{ar}(n - 1) \\ g &= g_a \text{ si } y_{ar}(n - 1) < d(n) \\ &= g_r \text{ sinon} \end{aligned}$$

2.4 Systèmes d'analyse – modification – synthèse

Contrairement aux traitements utilisant les techniques présentées jusqu'ici (filtrage, modulation de lignes à retard, traitements temporels) qui ne font aucune hypothèse sur le signal (ie. sur sa nature) et peuvent être utilisés pour tout type de son, les systèmes d'analyse–transformation–synthèse sont fondés sur des modèles de signal. L'hypothèse faite sur le signal permet d'aboutir à des méthodes de modification mieux adaptées, et souvent plus performantes. Quatre méthodes (ou modèles) sont présentés :

- les méthodes **segment-temporelles**, ne faisant pas d'hypothèse sur le signal (méthode non paramétrique), utilisées uniquement pour des changements d'échelle temporelle (dilatation/contraction) ou fréquentielle (transposition) ;
- le **vocodeur de phase** (méthode temps-fréquence), avec pour hypothèses : le signal est d'énergie finie, à variations lentes et décomposable en grains de sons stationnaires (méthode non-paramétrique) ;
- le **modèle additif** (spectral), où le signal est modélisé comme une somme de sinusoïdes de modules et fréquences variant lentement dans le temps, plus une composante transitoire d'attaque et une composante bruitée résiduelle pour le modèle le plus raffiné (méthode paramétrique) ;
- le **modèle soustractif** ou source-filtre (ou source-filtre), où le signal provient d'une source (ou excitation) riche en harmoniques, filtrée par un filtre, encore appelé résonance (méthode paramétrique).

Très peu de nouveaux effets ont été proposé utilisant la technique des ondelettes, ce qui explique le fait qu'elle ne soit pas présentée ici.

Nous verrons au chapitre 3 les utilisations que l'on peut faire de ces quatre modèles en vue de transformations sonores.

2.4.1 Méthodes segment-temporelles

Nous présentons deux techniques ou méthodes segment-temporelles : celle de la mémoire circulaire, la plus ancienne (dérivant d'un système analogique des années 50), et celle de l'ajout-superposition (*OLA, overlap-add*) temporel synchrone à la hauteur (TD-PSOLA).

Technique de la mémoire circulaire

L'origine analogique La technique de la mémoire circulaire est la plus simple et la plus ancienne des techniques de dilatation/contraction temporelle et de transposition fréquentielle. Il s'agit d'une méthode fonctionnant dans le domaine temporel. Cette technique dérive d'un système analogique proposé dans les années 50 par Fairbanks [Fairbansk *et al.*, 1954]. Elle consiste à utiliser un magnétophone muni d'une tête rotative. La bande en boucle fermée s'enroule sur la moitié du cylindre (comme pour les magnétoscopes ou les DAT) et défile à vitesse constante. Le cylindre est muni de deux têtes de lectures diamétralement opposées dont les signaux sont mélangés avec un gain identique. Il est possible de contrôler le sens de la rotation et la vitesse du cylindre.

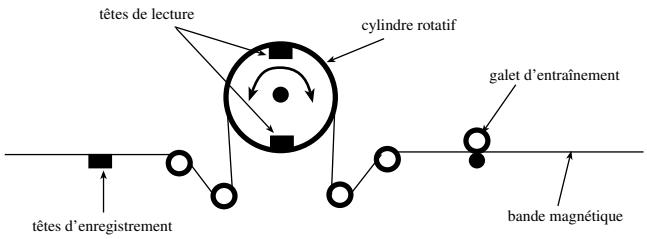


FIG. 2.16 – *Système analogique de la mémoire circulaire avec tête rotative.*

Lorsque le cylindre est immobile, la bande défile de façon identique devant la tête d'enregistrement et devant l'une des têtes de lecture. Le signal lu est donc identique au signal enregistré, aux erreurs d'enregistrement près. Lorsque le cylindre tourne en sens inverse du défilement de la bande, la vitesse relative V_r de défilement de la bande par rapport à la tête de lecture est supérieure à sa vitesse de défilement absolue V_a. Pendant la durée du contact entre la tête de lecture et la bande, le signal est donc lu plus rapidement qu'il n'a été enregistré, ce qui correspond à une dilatation de l'axe des fréquences (transposition vers le haut). La présence des deux têtes assure la continuité grâce à un fondu-enchaîné naturel (lorsqu'une tête quitte la bande, l'autre s'en rapproche, de sorte que l'amplitude du signal total ne diminue pas). On remarque que certaines portions du signal peuvent être lues deux ou plusieurs fois, en fonction de la vitesse de rotation de la tête. C'est cette relecture qui permet de conserver la durée du signal original. A l'inverse, lorsque le cylindre tourne dans le sens de défilement de la bande, le contenu en fréquence du signal est contracté vers l'origine (transposition vers le bas) puisque la bande est lue à une vitesse moindre qu'elle n'est enregistrée. Dans ce cas, certaines portions du signal peuvent ne pas être lues du tout.

Le rapport de l'homothétie en fréquences s'exprime par :

$$\alpha = \frac{V_r}{V_a} = 1 + \frac{R\,\Omega_{cyl}}{V_a} \tag{2.28}$$

avec V_a la vitesse de défilement de la bande devant la tête d'enregistrement, V_r la vitesse relative de la bande par rapport à la tête de lecture, Ω_{cyl} la vitesse de rotation du cylindre en $rad\ s^{-1}$ et R le rayon du cylindre. Dans tous les cas, l'alternance régulière des deux têtes se traduit par un "bruit" périodique de fréquence $f_{bruit} = \Omega_{cyl}/\pi$.

Les dilatations/contractions temporelles du signal sont obtenues par exemple en enregistrant le signal une première fois sur la bande, puis en la rejouant avec une vitesse de défilement de bande multipliée par un facteur α. En l'absence de rotation de la tête de lecture, le hauteur du signal est bien sûr multipliée par le facteur α, ce que l'on cherche à éviter. On compense donc le changement de hauteur par une rotation appropriée de la tête de lecture.

Mise en œuvre numérique La plupart des transposeurs disponibles dans le commerce sont basés sur une réalisation numérique du système décrit ci-dessus. La bande magnétique est remplacée par une mémoire circulaire dans laquelle on place les échantillons du signal en entrée. Cette mémoire circulaire est lue par deux pointeurs diamétralement opposés [Lee, 1972; Laroche, 1998].

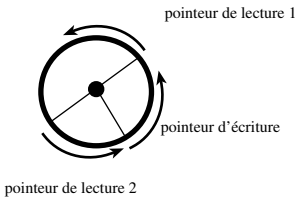


FIG. 2.17 – *Système numérique de mémoire circulaire.*

La mémoire circulaire correspond à une ligne à retard. A chaque échantillon écrit dans la mémoire (toutes les Δt secondes), on avance les pointeurs de lecture de $\alpha\Delta t$ secondes, où α est le taux de modification, puis on lit un échantillon dans la mémoire. Pour des valeurs de α non entière, on se retrouve dans la problématique des lignes à retard fractionnaires 2.2.4. Ainsi, le signal est lu avec une fréquence d'échantillonnage différente de celle à laquelle il a été écrit, ce qui provoque une modification de l'échelle des fréquences (transposition) de taux α. Un problème se pose quand le pointeur de lecture rattrape (lorsque $\alpha > 1$) ou est rattrapé par (lorsque $\alpha < 1$) le pointeur d'écriture. Comme dans l'équivalent analogique, la continuité est assurée par un mixage des deux pointeurs au moment où se produit la rencontre (fondu-enchaîné) : l'échantillon lu par le pointeur de lecture courant subit une pondération décroissante tandis que celui lu par l'autre pointeur subit

une pondération croissante. Finalement, le second pointeur devient le pointeur courant et garde sa pondération maximale jusqu'à ce que le pointeur d'écriture s'en rapproche.

Mis en œuvre de cette manière, le changement de hauteur par mémoire circulaire a un comportement équivalent à celui de son homologue analogique, à ceci près qu'il est plus facilement paramétrable : taille de la mémoire ou de la tête rotative, fonction de pondération, etc. Sa mise en œuvre en temps-réel ne pose pas de problème particulier, puisqu'il réclame très peu de calculs. Il produit malheureusement un bruit artificiel qui provient du mixage périodique de deux pointeurs de lecture. Pour tenter d'améliorer la qualité obtenue, on cherche à mieux raccorder les signaux lus par les pointeurs de lecture, par exemple en utilisant la fonction d'autocorrélation du signal pour déterminer l'endroit optimal pour le fondu-enchaîné. C'est le principe de la méthode temporelle synchrone à la hauteur, encore appelée méthode SOLA (*Synchronous OverLap Add*) [Roucos and Wilgus, 1985; Makhoul and El-Jaroudi, 1986].

Méthode temporelle "pitch-synchrone" TD-PSOLA

La méthode temporelle "pitch-synchrone" (ou *TD-PSOLA, Temporal Domain Pitch Synchronous OverLap Add*) suppose que l'on traite un signal de parole dont on connaît la période (déterminée par exemple par le premier pic de la fonction d'autocorrélation du signal) [Moulines and Charpentier, 1990]. L'idée est fondée sur l'hypothèse que le signal de parole est constitué d'impulsions glottales filtrées par le conduit vocal. On observe une succession de réponses impulsionnelles, positionnées en des temps multiples de la période (hypothèse du peigne temporel convolué avec une réponse impulsionnelle de conduit vocal). On définit d'abord des marques d'analyse synchrones de la fréquence fondamentale pour les parties voisées, positionnées sur la forme d'onde à chaque période. On peut alors effectuer des dilatations/contractions temporelles et des transpositions fréquentielles. Cette technique est une amélioration de la méthode SOLA.

Dilatation/contraction temporelle Pour modifier la durée d'un signal sans en altérer la fréquence fondamentale, on va simplement dupliquer (étirement temporel) ou éliminer (compression temporelle) des périodes de la forme d'onde, en fonction du taux de modification désiré. On est donc conduit à définir des marques de synthèse également synchrones à la hauteur, associées aux marques d'analyse (de façon non-bijective puisque certaines marques sont dupliquées ou éliminées).

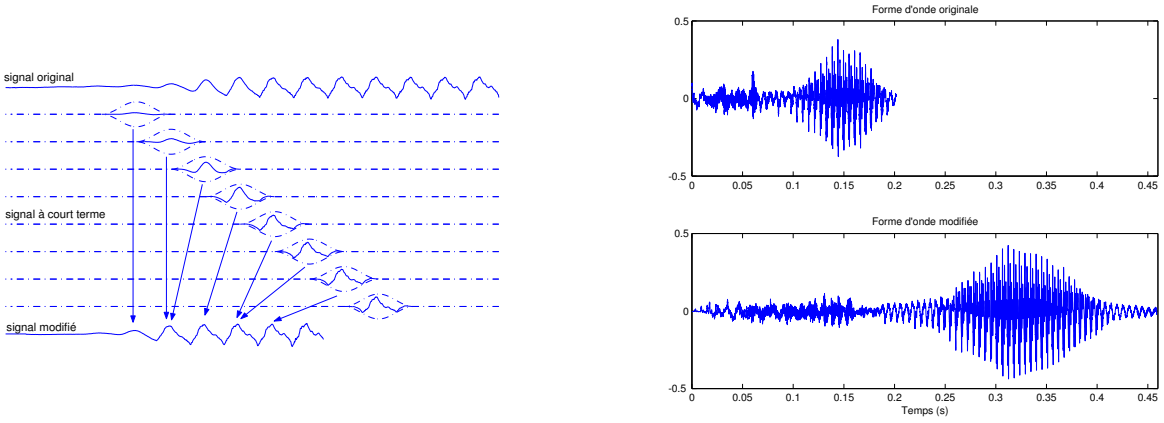


FIG. 2.18 – *Dilatation/contraction temporelle par TD-PSOLA, à gauche, principe pour un facteur 2/3, à droite pour un facteur 2, pour la syllabe "si".*

Les signaux à court-terme situés autour de chaque marque d'analyse sont alors extraits (par l'utilisation d'une fenêtre temporelle, par exemple de Hanning, de durée égale à deux périodes centrée sur la marque d'analyse) et recopiés autour des marques de synthèse correspondantes ; le signal modifié est obtenu par une simple méthode de superposition-ajout. La figure *fig.* 2.18 gauche

illustre le principe de cette méthode pour un taux d'étirement local de 2/3 (une contraction). On voit que deux périodes du signal original ont donné naissance à une période du signal de synthèse sur la figure de gauche, ce qui correspond bien à un étirement temporel, mais la durée de la période n'est pas modifiée (l'écartement des marques de synthèse est le même que celui des marques d'analyse), la fréquence fondamentale du signal est conservée. La figure *fig.* 2.18 droite donne un exemple d'application à la syllabe "si". on remarque la partie non-voisée à gauche de chaque forme d'onde (le son [s]), suivie de la partie voisée [i].

Transposition fréquentielle Si l'on est capable de positionner dans le signal les marques d'analyse exactement sur le début de chaque onde glottale (réponse impulsionnelle du conduit vocal se produisant à chaque fermeture glottale), on conçoit que diminuer (respectivement augmenter) l'intervalle de temps séparant deux marques d'analyse consécutives va permettre d'augmenter (respectivement de diminuer) la fréquence du fondamental, sans que les formants soient modifiés (la réponse impulsionnelle n'est pas modifiée, en particulier sa décroissance temporelle et ses fréquences de résonances, ou formants).

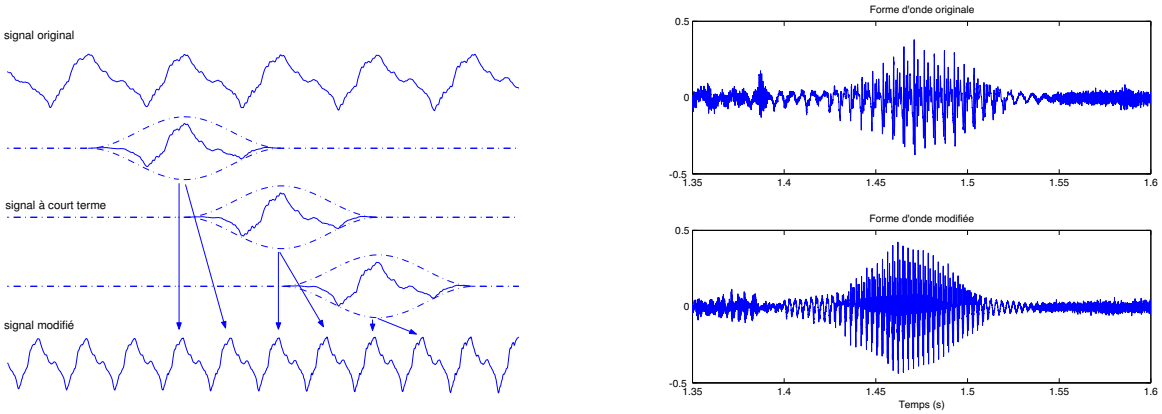


FIG. 2.19 – *Transposition fréquentielle par TD-PSOLA pour un facteur 2 : principe (gauche) et exemple (droite).*

On est ainsi conduit à définir des marques de synthèse correspondant à la valeur modifiée du fondamental, et à les associer aux marques d'analyse comme précédemment. Puisque les marques de synthèse sont plus serrées (élévation du fondamental) ou écartées (abaissement du fondamental) que dans le signal original, il faut dupliquer ou éliminer certaines marques si l'on veut conserver la durée du signal. La figure *fig.* 2.19 illustre le principe de cette méthode. On constate que les marques de synthèse étant plus resserrées que les marques d'analyse, la période du signal est raccourcie : il est nécessaire de dupliquer périodiquement certains signaux à court-terme pour conserver la durée du son. Lorsque le signal ne possède plus de fréquence fondamentale bien précise (cas des bruits et de certaines consonnes), la modification est réalisée de façon non synchrone, jusqu'à ce que l'on retrouve une région où le fondamental est présent.

La méthode décrite ci-dessus est appliquée principalement à la parole, et réalise des modifications de très bonne qualité. Par sa simplicité, elle peut faire l'objet d'une mise en œuvre temps-réel. En revanche, son application à des sons plus complexes (polyphoniques) ou dénués de hauteur (bruits) pose de sérieux problèmes. Les modifications du fondamental sont très sensibles à la position des marques d'analyse. Pour rendre la méthode plus robuste, les modifications de l'échelle fréquentielle peuvent être réalisées dans le domaine des fréquences (*FD-PSOLA, Frequential Domain PSOLA*).

2.4.2 Transformée de Fourier à court-terme et vocodeur de phase

Transformée de Fourier à court-terme La Transformée de Fourier à Court-Terme (TFCT), ou représentation de Fourier à court-terme d'un signal discret $x(n)$, est définie par :

$$\begin{aligned} X(n,k) &= \sum_{m=-\infty}^{+\infty} x(m)\, h_a(n-m)\, W_N^{mk} \text{ avec } W_N = e^{-2j\pi/N} \text{ k=0,...,N-1} \\ &= |X(n,k)|e^{j\phi(n,k)} = X_R(n,k) + jX_I(n,k) \end{aligned}$$

C'est un nombre complexe représentant la magnitude $|X(n,k)|$ et la phase $\phi(n,k)$ d'un spectre variant dans le temps, avec $k = 0, \ldots, N-1$ l'index du panier de fréquence (ou *bin* de fréquence) et n l'index de temps. Pour chaque valeur de n, le signal $x(m)$ est pondéré par une fenêtre d'analyse de longueur finie M : $h_a(n-m)$.

Transformée de Fourier à court-terme par bloc (vocodeur de phase) Le vocodeur de phase a été développé à l'origine dans le but d'effectuer du codage compressif de signaux sonores [Portnoff, 1976]. Du fait que cette technique utilise au moins deux fois plus de données que le signal original, les applications de compression ont été abandonnées. Cependant, de nombreuses applications musicales sont possibles, comme nous le verrons. Partons de la TFCT. Si l'on écrit maintenant le traitement effectué par bloc, on obtient :

$$\begin{aligned} X(sR_a,k) &= \sum_{m=-\infty}^{+\infty} x(m)\, h_a(sR_a-m)\, W_N^{mk} \\ &= W_N^{sR_ak} \sum_{m=-\infty}^{+\infty} x(m)\, h_a(sR_a-m)\, W_N^{-(sR_a-m)k} \\ &= W_N^{sR_ak} \tilde{X}(sR_a,k) \end{aligned}$$

La TFCT est échantillonnée tous les R_a échantillons en temps ; s est l'index de temps de la TFCT au taux d'échantillonnage de décimation. L'index de temps est $n = sR_a$ avec R_a le pas d'analyse ; $h_a(n)$ est la fenêtre d'analyse.

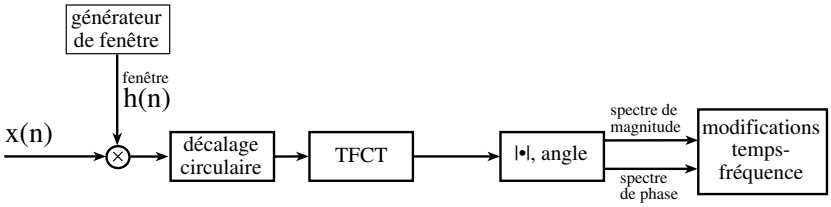


FIG. 2.20 – *Diagramme de l'analyse par vocodeur de phase.*

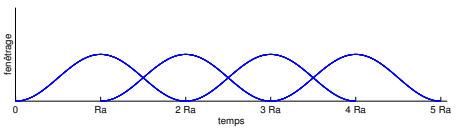


FIG. 2.21 – *Fenêtrage lors de l'analyse.*

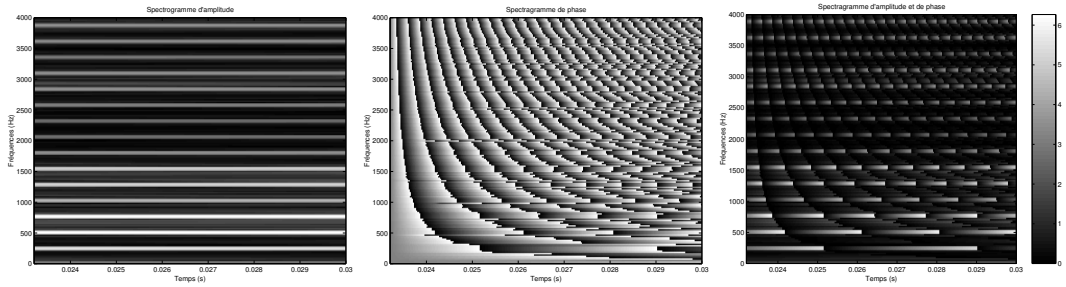


FIG. 2.22 – *Spectrogramme d'amplitude (sonagramme, à gauche) d'un signal harmonique, avec un pas $R_A = 1$, $L_w = 2048$; spectrogramme de phase (phasogramme, au milieu) et combinaison des deux (sonaphasogramme, à droite).*

Il est important d'avoir l'origine des temps de la TF au centre de la fenêtre d'analyse à la fois à l'analyse et à la synthèse [Serra and Smith, 1990]. De cette manière, un signal impulsionnel centré sur la fenêtre temporelle aura ses phases à zéro. Ceci est obtenu par décalage circulaire d'une demi-fenêtre avant application de la TF. Le décalage circulaire d'une demi-fenêtre consiste à prendre la moitié de droite et à l'inverser avec la moitié de gauche. Il est donc identique dans les deux sens. Notons $\hat{x}(n) = x(n - N/2)$ le signal après décalage circulaire. Sa transformée de Fourier discrète est donc $\hat{X}\left(e^{j\Omega}\right) = e^{-j\Omega\frac{N}{2}} X\left(e^{j\Omega}\right)$. En exprimant la phase $\Omega_k = \frac{2\pi}{N}\, k$, on obtient $\hat{X}(k) = e^{-j\frac{2\pi}{N}k\frac{N}{2}} X(k) = (-1)^k X(n)$. Il suffit donc de multiplier chaque valeur de la transformée de Fourier $X(k)$ par $(-1)^k$ pour effectuer le décalage circulaire et mettre les phases à zéros au centre du grain.

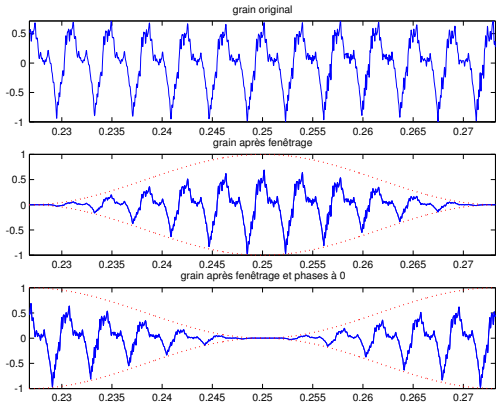


FIG. 2.23 – *Décalage circulaire d'une demi-fenêtre avant application de la TF.*

Après modification temps-fréquence (sur les modules et phases), la transformée de Fourier $X\left(sR_a, k\right)$ est modifiée en $Y\left(sR_s, k\right)$, avec R_s le pas de synthèse. Le signal temporel est obtenu par convolution de la transformée de Fourier à court-terme inverse avec la fenêtre de synthèse :

$$y(n) = \sum_{s=-\infty}^{+\infty} y_s(n - sR_s)\, h_s(n - sR_s) \tag{2.29}$$

$$y_s(n - sR_s) = \frac{1}{N} \sum_{k=0}^{N-1} \left(W_N^{-sR_s k}\, Y(sR_s, k) \right)\, W_N^{-nk} \tag{2.30}$$

Tant que les modifications en module et en phase sont linéaires, le fenêtrage lors de la synthèse n'est pas nécessaire. Cependant, pour toute transformation non linéaire, le fenêtrage devient néces-

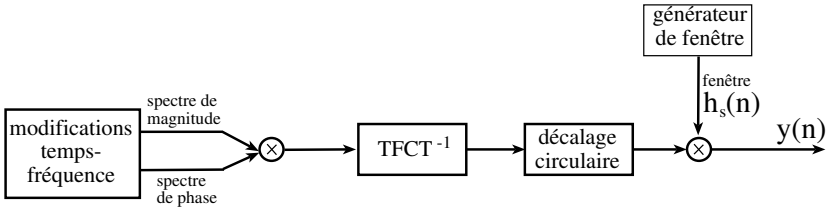


FIG. 2.24 – *Diagramme de la synthèse par vocodeur de phase.*

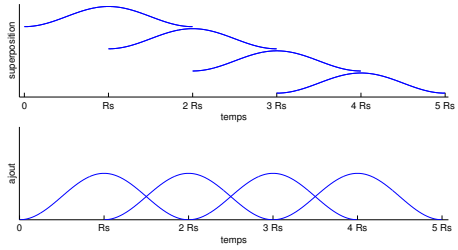


FIG. 2.25 – *Ajout et superposition des fenêtres lors de la resynthèse.*

saire, les grains de synthèse n'ayant plus nécessairement une enveloppe permettant la superposition sans problème de normalisation.

Déroulement de la phase Lors du processus de modification des modules et des phases de la TFCT, il est nécessaire de calculer le déroulement de la phase entre deux fenêtres de synthèse successives (et donc entre deux fenêtres d'analyse), de manière à bien interpoler les phases et calculer les fréquences instantanées de chaque panier de fréquence de la TFCT. La phase déroulée $\tilde{\phi}(n,k)$ se calcule à partir de la phase $\phi(n,k)$ du grain d'analyse par la formule :

$$\tilde{\phi}(n,k) = \Omega_k n + \phi(n,k) = \frac{2\ \pi\ k}{N} n + \phi(n,k) \tag{2.31}$$

Nous avons besoin d'une fonction qui transforme toute phase en son argument principal dans l'intervalle $]-\pi;\pi]$, définie par :

$$y = \text{princarg}(2\pi m + \phi_x) = \phi_x \in]-\pi;\pi] \tag{2.32}$$

On peut exprimer cette fonction "princarg" d'après la fonction "modulo" :

$$y = \text{mod}(\phi_y + \pi, -2\pi) + \pi \tag{2.33}$$

Etant données $\tilde{\phi}\,(sR_a,k)$ et $\tilde{\phi}\,((s+1)R_a,k)$ les phases de deux trames successives, on a :

$$\tilde{\phi}\,((s+1)R_a,k) = \Omega_k R_a + \tilde{\phi}\,(sR_a,k) \tag{2.34}$$

dans le cas où le panier de fréquence numéro k a une sinusoïde (c'est-à-dire un pic) de fréquence Ω_k. La phase déroulée s'exprime alors :

$$\tilde{\phi}_u\,((s+1)R_a,k) = \tilde{\phi}_c\,((s+1)R_a,k) + \tilde{\phi}_d\,((s+1)R_a,k) \tag{2.35}$$

avec $\tilde{\phi}_c\,((s+1)R_a,k)$ la phase cible et $\tilde{\phi}_d\,((s+1)R_a,k)$ la déviation par rapport à cette cible. Cette déviation s'exprime comme la différence entre la phase au temps $(s+1)\ R_a$ et sa cible :

$$\tilde{\phi}_d\,((s+1)R_a,k) = \text{princarg}\left(\tilde{\phi}\,((s+1)R_a,k) - \tilde{\phi}_c\,((s+1)R_a,k)\right) \tag{2.36}$$

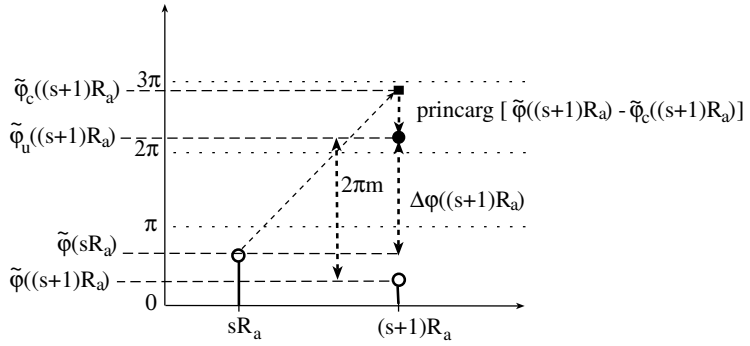


FIG. 2.26 – *Déroulement de phase (vocodeur de phase).*

Remplaçons maintenant $\tilde{\phi}_t\left((s+1)R_a, k\right)$ par sa valeur donnée en (2.34) :

$$\begin{aligned}\tilde{\phi}_u\left((s+1)R_a, k\right) &= \tilde{\phi}\left(sR_a, k\right) + \Omega_k R_a \\ &+ \text{princarg}\left(\tilde{\phi}\left((s+1)R_a, k\right) - \tilde{\phi}\left(sR_a, k\right) - \Omega_k n R_a\right) \end{aligned} \tag{2.37}$$

La différence de phases déroulées est alors :

$$\Delta\phi\left((s+1)R_a, k\right) = \tilde{\phi}_u\left((s+1)R_a, k\right) - \tilde{\phi}\left(sR_a, k\right) \tag{2.38}$$

$$= \Omega_k R_a + \text{princarg}\left(\tilde{\phi}\left((s+1)R_a, k\right) - \tilde{\phi}\left(sR_a, k\right) - \Omega_k n R_a\right) \tag{2.39}$$

On obtient la fréquence instantanée :

$$f_i\left((s+1)R_a, k\right) = \frac{1}{2\pi}\,\frac{\Delta\phi\left((s+1)R_a, k\right)}{R_a}\,F_e \tag{2.40}$$

Effets de phase et verrouillage de phase Un défaut du vocodeur de phase réside dans le fait que la phase de chaque panier de fréquence est supposé tourner à la fréquence centrale instantanée du panier. Ainsi, le pic dans une TCFT qui représente une seule sinusoïde est modélisé par plusieurs sinusoïdes de fréquences proches, ce qui donne au son ralenti un effet de réverbération artificielle, de glissement de phase. Une solution consiste à verrouiller les phases [Laroche and Dolson, 1997; Laroche and Dolson, 1999], c'est à dire à imposer que pour un pic fréquentiel, ses voisins aient une phase identique à la phase réelle de la sinusoïde. On peut aussi utiliser un vocodeur de phase pisteur, qui extraie les fréquences des partiels [Dolson, 1986] : c'est un prémisse de l'analyse additive, que nous allons présenter maintenant.

2.4.3 Analyse-synthèse additive

Le modèle additif se base sur l'observation suivante : les sons harmoniques peuvent être considérés comme la somme de sinusoïdes, dont l'amplitude et la fréquence évoluent dans le temps. De cette observation découle le modèle de McAulay et Quatieri [McAulay and Quatieri, 1986]. Si l'on écoute un signal analysé puis resynthétisé selon cette méthode, on remarque que soit le bruit caractéristique du signal a disparu (si on impose des rapports harmoniques entre les partiels extraits), soit le bruit est modélisé lui aussi par des sinusoïdes : la représentation n'est pas suffisamment complète. Dans ce but, on modélise le résidu comme un bruit filtré [Serra and Smith, 1990; Serra, 1996], selon le modèle source filtre, à l'aide de son enveloppe spectrale. Un raffinement possible consiste à différencier dans le résidu les transitoires d'attaque du bruit résiduel [Verma *et al.*, 1997]. La représentation additive du son est particulièrement adaptée à la voix, aux sons de cordes (frottées et pincées), de vents (colonnes d'air), de cuivres, et permet des transformations sur les paramètres de l'analyse.

Somme de sinusoïdes (McAulay, Quatieri, 1986)

Modèle Ce modèle est fondé sur l'analyse-synthèse de la voix humaine à l'aide d'une somme de sinusoïdes [McAulay and Quatieri, 1986]. Les sons quasi-stationnaires peuvent se modéliser comme une somme de partiels, sinusoïdes évoluant lentement dans le temps (en fréquence, phase et amplitude). Le modèle est donc le suivant :

$$x(n) = \sum_i A_i(n)\ \cos[\Theta_i(n)] \tag{2.41}$$

avec la phase :

$$\Theta_i(n) = \int_0^{nT} \omega_i(\tau)\ d\tau + \Phi_i \tag{2.42}$$

et avec $A_i(n)$ l'amplitude instantanée (enveloppe variant dans le temps), $\omega_i(n)$ la fréquence instantanée (piste fréquentielle de la i^{eme} sinusoïde, en radians) et $\Theta_i(n)$ la phase instantanée.

Analyse On applique une transformée de Fourier à court-terme. Les phases sont calculées à partir du spectre de phase, aux valeurs données par les pics (amplitude et fréquence) données dans le spectre de magnitude (*cf. fig.* 2.27).

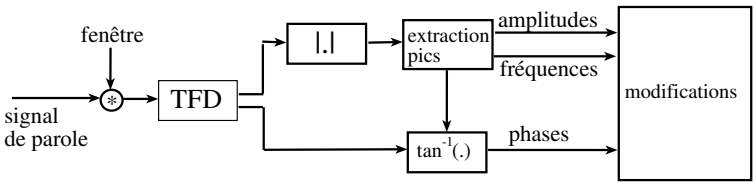


FIG. 2.27 – *Diagramme de l'analyse additive "somme de sinusoïdes".*

Synthèse La synthèse se fait elle aussi fenêtre par fenêtre. Les magnitudes sont interpolées linéairement, et servent d'enveloppes à des sinusoïdes générées après déroulement de phase et interpolation (linéaire ou non) entre deux phases successives dans le temps. Les sinusoïdes sont ensuite additionnées pour former le signal de synthèse (*cf. fig.* 2.28).

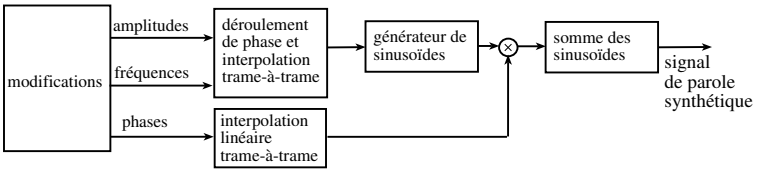


FIG. 2.28 – *Diagramme de la synthèse additive "somme de sinusoïdes".*

Modifications Les modifications peuvent avoir lieu sur tous les paramètres extraits par le modèle : enveloppe, trajectoires fréquentielles des partiels. C'est à cet endroit que nous pouvons intervenir.

Somme de sinusoïdes + résidu (Serra, Smith, 1989)

Un raffinement du modèle précédent a été proposé comme suit [Serra and Smith, 1990] : le signal est modélisé comme la somme de sinusoïdes (partie déterministe) et d'un bruit résiduel (partie stochastique). Ceci permet une meilleur représentation des sons non stationnaires. Le modèle s'appelle SMS, pour *Spectral Modeling Synthesis*.

Modèle Les sons peuvent se modéliser comme une somme de sinusoïdes évoluant lentement dans le temps (en fréquence, phase et amplitude) et d'un bruit filtré, dont l'enveloppe spectrale évolue lentement dans le temps. Le modèle est donc le suivant :

$$x(n) = \sum_{i}^{L} A_i(n) \ \cos[\Theta_i(n)] + e(n) \tag{2.43}$$

avec $A_i(n)$ l'amplitude instantanée, $\omega_i(n)$ la fréquence instantanée, $\Theta_i(n)$ la phase instantanée et $e(n)$ le bruit filtré à l'instant n. L'hypothèse de base du modèle est une variation lente en amplitude et en fréquence de la partie déterministe du signal.

Analyse L'analyse se fait par extraction des pics, calcul de la partie déterministe (somme de sinusoïdes) et soustraction au signal (temporelle, échantillon par échantillon), afin d'analyser ensuite la partie stochastique.

Dans chaque fenêtre, on applique une fenêtre de Blackman-Harris, dont la propriété est d'avoir un rapport lobe primaire sur lobes secondaires de 92 dB : la modélisation du premier lobe suffit donc, puisque les contributions des autres lobes est noyée dans le bruit de quantification du signal codé sur 16 bits. De plus, l'interpolation entre paniers de fréquence en décibels se fera par une parabole.

Pour éviter les effets de phase linéaire introduits par la fenêtre, chaque trame temporelle se voit appliquer un centrage autour de l'origine avant l'application de la transformée de Fourier. Après extraction des pics, la fréquence fondamentale est calculée (si elle existe) et sert à adapter la taille de la fenêtre d'analyse (analyse *pitch-synchrone*). Une procédure d'extraction de pics permet de donner les trajectoires des partiels (ie. attribuer à chacun des pics de la trame $n-1$ les bons successeurs, et à chacun des pics de la trame n les bons prédécesseurs) : il s'agit de l'algorithme de Maher et Beauchamp permettant d'estimer la fondamentale tout en effectuant le suivi des partiels [Maher and Beauchamp, 1994].

Le résidu est calculé par resynthèse de la partie déterministe et soustraction au signal original dans le domaine temporel. Après application de la transformée de Fourier à court-terme sur le résidu, on le modélise par interpolation (splines), par approximation (segments de droites), par codage par prédiction linéaire (LPC).

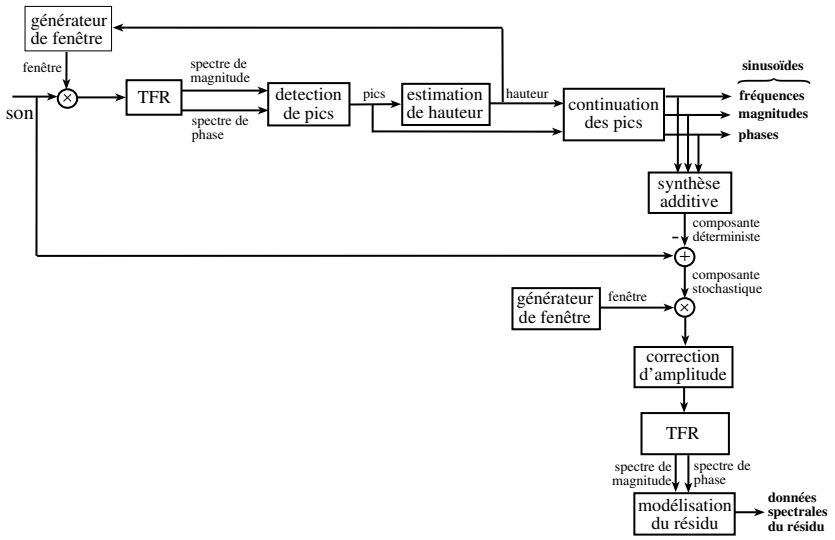


FIG. 2.29 – *Diagramme de l'analyse additive "somme de sinusoïdes + résidu".*

Synthèse La partie déterministe est générée par une synthèse additive sans information de phase, ce qui permet de synthétiser le signal dans le domaine temporel (synthèse de sinusoïdes) comme dans le domaine fréquentiel (basé sur le transformée de Fourier inverse). En terme de calculs, cette dernière méthode est la moins coûteuse (et donc la plus rapide). Notons qu'il est aussi possible de synthétiser la partie déterministe par TFCT inverse en tenant compte de l'information de phase.

La synthèse de la partie stochastique peut être vue comme un bruit blanc passant dans un filtre variant dans le temps, ce qui se met en œuvre généralement comme la convolution dans le domaine temporel d'un bruit blanc avec une réponse impulsionnelle de filtre. En pratique, on applique une FFT inverse à un spectre complexe, constitué de la forme spectrale du résidu et de phases aléatoires.

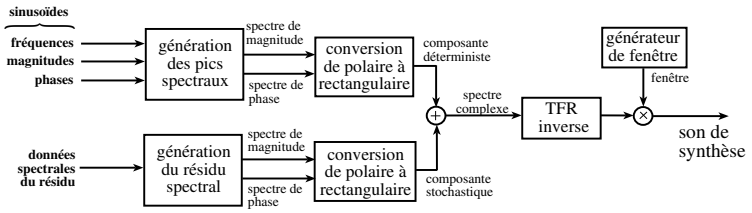


FIG. 2.30 – *Diagramme de la synthèse additive "somme de sinusoïdes + résidu".*

Les deux composantes peuvent s'ajouter dans le domaine fréquentiel (une seule FFT inverse à calculer) ou dans le domaine temporel.

Modifications Les modifications que l'on peut effectuer portent sur la partie stochastique (modifications d'enveloppe temporelle ou fréquentielle) et sur la partie déterministe (identiques à celles du modèle "somme de sinusoïdes").

Remarques Un grand intérêt de cette méthode est que l'analyse est rendue indépendante de la synthèse. Ainsi, les paramètres d'analyse (taille et type de fenêtre, taille de FFT, pas entre deux trames) peuvent être choisis de façon à optimiser l'analyse. Plusieurs mises en œuvre de ce modèle existent, dont **SMS** [Music Tech. Group, 2000] (renommé récemment CLAM [Music Tech. Group, 2002]) du Music Technology Group de Barcelone [Serra and Smith, 1990], **Additive** de l'Ircam, **SAS** développé au LaBRI par Sylvain Marchand [Desainte-Catherine and Marchand, 1999].

Somme de sinusoïdes + transitoire + résidu (Verma, Levine, Meng 1997)

Une extension intéressante du modèle précédent consiste à prendre en compte les transitoires dans la partie résiduelle et de les séparer de celle-ci, en les modélisant par un petit jeu de paramètres [Verma *et al.*, 1997]. Ainsi, chacune des trois composantes du signal peut faire l'objet de modifications distinctes, indépendantes.

Modèle

$$x(n) = \sum_{i}^{L} A_i(n)\ \cos[\Theta_i(n)] + r_t(n) + r_b(n) \tag{2.44}$$

avec $r_t(n)$ le résidu concernant la partie transitoire, et $r_b(n)$ la partie bruitée.

Analyse Une fois l'analyse de type SMS effectuée, le premier résidu obtenu est analysé. Un éventuel transitoire est détecté, paramétrisé, synthétisé et soustrait au résidu. Le second résidu obtenu est la composante stochastique de ce modèle. Cette composante peut alors être modélisée à son tour, selon la méthodologie expliquée pour le modèle "somme de sinusoïdes + bruit".

L'algorithme utilisé pour la modélisation du transitoire tient compte de la dualité entre sinusoïdes et transitoire. Une sinusoïde variant lentement dans le domaine temporel correspond à un

pic, une impulsion dans le domaine fréquentiel. C'est ainsi qu'on peut la détecter dans une TFCT, à l'aide du spectre de magnitude. Des signaux impulsifs dans le domaine temporel deviennent oscillants dans le domaine fréquentiel.

La première étape consiste donc à mettre en correspondance le domaine temporel des transitoires avec des signaux sinusoïdaux dans un certain domaine fréquentiel. La transformée en cosinus discret (TCD, *DCT : discrete cosine transform)* permet ce type de correspondance.

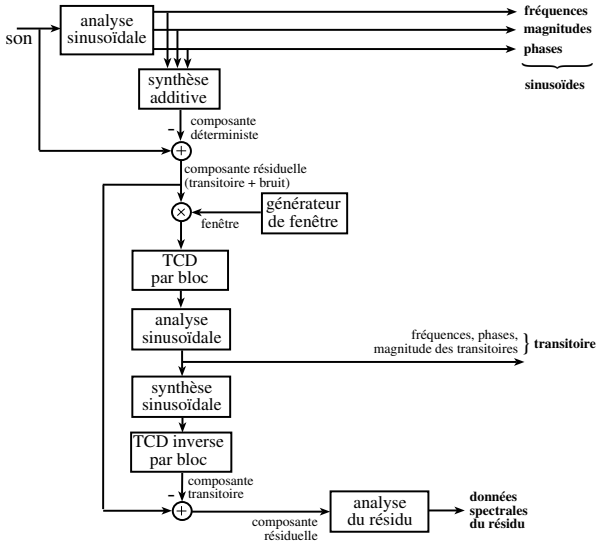


FIG. 2.31 – *Diagramme de l'analyse additive "somme de sinusoïdes + transitoire + résidu".*

La transformée en cosinus discret est définie de la manière suivante :

$$C(k) = \beta(k) \sum_{n=0}^{N-1} x(n) \cos\left[\frac{(2n+1)\,\pi}{2\,N}\right], \quad n, k \in \{0, 1, \ldots, N-1\} \tag{2.45}$$

avec $\beta(k) = \sqrt{1/N}$ si $k = 1$ et $\beta(k) = \sqrt{2/N}$ sinon. Une impulsion au début d'une trame temporelle correspondra à la présence d'un cosinus basse fréquence, alors qu'une impulsion à la fin d'une trame temporelle correspondra à la présence d'un cosinus haute fréquence. Inversement, lorsque l'on modifie la fréquence de la sinusoïde correspondant à un transitoire, on le déplacera dans le temps. L'algorithme consiste à prendre des blocs qui ne se superposent pas, calculer une TCD sur chaque bloc, extraire les paramètres de la transitoire en appliquant un modèle sinusoïdal sur chaque trame DCT (ie. sur chaque bloc).

Synthèse La synthèse se fait par reconstruction des sinusoïdes du domaine DCT et l'utilisation de la transformée inverse (TCDI). Le transitoire synthétisé est ajouté à la partie déterministe et à la partie stochastique de synthèse.

Modifications Elles peuvent ne s'appliquer que sur l'une des composantes, ou plusieurs à la fois. Attention, lors du changement d'échelle temporelle, de bien appliquer le même facteur aux trois composantes, afin de s'assurer que la synchronisation du transitoire avec le reste ait bien lieu !

2.4.4 Analyse-synthèse soustractive, ou source-filtre

En partant d'un modèle de la voix humaine constitué d'une source riche en harmoniques, filtrée dans plusieurs cavités successives, on obtient un modèle de synthèse soustractive. C'est un modèle très performant pour décrire la voix humaine : en effet, les cordes vocales produisent un train d'ondes

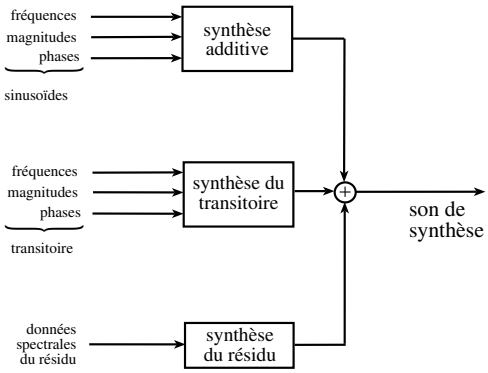


FIG. 2.32 – *Diagramme de la synthèse additive "somme de sinusoïdes + transitoire + résidu".*

riche en harmoniques, et le larynx, la fosse nasale et la cavité buccale sont des cavités filtrant le signal, responsables de l'apparition des formants dans le spectre d'un signal. La reconnaissance de la parole (reconnaissance naturelle faite par chacun d'entre nous, ou automatique à l'aide de systèmes informatiques), et particulièrement des voyelles, se fait grâce à ces formants. De ce fait, la reconnaissance d'autres sons (les sons instrumentaux, notamment) se fait aussi, en partie, à l'aide des formants. Le modèle soustractif ou source-filtre a alors été développé et utilisé pour la synthèse comme pour la modification de sons. Ainsi, en changeant la source ou les filtres, on modifie le timbre d'un son.

La grande difficulté réside en l'extraction de l'enveloppe spectrale. Pour les sons harmoniques, il s'agit de la courbe passant par les pics spectraux des sinusoïdes présentes dans le signal. Dans ce cas, se pose le problème de connaître précisément l'emplacement de ces pics (l'utilisation de la méthode additive répond à cette question), et de savoir quelle interpolation est la plus censée dès lors que l'on veut connaître la valeur de l'enveloppe spectrale entre les pics. Dans le cas de sons inharmoniques ou bruités, cette définition de l'enveloppe spectrale n'est plus valide, et la notion d'enveloppe spectrale est totalement dépendante de ce que l'on considère comme source, excitation et de ce que l'on considère comme filtre (ou résonance) dans le signal bruité ou inharmonique.

Une approximation de la notion d'enveloppe spectrale consiste à dire que c'est un lissage du spectre d'amplitude, qui tend à laisser de côté les lignes spectrales tout en conservant la forme générale du spectre.

Nous allons maintenant présenter deux principales manières d'estimer l'enveloppe spectrale : la prédiction linéaire et le cepstre.

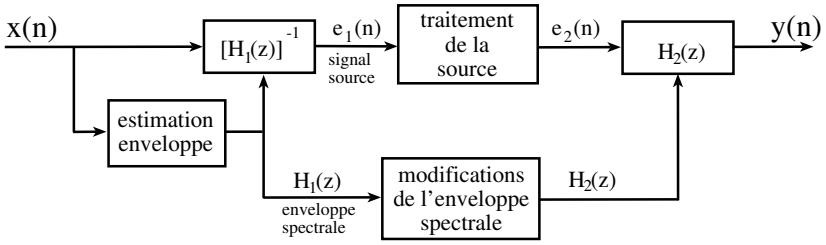


FIG. 2.33 – *Diagramme de l'analyse-synthèse soustractive.*

Prédiction linéaire (LPC)

La LPC est une technique d'analyse utilisable dans le cadre du modèle source-filtre. Le filtre utilisé est tout-pôles (il ne contient que des pôles) et représente l'enveloppe spectrale. Cette technique a été développée pour la voix, en considérant à juste titre que l'excitation (ou la source) des

sons voisés correspondent à un train d'ondes de période la fréquence fondamentale, tandis que la source des sons non voisés est un bruit [Markel and Gray, 1976; Makhoul, 1977]. Dans les deux cas, l'excitation est filtrée par les formants. Le **codage à prédiction linéaire** (ou *LPC, linear prediction coding*) modélise le filtre par un filtre linéaire. Ainsi, pour un signal sonore $x(n)$, le modèle utilisé est :

$$\hat{x}(n) = \sum_{k=1}^{p} a_k \, x(n-k) \tag{2.46}$$

avec p l'ordre du filtre, ou ordre de prédiction, et a_k les coefficients du filtre, ou coefficients de prédiction. La qualité de cette prédiction est évaluée grâce à l'erreur de prédiction, ou résidu :

$$e(n) = x(n) - \hat{x}(n) \tag{2.47}$$

La transformée en z du filtre de prédiction est donnée par :

$$P(z) = \sum_{k=1}^{p} a_k \, z^{-k} \tag{2.48}$$

et l'erreur de prédiction dans le domaine de la transformée en z :

$$E(z) = X(z) - \hat{X}(z) = X(z) \; (1 - P(z)) \tag{2.49}$$

FIG. 2.34 – *Diagramme de l'analyse et de la synthèse soustractive par LPC.*

Comme on peut le voir en figure *fig.* 2.34, l'analyse LPC correspond à une prédiction "*feed-forward*". Le filtre inverse, ou filtre d'erreur de prédiction est $A(z) = 1 - P(z)$, et l'erreur s'exprime alors :

$$E(z) = X(z) \; A(z) \tag{2.50}$$

Le son est obtenu en bout de chaîne de traitement à l'aide du signal d'excitation (éventuellement modifié) $\tilde{x}(n)$ en entrée du filtre tout pôle $H(z) = \frac{1}{A(z)} = \frac{1}{1-P(z)}$, d'où :

$$Y(z) = \tilde{E}(z) \; H(z) \tag{2.51}$$

avec $H(z)$ obtenu par une boucle de réinjection, *cf. fig.* 2.34 à droite. Le filtre RII (à réponse impulsionnelle infinie) $H(z)$ est appelé filtre LPC ou filtre de synthèse, et représente le modèle spectral (à un facteur gain près) de $x(n)$.

Si le filtre est optimal, l'erreur résiduelle est minimisée. Cette méthode est utilisée en télécommunications, où le résidu quantifié $\tilde{e}(n) = Q(e(n))$ est utilisé comme signal d'excitation du filtre LPC.

Calcul des coefficients Il existe différentes méthodes de calcul des coefficients du filtre LPC ; la plupart des mises en œuvre les calculent à partir d'un grain du signal $x(n)$, et remettent à jour le filtre régulièrement. C'est le cas des méthodes de covariance (filtres non stables), d'autocorrélation et l'algorithme de Burg. Les trois méthodes ont une interprétation en terme de filtres en treillis, mais seule la méthode de Burg utilise cette structure pour effectuer l'analyse.

Nous présentons ici la méthode d'autocorrélation. L'énergie de l'erreur de prédiction est donnée par l'espérance mathématique de l'erreur : $E_p = E\{e^2(n)\}$. Le filtre est optimal lorsque l'énergie de l'erreur de prédiction est minimale, donc lorsque sa dérivée est nulle. Sa dérivée est donnée par :

$$\frac{\partial E_p}{\partial a_i} = 2\ E\left\{e(n).\frac{\partial e(n)}{\partial a_i}\right\} = -2E\left\{e(n)\ x(n-i)\right\} \tag{2.52}$$

La condition d'optimalité $\frac{\partial E_p}{\partial a_i} = 0$ s'écrit :

$$-\ 2\ E\left\{\left(x(n) - \sum_{k=1}^{p} a_k\ x(n-k)\right)\ x(n-i)\right\} \quad = \quad 0 \tag{2.53}$$

D'où la formulation des équations normales :

$$\sum_{k=1}^{p} a_k\ E\left\{x(n-k)\ x(n-i)\right\} = E\left\{x(n)\ x(n-i)\right\} \tag{2.54}$$

L'autocorrélation temporelle d'une séquence de N valeurs de $u(n)$ est définie par :

$$r_{xx}(i) = \sum_{n=i}^{N-1} u(n)\ u(n-i) \tag{2.55}$$

avec $u(n) = x(n)\ w(n)$ le fenêtrage de x par la fenêtre $w(n), n = 0, \ldots, N-1$ (dans notre cas, on utilise la fenêtre de Hamming). On peut remplacer les espérances par leurs approximations en utilisant l'autocorrélation, d'où la nouvelle formulation des équations normales :

$$\sum_{k=1}^{p} a_k r_{xx}(i-k) = r_{xx}(i) \qquad i = 1, \ldots, p \tag{2.56}$$

Les coefficients a_k sont obtenus par résolution de ces équations normales, à l'aide de l'algorithme récursif de Levinson-Durbin. Minimiser l'énergie du résidu revient à trouver la meilleur correspondance spectrale dans le domaine fréquentiel, à un facteur gain près. Dans ce cas, le modèle d'entrée $x(n)$ est modélisé par le filtre :

$$H_g(z) = G.H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{2.57}$$

avec G le facteur gain. Avec ce filtre de synthèse modifié, le signal original est modélisé en utilisant comme signal d'excitation un bruit blanc de variance unité. Pour la méthode d'autocorrélation, le facteur de gain est défini par :

$$G^2 = r_{xx}(o) - \sum_{k=1}^{p} a_k\ r_{xx}(k) \tag{2.58}$$

Remarquons que ce gain dépend de l'énergie de l'erreur de prédiction.

Cepstre

Inventée par Tuckey en 1964 puis utilisé avec Noll, la méthode du **cepstre**[1] est une méthode qui permet l'estimation de l'enveloppe spectrale à l'aide de la FFT d'une trame temporelle du signal fenêtrée [Noll, 1964]. Le cepstre complexe est défini comme la transformée de Fourier à court-terme inverse du logarithme de la TFCT $X(n)$ du signal :

$$c(n) = \mathcal{F}^{-1}\left(\log\left(X\right)\right) \tag{2.59}$$

[1] Cepstre : terme obtenu par inversion des 4 premières lettres du mot "spectre".

Le cepstre réel est défini comme la TFCT inverse du logarithme du module $|X|$ de la TFCT du signal :

$$c_R(n) = \mathcal{F}^{-1}\left(\log|X|\right) \tag{2.60}$$

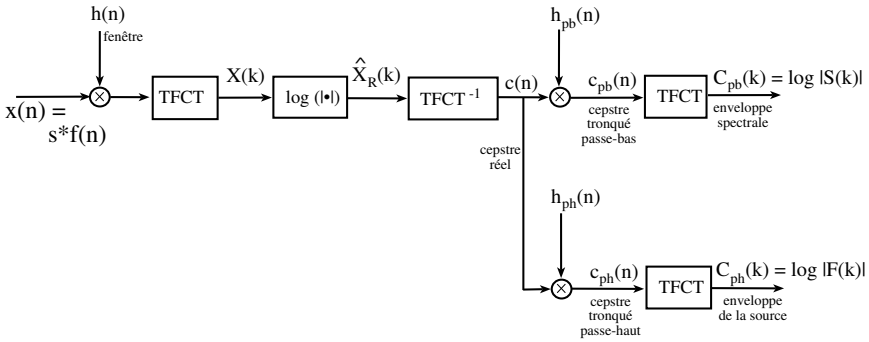


FIG. 2.35 – *Diagramme de l'analyse soustractive par cepstre.*

Lorsque l'on applique un filtre passe-bas au cepstre, on obtient un lissage du spectre (appelé **liftrage**[2]). Si le filtre est bien choisi, le lissage correspond à une bonne approximation de l'enveloppe spectrale. Soit $h_{pb}(n)$ la fenêtre utilisée pour le filtrage passe-bas, définie par :

$$h_{pb}(n) = \begin{cases} 1 & n = 0, N/2 \\ 2 & 1 \leq n < N/2 \\ 0 & N/2 < n \leq N-1 \end{cases} \tag{2.61}$$

Pour la définition de cette fenêtre, on utilise le fait que le cepstre réel est une fonction paire, donc sa TFCT inverse $c(n)$ est liée au cepstre complexe par la formule $c_R(n) = \frac{c(n)+c(-n)}{2}$. Le filtrage passe-bas $c_{pb}(n)$ du cepstre réel est obtenu par convolution de la fenêtre $h_{pb}(n)$ avec le cepstre réel $c(n)$. L'enveloppe spectrale est obtenue par TFCT de $c_{pb}(n)$:

$$\mathcal{S} = \exp\left(\mathcal{F}\left(c_{pb}\right)\right) \tag{2.62}$$

En calculant le cepstre réel en dB $c_R = \mathcal{F}^{-1}\left(20\ \log_{10}\left(|X|\right)\right)$, on obtient l'enveloppe spectrale en dB sans avoir à appliquer de fonction exponentielle : $\mathcal{S} = \mathcal{F}\left(c_{pb}\right)$. Si l'on considère le signal $x(n)$ constitué comme la convolution d'une source $s(n)$ et d'un filtre $f(n)$, sa transformée de Fourier est donnée comme le produit des transformées de Fourier de s et de f : $\mathcal{F}(x) = \mathcal{F}(s).\mathcal{F}(f)$, ce qui reste vrai pour les modules :

$$|\mathcal{F}(x)| = |\mathcal{F}(s)|.|\mathcal{F}(f)| \tag{2.63}$$

En échelle logarithmique, un produit devient une somme :

$$\log\left(|\mathcal{F}(x)|\right) = \log\left(|\mathcal{F}(s)|\right) + \log\left(|\mathcal{F}(f)|\right) \tag{2.64}$$

La séparation source-filtre s'obtient alors aisément avec la méthode du cepstre, par l'utilisation de deux fenêtres, l'une h_{pb} passe-bas pour extraire l'enveloppe spectrale, l'autre passe-haut h_{ph} pour extraire la source. Les valeurs du cepstre sont repérées par des temps ; aussi, les valeurs de temps basses, appelées quéfrences[3], correspondent à des variations lentes du spectre de magnitude (donc à l'enveloppe spectrale), et les hautes quéfrences correspondent aux variations rapides, aux pics du spectre de magnitude (soit à la source).

[2] Liftrage : terme obtenu par inversion des 3 premières lettres du mot "filtrage"
[3] Quéfrence : terme obtenu par inversion des deux premiers groupes de consonnes du mot "fréquence"

Seuillage Afin de trouver un bon seuil de quéfrence pour le calcul de la source et du filtre par la méthode du cepstre, on peut utiliser le fait que les quéfrences sont des variables temporelles, donc si le signal est harmonique, le cepstre sera périodique, la période correspondant à la fréquence fondamentale du signal harmonique[Noll, 1967]. Dans ce cas, cette valeur est la quéfrence maximale définissant l'enveloppe. Une valeur inférieure lissera l'enveloppe, une valeur supérieure inclura une partie des pics des harmoniques (ou des partiels) dans l'enveloppe. Notons aussi que la troncature effectuée lors du filtrage passe-bas correspond à un filtrage dans le domaine fréquentiel par $\frac{\sin(f)}{f}$. Une alternative est d'utiliser une fenêtre dont la transition est plus lisse.

Utilisation du logarithme Elle donne des valeurs variant entre $-\infty$ et 0 dB. Pour éviter les problèmes de calculs dus à des valeurs infinies, la valeur de l'opérande est limitée par l'addition d'une valeur très faible (10^{-5}), afin d'obtenir -100 dB comme limite inférieure.

Itérations Une manière d'améliorer le calcul de l'enveloppe spectrale consiste à utiliser un algorithme itératif qui ne calcule que la différence positive entre le spectre instantané et l'enveloppe spectrale estimée à chaque itération.

Notons enfin que, même si l'on utilise très couramment le spectre réel, le spectre complexe peut lui aussi être utilisé pour réaliser l'estimation de l'enveloppe spectrale. Dans ce cas, l'enveloppe spectrale est définie par une TFCT complexe.

2.4.5 Remarque sur les ondelettes

L'analyse-synthèse par ondelettes a fait ses preuves pour l'analyse et la resynthèse de signaux sonores [Kronland-Martinet, 1988; Guillemain, 1994; Evangelista, 1997], ainsi que pour les applications de compression et décompression sonore et visuels. Elle permet des traitements sonores tels que la dilatation temporelle dans transposition ainsi que la transformation duale, la dispersion, le glissando, la modulation d'amplitude, la synthèse croisée. Cependant, la lecture des images temps-fréquence et temps-échelle obtenue est rendue difficile du fait de la bilinéarité de la fonction utilisée comme noyau reproduisant. Ce problème existe déjà pour le vocodeur de phase par exemple, où modifier arbitrairement le spectre d'amplitude ne signifie pas que l'on crée une image valide (on peut le vérifier en appliquant une transformée de Fourier inverse, puis à nouveau une transformée de Fourier, et comparer les deux images). La complexité est encore plus grande pour les ondelettes, car il existe des interférences croisées dans les images. Nous n'avons donc pas investigué dans cette direction.

2.5 Repliement du spectre

La notion de repliement du spectre (*foldover* ou *aliasing*) est essentielle dans le traitement du signal numérique sonore, du fait que le repliement du spectre s'entend. De même, le repliement du spectre en image se voit, par exemple lorsqu'une roue de voiture tourne plus vite que la fréquence maximale codée par le signal, et semble alors tourner très lentement, voire à l'envers.

Pour les sons, si l'échantillonnage se fait à F_e Hz (fréquence d'échantillonnage), on ne pourra pas coder des fréquences supérieures à $F_e/2$ (fréquence de Nyquist), car il faut au moins deux échantillons pour définir une sinusoïde (sa fréquence, son amplitude et sa phase à l'origine). Généralement, le problème de repliement se pose lorsque l'on veut numériser un signal ; il faut alors le filtrer passe-bas à la demi-fréquence d'échantillonnage avant de pratiquer l'échantillonnage. Le problème peut aussi se poser avec un signal déjà numérisé, par exemple quand on le ré-échantillonne, quand on modifie son spectre (par exemple en le décalant, en l'étirant) sans précaution. Nous aurons donc ceci bien à l'esprit lors de la modification d'algorithmes pour les rendre adaptatifs.

Chapitre 3

Effets et transformations selon la taxonomie perceptive

Je suis partagé entre mon goût pour les faits et mon goût pour l'effet.
Louis Scutenaire [Scutenaire, 1984]

Sommaire

3.1 Introduction

Nous présentons ici un ensemble assez complet, bien que non exhaustif, des effets et transformations audionumériques possibles, d'après de nombreuses lectures, parmi [Laroche, 1995; Mathews and Pierce, 1989; Orfanidis, 1996; Roads, 1998; Zoelzer, 1997; Zoelzer, 2002]. Certains peuvent s'obtenir par plusieurs procédés, pour lesquels des différences audibles de résultat existent. Nous nous limiterons à présenter les familles d'effets par paramètres perceptifs principalement modifiés (taxonomie perceptive), *cf. tab.* 3.1, ainsi qu'au moins une des méthodes d'obtention de chaque effet. Remarquons que la plupart des effets peuvent aujourd'hui se mettre en œuvre en temps-réel, sauf quelques uns (avec une particularité pour le brassage et la granulation, qui techniquement se font en temps-différé, mais peuvent être adaptés en temps-réel en ne traitant que le signal passé).

nom	**dynamique**	**durée**	**hauteur**	**espace**	**timbre**	**temps-réel**
amplification	×					oui
normalisation	×					non
expandeur, *noise gate*	×					oui
compresseur, limiteur	×					oui
trémolo	×					oui
dilatation/contraction		×				oui/non
inversion		×				non
transposition (sans formants)			×		(×)	oui
transposition (avec formants)			×			oui
discrétisation de hauteur			×			oui
harmoniseur			×			oui
écho				×		oui
panoramisation				×		oui
effet de précédence				×		non
réverbération				×		oui
distance				×		oui
effet Doppler				×		oui
Leslie/Rotary				×		oui
rendu 3D avec casque				×		oui
rendu 3D avec haut-parleurs				×		oui
directivité				×		oui
effets sur l'enveloppe						
modifications de l'enveloppe spectrale					×	oui
équaliseur					×	oui
filtre en peigne					×	oui
filtre résonant					×	oui
wha-wha					×	oui
filtre avec résolution arbitraire					×	oui
effets sur la phase						
flanger					×	oui
chorus					×	oui
phaser					×	oui
effets sur le spectre et sa structure						
décalage du spectre			(×)		×	oui
vibrato			(×)		×	oui
effets sur le spectre et l'enveloppe						
distorsion					×	oui
conformation spectrale					×	oui
synthèse croisée, *morphing*					×	oui
chuchotement					×	non
overdrive					×	oui
fuzz					×	oui
exciter					×	oui
enhancer					×	oui
débruitage					×	oui
décliquage					×	oui
robotisation			×		×	oui
ré-échantillonnage			×		×	non
brassage, granulation	×				×	oui/non
changement de genre	×		×		×	non

TAB. 3.1 – *Tableau des principaux effets audionumériques usuels, indiquant le ou les paramètres perceptifs modifiés par l'effet, et la possibilité de les mettre en œuvre en temps-réel ou non.*

3.2 Traitements modifiant la dynamique

Les effets modifiant la dynamique peuvent être linéaires (amplification, normalisation), non linéaires. Certains peuvent être considérés comme les premiers effets adaptatifs : ils effectuent dans un premier temps une mesure du niveau d'énergie du signal d'entrée, et modifient ensuite le gain appliqué à ce signal (expandeur, *noise gate*, compresseur, limiteur).

3.2.1 Amplification

L'amplification d'un signal consiste à augmenter ou diminuer l'amplitude des échantillons $x(n)$, le gain g_{ampli} (en décibel) pouvant varier dans le temps :

$$y(n) = 10^{g_{ampli}/20} x(n) \tag{3.1}$$

Ceci constitue une amplification de base. En effet, les amplificateurs haute-fidélité utilisent différents étages de traitement, entre autres pour apporter des corrections au spectre (équalisation). Ceci n'est pas abordé dans cette partie.

3.2.2 Normalisation

La normalisation consiste à appliquer un gain à un signal de sorte que son niveau maximal soit de 0 dB. Elle se réalise hors temps-réel. On calcule le maximum de la valeur absolue du signal, et l'on divise la valeur de chaque échantillon par ce maximum. Soit $x(n)$ le signal à normaliser, le maximum M est donné par :

$$M = \max_n |x(n)| \tag{3.2}$$

Le signal normalisé est donné par :

$$y(n) = \frac{x(n)}{M} \tag{3.3}$$

Cet effet sert à utiliser toute la plage de dynamique que le codage permet lors de l'écriture en fichier ou le stockage en mémoire.

3.2.3 Expandeur, *noise gate*

Définition L'expandeur est un traitement non linéaire qui augmente la dynamique d'un signal de façon à ce que les signaux de bas niveau sonore soient atténués sans modifier les signaux de haut niveau sonore. Le *noise gate* correspond à un expandeur poussé à l'extrême : l'entrée de faible niveau est atténuée fortement, voire éliminée et remplacée par du silence.

Fonctionnement L'expandeur est un amplificateur à gain variable (le gain ne dépassant jamais 1 en échelle linéaire), contrôlé par le niveau d'énergie du signal d'entrée et par une fonction de conformation non-linéaire (sans doute le premier effet adaptatif de l'histoire). Lorsque le niveau d'entrée est élevé, l'expandeur a un gain unitaire, tandis que lorsque le niveau est faible, le gain décroît, rendant le niveau du signal encore plus bas. La figure *fig.* 3.1 montre la structure d'un expandeur, qui est aussi celle d'un compresseur limiteur.

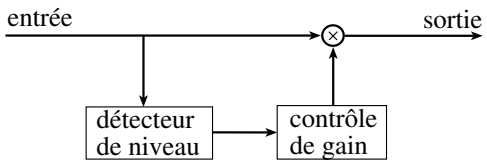


FIG. 3.1 – *Diagramme de l'expandeur et du compresseur/limiteur.*

Le niveau d'énergie correspond à une mesure pondérée du signal. Par exemple, le RMS (*cf.* 4.5.1) est un bon indicateur du niveau d'énergie d'un signal.
On peut représenter la relation entrée-sortie d'un expandeur à l'aide d'une fonction de conformation tel que *fig.* 3.2. Le niveau d'entrée est représenté en abscisse, le niveau de sortie en ordonnée. Le point de la courbe S_{exp} où la pente change est le **seuil** de l'expandeur ; il est ajustable par l'utilisateur. C'est la valeur du niveau d'entrée à partir de laquelle le contrôle de gain est effectif. Le rapport

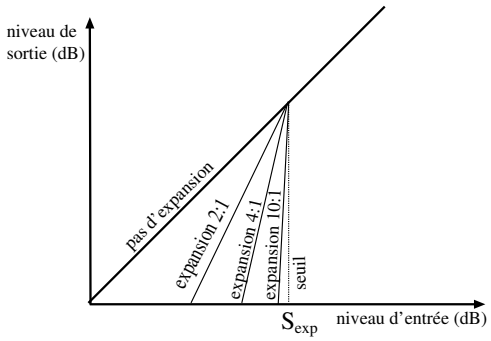


FIG. 3.2 – *Fonction de conformation de l'expandeur (relation entrée-sortie).*

d'expansion R_{exp} indiqué (par exemple 4 : 1) s'applique en Décibel (dB), puisque c'est l'unité de mesure de niveau sonore. Le *noise gate* est un expandeur de rapport d'atténuation de 10 : 1 ou plus.

$$N_{sortie} = \begin{cases} N_{entree} & \text{si } N_{sortie} \geq S_{exp} \\ R_{exp}N_{entree} + (1 - R_{exp})S_{exp} & \text{si } N_{sortie} \leq S_{exp} \end{cases} \tag{3.4}$$

La fonction d'évaluation du niveau sonore s'effectuant sur une moyenne du signal entrant, il existe un temps de retard entre un changement de niveau dans le signal d'entrée et le déclenchement d'un changement de gain. Le **temps d'attaque** (noté $\tau_{a,exp}$) est le temps nécessaire à l'expandeur pour restaurer à 1 le gain d'un signal lorsque le niveau d'entrée dépasse le seuil. De la même manière, le **temps de relaxation** $\tau_{r,exp}$ est le temps nécessaire à l'expandeur pour réduire le gain après que le signal est passé en dessous du seuil. Le comportement de l'expandeur est celui d'un système à hystérésis, puisque lors de la montée et de la descente, le point représentant le niveau de sortie en fonction du niveau d'entrée n'utilise pas le même chemin.

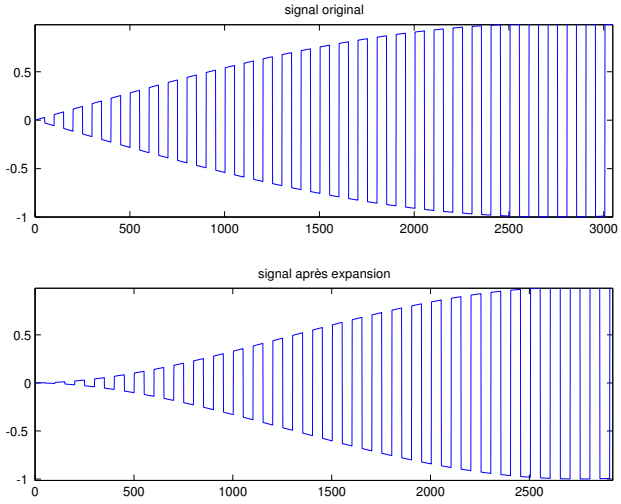


FIG. 3.3 – *Signal carré modulé en amplitude avant et après expansion.*

Intérêt de cet effet La plus grande application de l'expandeur est sans doute la réduction de bruit de fond. Le *noise gate* est utilisée pour éliminer les bruits ou masquer des bruits risquant d'être amplifiés ou entendus lorsque un instrument ne joue pas, et pour réduire les risques d'effets de larsen (*feedback*). Le seuil doit alors être suffisamment élevé pour supprimer les bruits indésirables, mais pas trop non plus pour éviter de couper intempestivement des notes instrumentales ou prématurément

les parties décroissantes de notes. Les expandeurs peuvent être utilisés conjointement aux compresseurs pour réduire les effets de bruit de fond d'une chaîne de transmission de signal, audio ou autre. Tout canal de transmission ayant une bande passante dynamique réduite, on compresse le signal à transmettre afin d'élever le niveau moyen du signal par rapport au niveau du bruit. Un expandeur est utilisé à la réception afin de rendre au signal sa dynamique originale : il s'agit du processus de compression-expansion ou *companding* (ex. la technique de réduction de bruit Dolby A utilisé pour les cassettes).

Compléments d'information Le placement de cet effet dans une chaîne d'effets est d'une grande importance. La plupart des effets produisant du bruit, on préfère placer l'expandeur en fin de chaîne, de façon à ce que le bruit généré et éventuellement amplifié soit inaudible lorsqu'aucun signal n'est envoyé en entrée (le musicien utilisant la chaîne d'effets de joue pas). Cependant, il est préférable de le placer avant les effets d'écho et de réverbération, pour éviter de dupliquer le bruit, et aussi afin d'éviter une rupture brutale de derniers échos ou des dernières répliques de la réverbération.

3.2.4 Compresseur, limiteur

Définition Le compresseur est un traitement non linéaire qui réduit la plage dynamique d'un signal. Il est très utilisé en enregistrements audio, en production, pour la réduction de bruit de fond, et pour les performances, mais doit être manipulé avec soin.

Fonctionnement Le diagramme de fonctionnement du compresseur est le même que celui de l'expandeur (*cf. fig.* 3.1), seule la fonction de conformation de gain diffère (*cf. fig.* 3.4). Le gain est réduit lorsque le niveau du signal est élevé, ce qui rend les passages forts un peu plus doux, en réduisant la plage dynamique. Elle s'explique de la même manière que la relation entrée-sortie de

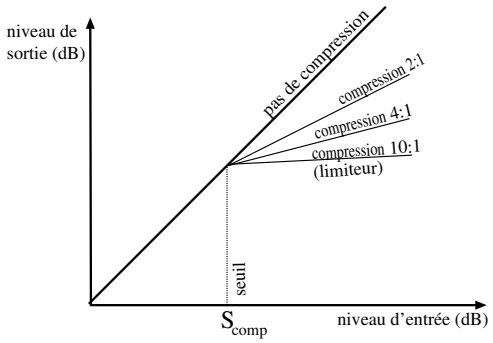


FIG. 3.4 – *Fonction de conformation du compresseur.*

l'expandeur : au dessus du seuil S_{comp}, un rapport de compression R_{comp} de 4 : 1 signifie qu'il faudra que le niveau d'entrée augmente de $4dB$ pour que celui de sortie augmente de $1dB$. Le **limiteur** est un compresseur avec un taux de compression extrême ($R_{comp} \geq 10$).

$$N_{sortie} = \begin{cases} N_{entree} & \text{si } N_{sortie} \leq S_{comp} \\ \frac{1}{R_{comp}}(N_{entree} - S_{comp}) + S_{comp} & \text{si } N_{sortie} \geq S_{comp} \end{cases} \quad (3.5)$$

Le niveau d'énergie est mesuré par moyenne sur une portion courte du signal (*cf.* 2.3.3), mais peut aussi l'être par pic instantané ou valeur échantillonnée (*cf.* 2.3.3). Le temps d'attaque ($\tau_{a,comp}$) est le temps nécessaire au compresseur pour répondre à un accroissement de niveau franchissant le seuil S_{comp}. Le temps de relaxation ($\tau_{r,comp}$) correspond au temps nécessaire au compresseur pour répondre à une diminution de niveau passant en dessous du seuil, temps généralement plus grand que le temps d'attaque.

Selon l'application, il peut être préférable d'avoir un faible temps d'attaque ou un faible temps de relaxation, ce qui requiert une rapide variation de gain : cela sera perçu comme un effet de "respiration" (*breathing*) ou de "pompage" (*pumping*). Lorsque le niveau sonore est inférieur au seuil, le gain augmente et passe à 1. Le signal est alors plus proche du niveau de bruit du système, ce qui rend le bruit audible. Un compresseur plus sophistiqué étudiera plus en détail le signal d'entrée et ajustera le gain lorsque le niveau d'entrée atteint momentanément 0, afin de limiter cet effet indésirable de "respiration".

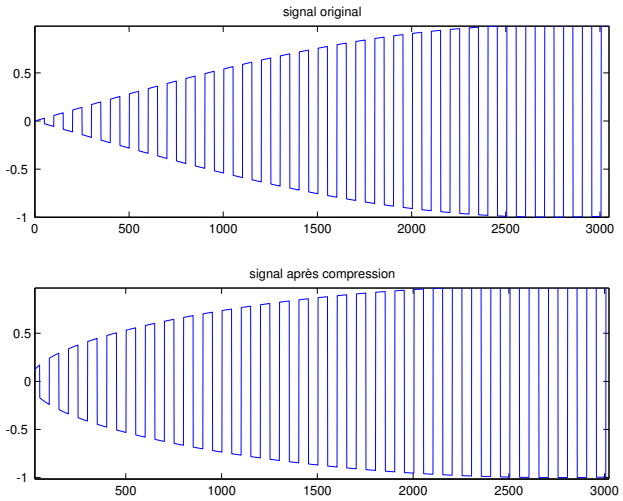


FIG. 3.5 – *Signal carré modulé en amplitude avant et après compression et re-normalisation.*

Intérêt de cet effet Limiter le niveau de certains signaux permet d'éviter que le signal d'une piste magnétique ne s'étende sur les pistes voisines (magnétisation trop forte d'une piste qui magnétise partiellement les particules des pistes voisines) lors d'enregistrement studios. Cela permet aussi d'éviter les phénomènes de distorsions indésirés, du fait que les appareils électroniques, analogiques ou numériques recevant les signaux ont un niveau d'entrée maximal (protégeant en même temps le matériel). En studio, la compression est un outil pratique pour découper des pistes et ajuster des mixages (en diminuant les différences excessives de niveau d'une piste, par exemple). En utilisant un temps d'attaque approprié, on conservera le naturel de l'attaque d'un son instrumental après compression. Dans certains cas, la compression permet de réduire le besoin d'égalisation.

Une autre utilisation du compresseur est l'augmentation de tenue (*sustain*) d'un instrument. Le compresseur peut maintenir un niveau quasi constant pour le signal de sortie en amplifiant convenablement le signal d'entrée. Une faible compression empêchera le niveau de l'instrument de changer trop radicalement, ce qui est ressenti comme une augmentation de tenue, ou un "lissage" du son. Un plus grand temps de relaxation que le temps de décroissance du son instrumental permet de conserver les propriétés du son de l'instrument après compression.

Compléments d'information Le limiteur croisé (*ducking/cross limiting*) correspond à un limiteur appliqué à un signal (en l'occurrence la musique) et piloté par un autre signal, la voix (premier exemple d'effet adaptatif croisé). Il permet par exemple à un animateur ou un *Disk Jokey* (D.J.) de gérer automatiquement le gain du fond sonore lors des prises de parole. De plus, il permet de donner de l'importance à certains éléments (tels les percussions) lors de l'enregistrement studio, augmentant leur présence en baissant le niveau des autres pistes. En appliquant le limiteur à certaines bandes fréquentielles, on obtient un *de-esser* (cet effet supprime les "s"). Lorsqu'on utilise le compresseur avec d'autres effets, il est préférable de la placer en premier dans la chaîne, tout d'abord pour des raisons de bruit. Un compresseur en fonctionnement (réduisant la gamme dynamique de la sortie)

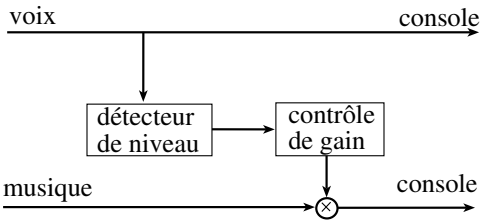


FIG. 3.6 – *Diagramme du limiteur croisé.*

suivi d'un gain additionnel sur la sortie amplifie le bruit en même temps que le son de l'instrument. D'autres effets peuvent introduire du bruit de fond, que l'on préfère ne pas amplifier.

3.2.5 Trémolo

Le trémolo correspond à une modulation d'amplitude en dessous de 20 Hz, audible dans le domaine temporel. Il est utilisé en tant que tel comme mode de jeu, et est aussi présent dans le vibrato (volontairement ou non de la part du musicien). Soit $x(n)$ le signal à traiter. Le signal modulé est donné par :

$$y(n) = x(n).\left[1 - d_{tr}\left(1 - \frac{\sin(2\pi f_{tr} n/F_e)}{2}\right)\right] \tag{3.6}$$

si on applique le trémolo en échelle linéaire, et par :

$$y(n) = x(n).\left[1 - 10^{-\left(d_{tr}\frac{1-\sin(2\pi f_{tr} n/F_e)}{40}\right)}\right] \tag{3.7}$$

si on applique le trémolo en échelle logarithmique, ce qui fait plus de sens à la perception, avec $d_{tr} \in [0;1]$ la profondeur.

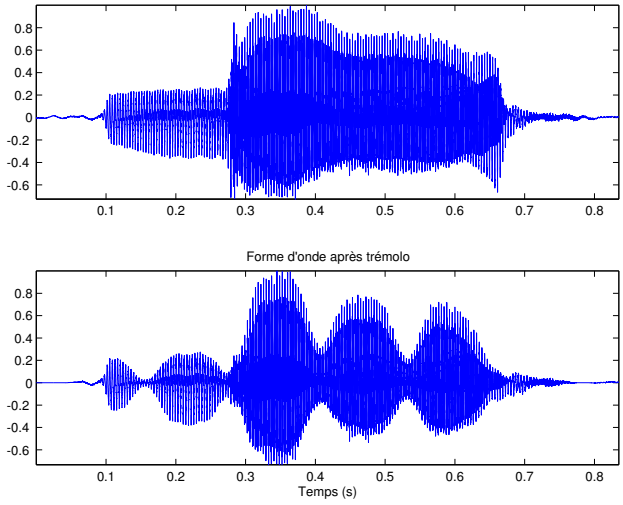


FIG. 3.7 – *Formes d'ondes d'un son de voix avant et après application d'un trémolo de fréquence* 6 Hz *et de profondeur 0.7 (*-3 dB*).*

3.3 Traitements modifiant l'échelle temporelle

Le temps et la fréquence sont intrinsèquement liés. Ainsi, prenons une sinusoïde à 440 Hz échantillonnée à 44100 Hz. Sur une période de 1 seconde, on a 44100 valeurs de cette sinusoïde

discrétisée, avec 440 périodes durant cette seconde. Si l'on procède à un changement de fréquence d'échantillonnage (par exemple 22500 *Hz*), on n'aura plus que 220 périodes en une seconde (donc un son à 220 *Hz*, une octave en dessous), et il faudra deux fois plus de temps pour lire le même signal.

Ce problème de changer l'échelle temporelle indépendamment du contenu fréquentiel est classique en traitement du son. Il est lié à la perception. La somme de deux sinusoïdes peut être vue comme le produit d'une porteuse et d'une modulante. Une fois l'étirement temporel effectué, doit-on avoir la somme des deux mêmes sinusoïdes ou la même porteuse multipliée par une modulante différente ? En dessous de 10 *Hz*, une modulation est entendue comme un trémolo ou un vibrato ; au-dessus, cela donne des fréquences audibles.

Il faut donc mettre en œuvre des méthodes d'analyse-synthèse afin de pouvoir modifier le temps sans modifier les fréquences. L'idée de base est de répéter une portion de son : si je prends un grain de son fenêtré, et que je l'ajoute en sortie avec un décalage temporel, pour peu que les phases soient en continuité, j'aurai un son plus long, de même hauteur. Ce principe peut être appliqué dans le domaine temporel TD-PSOLA, *cf.* 2.4.1), dans le domaine temps-fréquence (vocodeur de phase, *cf.* 2.4.2) ou dans le domaine fréquentiel (modèle spectral, *cf.* 2.4.3).

On appelle **dilatation/contraction temporelle** le fait de modifier l'échelle temporelle. Cette modification de l'échelle temporelle ou de la durée du son peut se faire par dilatation (étirement, ralenti, allongement) ou par contraction (compression, accéléré). Le terme anglais utilisé est *time-stretching* ou *time scaling*, selon les auteurs. Par souci de simplicité, nous choisissons de n'utiliser qu'un seul de tous ces termes dans le document, à savoir "dilatation/contraction temporelle", sachant que le facteur de dilatation/contraction peut être supérieur à 1 (dilatation) ou compris entre 0 et 1 (contraction).

3.3.1 Dilatation/contraction temporelle linéaire sans conservation de la hauteur

La dilatation/contraction temporelle sans conservation de la hauteur (ni des formants) consiste à modifier la fréquence d'échantillonnage d'un signal, ce qui modifie la hauteur perçue ainsi que les formants et la durée du signal, c'est pourquoi il est expliqué en tant qu'effet modifiant plusieurs paramètres, en 3.7.2.

3.3.2 Dilatation/contraction temporelle linéaire avec conservation de la hauteur

Définition La dilatation/contraction temporelle linéaire avec conservation de la hauteur consiste à ralentir ou accélérer un son tout en préservant la hauteur d'origine.

Fonctionnement La manière de réaliser cet effet consiste à conserver l'information fréquentielle lorsque la dilatation/contraction temporelle est réalisé. Pour ce faire, plusieurs méthodes existent.

Les traitements segment-temporels (type SOLA, PSOLA) répètent ou suppriment des segments d'échantillons, de la taille d'une période fondamentale, afin de modifier la durée. Ceci fonctionne particulièrement bien pour des signaux harmoniques et surtout la partie stable.

Les traitements spectraux utilisent l'information de module et de fréquence des partiels analysés pour synthétiser un son de la durée requise, en interpolant les valeurs des modules et fréquences de synthèse à partir des valeurs d'analyse. Le bruit quant à lui est resynthétisé par filtrage, l'enveloppe spectrale du filtre étant déduite par interpolation des enveloppes spectrales d'analyse.

Les traitements temps-fréquence (vocodeur de phase) synthétisent le son en utilisant eux aussi un pas d'analyse différent du pas de synthèse, et en recalculant les phases à chaque itération de synthèse, de façon à ne pas provoquer de rupture de phase. En effet, changer les valeurs d'une transformée de Fourier de manière arbitraire n'est pas une opération valide (la transformée obtenue n'est pas forcément l'image d'un signal réel) : le son de synthèse peut alors être différent de ce que à quoi on s'attendrait. Un bon algorithme de transformation consiste à trouver une stratégie

qui préserve l'aspect de changement d'échelle temporelle sans introduire trop d'artefacts. Pour le vocodeur de phase (par bloc FFT/IFFT), l'idée est donc de conserver la magnitude et de modifier la phase de façon à préserver la fréquence instantanée f_i. La différence de phase du panier de fréquence k est donc $\Delta\psi(k) = \frac{R_s}{R_a}\Delta\phi(k)$, avec R_a le pas d'analyse, R_s le pas de synthèse, d'où $\frac{R_s}{R_a}$ le facteur de changement d'échelle temporelle. La différence de phase à l'analyse se calcule comme suit :

$$\Delta\phi(k) = \frac{2\pi R_a(k-1)}{N} + \text{princarg}\left(\tilde{\phi}\left((s+1)R_s, k\right) - \tilde{\phi}\left(s\ R_s, k\right) - \frac{2\pi R_a(k-1)}{N}\right) \tag{3.8}$$

d'où l'on obtient la nouvelle phase de synthèse :

$$\psi\left((s+1)R_s, k\right) = \text{princarg}\left(\psi\left(sR_s, k\right) + \Delta\phi(k)\frac{R_s}{R_a}\right) \tag{3.9}$$

Intérêt de cet effet Cet effet permet de modifier la durée d'un son, en donnant une expressivité toute autre que celle de départ. Il permet aussi de synchroniser deux signaux de durées initiales différentes ; il est aussi utilisé en post-production cinéma-vidéo, du fait que les formats sont différents (24 images/s pour l'un, 25 images/s pour l'autre), avec préservation des formants [Pallone *et al.*, 1999] (ceci sera évoqué avec la transformation duale, la transposition avec préservation de la durée et des formants, *cf.* 3.4.3).

Compléments d'information Plusieurs remarques quant à la mise en œuvre du changement d'échelle temporelle par vocodeur de phase sont à prendre en compte. Tout d'abord, le changement de phases correspond à l'application d'un filtre passe-tout. Il faut appliquer une fenêtre de resynthèse, sinon un effet de convolution circulaire (due au filtrage) implique des discontinuités aux extrémités des fenêtres. Ensuite, même avec l'application d'une fenêtre de synthèse, il reste un aspect de convolution circulaire (c'est la version repliée d'une FFT infinie) : il est nécessaire de procéder à un bourrage de 0 à l'analyse comme à la synthèse.

Pour s'assurer d'une reconstruction parfaite lorsque les pas et les fenêtres d'analyse et de synthèse sont identiques, la somme des carrés des fenêtres (soit leur puissance) espacées d'un pas doit toujours être de 1, pour chaque échantillon (on doit s'assurer de ne pas ajouter de modulation d'amplitude avec l'utilisation de fenêtres)[1]. Par exemple, pour une fenêtre de Hanning sans bourrage de 0, le pas de synthèse R_s doit être un diviseur de $\frac{N}{4}$. Pour les fenêtres de Hamming et Blackman, les lobes secondaires sont plus bas que pour la fenêtre de Hanning, mais leur valeur aux extrémités est non nulle, ce qui implique des problèmes aux extrémités si on les utilise seules. Il faut un pas de synthèse R_s diviseur de $\frac{N}{8}$. Pour une fenêtre de Gauss tronquée, la somme des fenêtres décalées d'un pas produit toujours des oscillations (modulations d'amplitude), mais qui peuvent être rendues en dessous du seuil de perception avec un pas bien choisi.

Remarquons enfin que lorsque le pas de synthèse est multiple entier du pas d'analyse : $\frac{R_s}{R_a} = k \in \mathbb{N}$, il n'est plus nécessaire de calculer le déroulement de phase, puisque la relation modulo 2π est conservée. Par contre, lorsque ce ratio est réel, il résulte un problème de dispersion des phases. En effet, toutes les phases sont forcées de tourner à la vitesse à laquelle elles tourneraient si une sinusoïde de fréquence égale à la fréquence centrale du panier de fréquence existait dans le signal. Ce problème se résout par le verrouillage de phase [Laroche and Dolson, 1997; Laroche and Dolson, 1999].

3.3.3 Inversion temporelle

L'inversion temporelle du signal est un traitement non temps-réel, qui consiste à faire se dérouler le temps à l'envers, en lisant les échantillons du dernier au premier. Ceci permet notamment de

[1] Ce problème de normalisation est tout aussi vrai pour des pas variables, ce qui est le cas des traitements adaptatifs, pour lesquels des adaptations devront être réalisées.

modifier complètement la perception du son. Par exemple, les sons percussifs ne seront plus perçus comme tels, puisque l'amplitude ira *crescendo* puis chutera brusquement. Cet effet a été utilisé depuis les bandes analogiques, et popularisé notamment par les Beatles (avec l'aide de leur ingénieur du son, George Martin), Jimi Hendrix à la fin des années 60.

3.4 Traitements modifiant la hauteur

Les différents traitements existant sur la hauteur sont basés sur la transposition (avec ou sans conservation des formants).

3.4.1 Transposition sans conservation de la durée ni des formants

La transposition sans conservation de la durée ni des formants (*cf.* 3.7.2) consiste à modifier l'échantillonnage d'un signal, de façon à modifier la hauteur perçue. Ce traitement ne conserve ni les formants, ni la durée du signal. C'est la transformation duale du changement d'échelle temporelle sans conservation de la hauteur.

3.4.2 Transposition sans conservation des formants

Définition La transposition sans conservation des formants consiste à modifier la hauteur (transposer) de manière brutale, sans modifier l'échelle temporelle et sans respecter l'enveloppe spectrale. Il s'agit avant tout d'un artefact de méthodes de changement de hauteur, basées sur le ré-échantillonnage. Cet effet conduit facilement à des effets de type "Donald Duck".

Fonctionnement La hauteur d'un signal étant donnée par sa répartition en fréquences (un signal harmonique composé d'harmoniques, partiels de fréquences multiples de la fréquence fondamentale), en changer la hauteur se fait en multipliant toutes les fréquences par un même facteur γ.

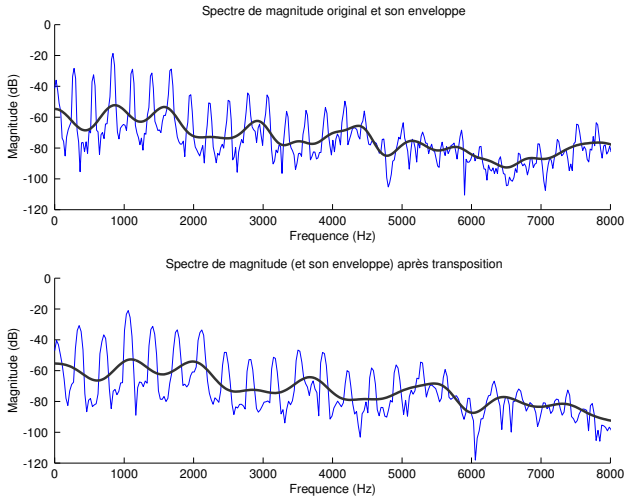


FIG. 3.8 – *Sonagramme d'un son voisé avant et après transposition sans conservation des formants (l'enveloppe spectrale est dilatée avec le spectre).*

L'utilisation d'un modèle spectral (SMS, SAS) permet aisément cette manipulation. Le vocodeur de phase quant à lui nécessite une mise en œuvre plus complexe. Le changement de hauteur est en lien direct avec l'algorithme de changement de durée avec le vocodeur de phase : si l'on applique un changement de durée d'un facteur $\gamma_{duree} = \frac{R_s}{R_a}$ puis un ré-échantillonnage avec le facteur inverse $\gamma_{hauteur} = \frac{1}{\gamma_{duree}}$, on obtient un son de même durée que l'original, mais dont la hauteur a changé

du fait du ré-échantillonnage. Soit $L_i = \frac{R_s}{R_a} L_a$ la durée du signal rallongé, la durée du signal de synthèse est, après ré-échantillonnage, $L_s = \frac{R_s}{R_a}\ L_i = \frac{R_a}{R_s}\frac{R_s}{R_a}\ L_a = L_a$.

Cet effet peut aussi s'obtenir à l'aide de traitements segment-temporels : par contraction-étirement temporel et ré-échantillonnage (comme expliqué précédemment), par lecture à vitesse variée (avec 2 ou 4 têtes de lectures tournantes).

3.4.3 Transposition avec conservation des formants

Définition La transposition avec conservation des formants consiste à appliquer un changement de hauteur tout en gardant l'information formantique du signal. Cette transposition est en générale plus proche de la transposition effectuée par un instrument de musique que le traitement ne préservant pas les formants. On l'appelle aussi "p-transposition" ou "transposition-p" pour préciser que la contrainte perceptive de timbre est respectée, et c'est la transformation duale de la "p-dilatation". Elle permet aussi d'éviter l'effet "Donald Duck" de la transposition d'une voix sans préservation des formants.

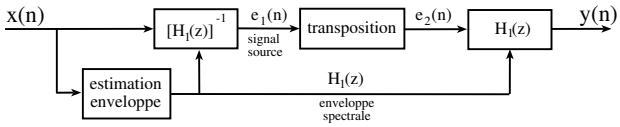


FIG. 3.9 – *Diagramme de la transposition avec conservation des formants.*

Fonctionnement De même qu'on ne peut directement modifier la hauteur d'un signal sans en modifier la durée, on ne peut modifier la hauteur d'un signal sans affecter les formants, à moins de mettre en œuvre des techniques particulières. Aussi, il faut d'abord calculer l'enveloppe spectrale du son (par cepstre ou LPC), effectuer la transposition (par TD-PSOLA, vocodeur de phase, modèle additif), puis corriger les amplitudes des pics fréquentiels en fonction de l'enveloppe spectrale d'origine, *cf. fig.* 3.9.

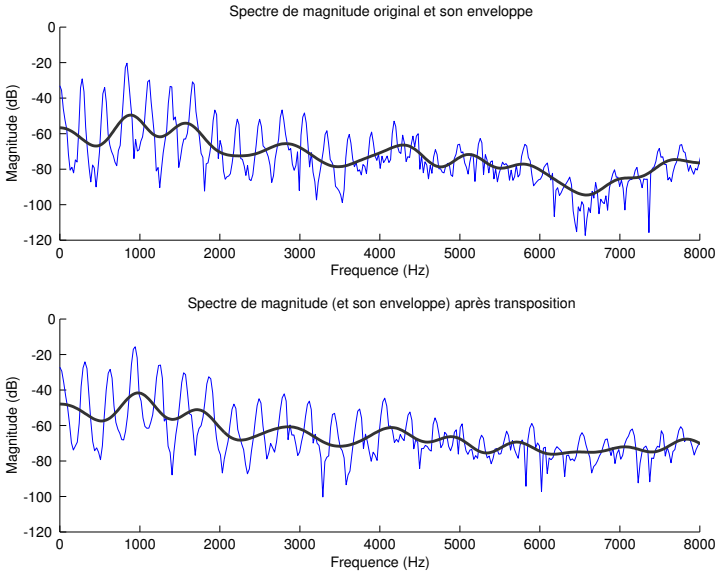


FIG. 3.10 – *Sonagramme d'un son voisé avant et après transposition avec conservation des formants (l'enveloppe spectrale est conservée).*

Mise en œuvre par le vocodeur de phase La transposition avec préservation des formants se fait à l'aide du modèle source-filtre [Arfib and Zoelzer, 2002a] et par ré-échantillonnage [Arfib and Zoelzer, 2002b]. Il existe deux manières de procéder : soit on applique une correction sur les formants (par dilatation/contraction) en sens inverse de celle due au ré-échantillonnage, puis on applique la transposition (solution que nous avons adoptée) ; soit on applique d'abord le ré-échantillonnage puis la correction sur les formants.

Concernant la première solution, l'étape de correction de l'enveloppe se fait dans le domaine fréquentiel, entre l'étape de dilatation/contraction et l'étape de ré-échantillonnage. Soit un grain de signal $x(n)$ et sa TFCT $X(n, f)$. On effectue la séparation source-filtre dans le domaine fréquentiel à l'aide de la méthode du cepstre, et on obtient $\mathcal{S}(f)$ la source et $\mathcal{E}(f)$ l'enveloppe. Ensuite, on passe dans le domaine temporel, et on ré-échantillonne la source $s(n) = TFCT^{-1}(\mathcal{S}(f))$ pour la transposer du facteur γ désiré. Ensuite, on revient dans le domaine fréquentiel pour multiplier la source transposée avec l'enveloppe spectrale. Ceci revient à convoluer la nouvelle source avec le filtre original. Enfin, on applique une TFCT inverse pour obtenir le grain temporel transposé (*cf. fig.* 3.11).

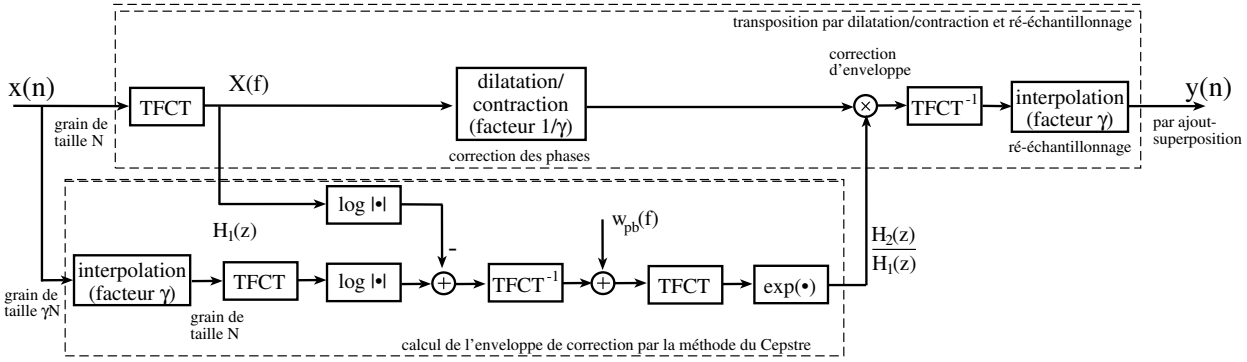


FIG. 3.11 – *Diagramme de la transposition avec préservation des formants par vocodeur de phase par dilatation/contraction temporelle puis ré-échantillonnage. La correction d'enveloppe se fait après la dilatation temporelle.*

Mise en œuvre par le modèle additif [Amatriain *et al.*, 2001] Soit la forme spectrale comme l'enveloppe décrite par les amplitudes et fréquences des partiels :

$$\mathcal{S} = \{(f_1, a_1), (f_2, a_2), ..., (f_H, a_H)\} \tag{3.10}$$

Le spectre obtenu après transposition sera donné par :

$$X_{transp}(f) = \sum_{i=1}^{H} \delta(f - \gamma f_i).\mathcal{S}(\gamma.f_i) \tag{3.11}$$

C'est encore mieux si on applique un filtre en peigne au résidu pour les sons harmoniques, de même hauteur que la hauteur d'origine. En effet, on supprime ainsi les bandes latérales aux sinusoïdes, qui ne sont pas toujours bien effacées.

Du fait que la transposition peut se faire en re-synthétisant chaque harmonique dans le domaine temporel (et aussi fréquentiel), échantillon par échantillon, il n'y a aucune contrainte pour que le facteur de transposition soit constant dans le temps ou non.

3.4.4 Discrétisation sur une échelle tempérée

La discrétisation sur une échelle tempérée consiste à modifier la hauteur d'un signal de telle sorte que la fréquence fondamentale appartienne toujours à la gamme tempérée. Ceci revient en quelque

sortes à accorder le signal sur une échelle prédéterminée. Cet effet s'appelle souvent *Auto-tune*, du fait qu'il accorde le signal. Cela permet par exemple de mixer ensemble des pistes qui n'étaient auparavant pas accordées.

Il s'agit donc dans un premier temps de calculer la hauteur (à défaut la fréquence fondamentale) du signal, puis de la modifier. Dans une certaine mesure, on pourra considérer que des écarts de hauteur entre la hauteur mesurée et la plus proche hauteur de l'échelle tempérée supérieurs à une tolérance impliqueront la modification, tandis qu'un écart inférieur à la tolérance ne sera pas modifié. C'est un autre exemple d'effet adaptatif, du fait qu'une analyse de la fréquence fondamentale sert à piloter la transposition.

3.4.5 Harmoniseur

Un harmoniseur est un effet qui ajoute à un signal sa version transposée. Selon le nombre de transpositions ajoutées et leur facteur de transposition, on obtient différents accords. Cet effet permet par exemple de remplacer une voix monophonique par un choeur harmonisé, et donc d'enjoliver des mélodies. Le principe est simple : une fois la règle d'harmonisation définie (par exemple, pour obtenir un accord parfait majeur, il faut transposer d'une tierce majeure et d'une quinte), les versions transposées $T_{\gamma_i}(x)(n)$ du signal sont calculées (avec γ_i le facteur de transposition), puis ajoutées au signal de départ $x(n)$.

Il existe maintenant des harmoniseurs adaptatifs, fonctionnant en temps-réel et basés sur le modèle additif, et permettant d'effectuer les bonnes transpositions pour respecter la tonalité selon des règles d'harmonie données par l'utilisateur [TC-Helicon, 2002].

3.5 Traitements modifiant la spatialisation

Nous abordons maintenant les effets portant sur les notions de localisation, de spatialisation, de distance entre l'auditeur et la source, les effets produits par la salle, par les mouvements de l'auditeur ou de la source, les effets de rayonnement et de directivité le son étant représentée dans un espace à trois dimensions (*cf. fig.* 3.12).

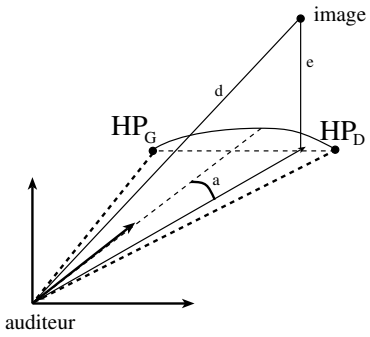


FIG. 3.12 – *Localisation par un rendu 3D d'un son : azimut (a), distance (d) et élévation (e).*

3.5.1 Écho

L'écho (ou délai) est l'un des effets les plus simples à réaliser. Il s'agit de créer une ou plusieurs répliques du son d'entrée décalées dans le temps. Le délai simple prend un signal en entrée et le décale dans le temps d'une durée τ_{delay}. Ce temps de retard peut varier entre quelques millisecondes et quelques secondes. La figure *fig.* 3.13 présente le diagramme de cet écho simple.
Un délai simple étant un effet limité, des systèmes de délai plus complexes sont réalisés avec une boucle de retour ou de réinjection (*feedback*/régénération), *cf. fig.* 3.14. La sortie du premier écho

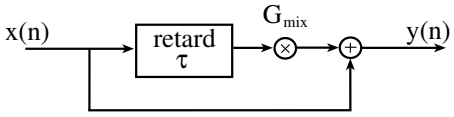


FIG. 3.13 – Diagramme du délai simple.

est mixée avec l'entrée, et pourvu que le gain soit inférieur à 1 (pour des raisons de stabilité de l'appareil : dans le cas contraire, la sortie sature très vite), on peut répéter durablement un son, dont le niveau baisse progressivement au fur et à mesure de son passage dans la boucle d'écho. Théoriquement, le son devrait se répéter indéfiniment avec la réinjection, mais il passe vite en dessous du niveau de bruit ambiant, et devient inaudible de ce fait.

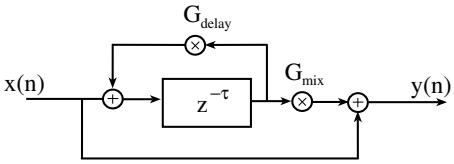


FIG. 3.14 – Diagramme du délai avec réinjection (*feedback*).

Intérêt de cet effet Les délais sont très utilisés pour enrichir un son instrumental. En utilisant de faibles temps de retard (50 à 100 millisecondes), il se crée un effet de doublage, comme si deux instruments jouaient à l'unisson. L'utilisation de plusieurs délais avec réinjection peut être utilisé pour créer un effet de réverbération, bien qu'une unité de réverbération classique crée un son plus complexe. Pour des temps de retard supérieurs à 100 millisecondes, l'effet est moins subtile, la répétition s'entend comme telle. Aussi, une possibilité intéressante consiste à faire correspondre ce temps au tempo, de façon à ce que les répétitions tombent exactement sur l'un des battements suivants. Les systèmes de bouclage et d'échantillonnage découlent de ceci, en prenant une portions de mélodie ou de progression harmonique ou mélodique et en la répétant indéfiniment (techno). Cela permet aussi d'improviser sur ses propres grilles, par exemple (jazz). Plusieurs autres styles de musiques populaires sont construites sur ce modèle. Les délais sont aussi d'un grand intérêt pour le mixage d'enregistrements stéréophoniques. Cela grandit le son du mixage, et permet une meilleure sensation de placement des instruments (en modifiant à la fois le gain de chaque piste et le délai, qui correspond au temps d'arrivée du son à chaque oreille de l'auditeur) dans un espace à deux dimensions (*cf.* 3.5.5). Un délai de 20 millisecondes peut donner une grande différence au mixage final.

Compléments d'information La mise en œuvre numérique d'un délai se fait à l'aide d'une mémoire tampon circulaire (*circular buffer*), dont la longueur est déterminée par le temps de retard τ_{delay}. Ainsi, pour un retard d'une seconde de son échantillonné à $44100Hz$, il faut un buffer de 44100 valeurs. Dans le cas où le temps de retard ne correspond pas à un nombre d'échantillons entier, on utilise des lignes à retard fractionnaire (*cf.* 2.2.4) [Laakso *et al.*, 1996].

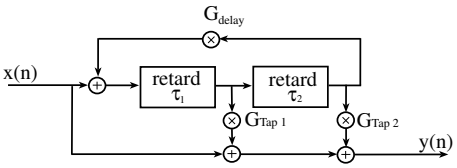


FIG. 3.15 – Diagramme du multi-délai.

D'autres types de délais, plus perfectionnés mais basés sur ces principes, existent. Le **slapback** est un délai simple à temps de retard très court (40 à 120 ms). Le **Multi-Tap delay** consiste à utiliser plusieurs délais avec réinjection (*cf. fig.* 3.15 pour un exemple à deux répétitions), combinés de façon à pouvoir créer des motif rythmiques complexes. Le **Ping-pong delay** crée un signal stéréophonique, et peut s'appliquer à un signal stéréophonique. Les échos de chaque canal sont envoyés dans l'autre canal, et ceci avec une réinjection. Cela crée un effet de rebond.

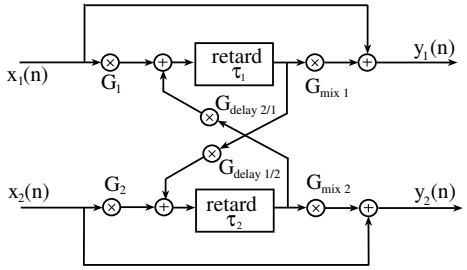


FIG. 3.16 – Diagramme du délai ping-pong.

3.5.2 Panoramisation

Définition La panoramisation permet de modifier l'azimut d'un son dans les systèmes multi haut-parleurs. Dans le cas de deux haut-parleurs par exemple, il s'agit de pouvoir produire un signal semblant provenir de n'importe quel point entre les deux haut-parleurs. Dans un système à N haut-parleurs situés en cercle autour des auditeurs, le positionnement du son en azimut se fait à partir de deux haut-parleurs seulement. L'auditeur est situé placé au centre du cercle de haut parleurs, ou du moins à égale distance des deux hauts-parleurs (en stéréophonie).

Fonctionnement Soit deux haut-parleurs formant un angle θ_{max} avec l'auditeur. La **panoramisation linéaire** est la plus simple ; les amplitudes correctrices du signal dans chaque haut parleur sont données par :

$$\alpha_G = \frac{\theta}{\theta_{max}} \tag{3.12}$$

$$\alpha_D = 1 - \frac{\theta - \theta_{max}}{\theta_{max}} \tag{3.13}$$

$$\tag{3.14}$$

Le grand défaut de cette panoramisation est qu'elle crée un "trou" au milieu, entre les deux haut-parleurs : le son semble être plus faible au centre, comme s'il était éloigné de l'auditeur. Au milieu de la panoramisation l'intensité vaut $I = \sqrt{\alpha_G^2 + \alpha_D^2} = \sqrt{0.5^2 + 0.5^2} = 0.707$. Cette chute d'intensité correspond à une chute de 3 dB, ce qui correspond environ à un changement de nuance en musique (de mf à f par exemple, sur l'échelle ppp, pp, p, mp, mf, f, ff, fff). La **panoramisation à puissance constante** utilise des courbes sinusoïdales pour contrôler l'amplitude du signal émis par chaque enceinte. Ceci corrige le défaut de la panoramisation linéaire :

$$\alpha_G = \frac{\sqrt{2}}{2}(\cos\theta + \sin\theta) \tag{3.15}$$

$$\alpha_D = \frac{\sqrt{2}}{2}(\cos\theta - \sin\theta) \tag{3.16}$$

Elle est adaptée à une situation où les enceintes sont placées avec un angle de $\pi/2$ radians, $\pi/4$

radians de part et d'autre de l'auditeur. L'intensité est donnée par :

$$I = \sqrt{\alpha_G^2 + \alpha_D^2} \tag{3.17}$$

$$= \sqrt{\frac{1}{2}\left((\cos\theta + \sin\theta)^2 + (\cos\theta - \sin\theta)^2\right)} = 1 \tag{3.18}$$

L'intensité est donc bien la même pour tout angle $\theta \in [-\frac{\pi}{4}; \frac{\pi}{4}]$.

3.5.3 Effet de précédence

Lorsque un signal identique est envoyé dans deux haut-parleurs avec des délais différents de quelques millisecondes, il est perçu comme venant d'un côté particulier : ceci est dû à l'effet perceptif de **précédence**. Pour un son effectivement latéralisé, la différence de temps d'arrivée aux oreille est non-nulle. Cet effet est un complément de la panoramisation en vue d'une latéralisation parfaite du signal. Il suffit d'utiliser deux lignes à retard, et de moduler leurs délais en fonction de l'angle de panoramisation désiré.

3.5.4 Réverbération

Le phénomène de réverbération a été présenté en 1.2.4.

Fonctionnement La manière de simuler une réverbération consiste à modéliser les trois constituants, soit distinctement, soit à l'aide d'un dispositif qui reproduit de facto ces constituants. On dénote quatre approches : par ligne à retard, par ensemble de filtres passe-tout / filtres en peigne, par réseau de lignes à retard avec réinjection, et enfin par convolution avec une réponse impulsionnelle de salle. Cet effet permet d'ajouter un effet de salle à un son brut, par exemple enregistré en chambre anéchoïque ou en champ très proche et dans une salle à parois absorbantes.

Réverbération par systèmes de filtres et lignes à retard Les premiers systèmes de réverbération numérique, décrits par Schroeder et Logan [Schrœder and Logan, 1961] aux laboratoires Bell dans les années soixante puis réalisés par Chowning [Chowning, 1971a] à Stanford University, utilisaient un filtre passe-tout basé sur une ligne à retard récursive (ou à réinjection). Ce filtre permet une réponse impulsionnelle dense et une réponse en fréquence plate. Jusqu'au début des années 80, cette structure simple, efficace et peu onéreuse en temps de calculs a été très largement utilisée. On considère souvent qu'un filtre passe-tout ne donne aucune coloration au son : ceci n'est vrai (d'un point de vue perceptif) que lorsque la ligne à retard est de longueur inférieure au temps d'intégration de l'oreille, à savoir environ 50 ms. Cette limitation explique la recherche d'autre modèles par la suite.

Dans les années soixante-dix, Gerzon [Gerzon, 1976] a généralisé le filtre passe-tout à une entrée et une sortie, pour obtenir une structure de type réseau à N entrées et N sorties. L'idée était d'augmenter la complexité de la réponse impulsionnelle sans introduire de coloration notable.

Moorer [Moorer, 1979] développa les études de Schroeder en mettant en lien les structures de calcul (lignes à retard, filtre en peigne et filtre passe-tout) avec les propriétés physiques des salles. Cela lui permit de proposer des modèles plus réalistes.

Smith [Smith, 1984; Smith, 1987] a proposé les réseaux de guides d'ondes numériques comme modèle, entre autres, des réverbérations. Un guide d'onde est modélisé par une ligne à retard bidirectionnelle, avec filtres. Enfin, Stautner et Puckette [Stautner and Puckette, 1982] ont jeté les bases de ce qui forme les réseaux de délais à réinjection, formalisés ensuite par Jot [Jot and Chaigne, 1991; Jot, 1992].

Réverbération par convolution avec réponse impulsionnelle Dès lorsque l'on a enregistré la réponse impulsionnelle d'une salle, on sait que le son réverbéré est obtenu par convolution de cette réponse impulsionnelle $h(n)$ avec le signal $x(n)$, d'où le signal obtenu : $y(n) = (x * h)(n)$. Parmi les défauts de cette méthode, on relève la nécessité de longues réponses impulsionnelles afin d'obtenir des réverbérations naturelles, et l'impossibilité de manipuler simplement (par interpolation) les réponses impulsionnelles, notamment pour passer de l'un à l'autre simplement. Les structures avec filtres et lignes à retard se prêtent beaucoup mieux à ce genre de manipulations.

3.5.5 Distance

La distance d'une source sonore est évalué par l'auditeur à l'aide de plusieurs mécanismes (*cf.* 1.2.4) : tout d'abord le niveau du signal émis, ensuite le filtrage passe-bas du son (effectué par l'air), puis l'écho et la réverbération.

Les mécanismes que nous pouvons modéliser sont la réverbération, expliquée en 3.5.4, le filtrage passe-bas et la puissance à l'aide d'un simple potentiomètre. Le filtrage passe-bas et le changement du gain correspondent à ce qui se passe en espace ouvert (à l'extérieur), lorsque le son traverse de grande étendues d'air. Dans ce cas, les hautes fréquences sont absorbées par l'agitation thermique des molécules présentes dans l'air. Les indications d'écho et de réverbération modélisent ce qui se passe dans un espace clos, tel qu'une salle de concert. Pour simuler une distance spécifique à l'intérieur d'une pièce, le plus simple consiste à garder le même niveau de réverbération $P_{reverb} = P^0_{reverb}$ et d'échelonner le signal direct de façon à ce qu'il soit inversement proportionnel à l'inverse de la distance. En effet, la puissance du signal direct décroît en proportion inverse de la distance $P_{direct} = \frac{P_0}{d}$, tandis que la puissance du signal réverbéré décroît en proportion de l'inverse de la racine carrée de la distance $P_{reverb} = \frac{P_0}{\sqrt{d}}$. Cette seconde modélisation est bien plus fidèle à la réalité.

Pour modéliser un espace fermé avec une ouverture sur un second espace (plus grand), on utilise une réverbération globale relativement faible dans l'ensemble des N haut-parleurs, et une réverbération locale plus forte dans deux haut-parleurs adjacents, simulant la direction de l'ouverture. Cette distinction entre réverbération locale et globale possède un autre intérêt : en utilisant la relation $R_{locale} > \left(1 - \frac{1}{d}\right) R_{globale}$ qui permet d'augmenter la réverbération locale en fonction de la distance, on évite l'effet de masquage qui apparaît à des distances où les amplitudes des signaux directs et réverbérés globaux sont égales. En effet, lorsque la distance tend vers 0 (la source est très proche de l'auditeur), la puissance de la réverbération locale tend vers celle de la réverbération globale, elle est distribuée uniformément, dans tous les canaux. Au contraire, lorsque la source s'éloigne de l'auditeur, le signal réverbérant se concentre dans la direction de la source.

3.5.6 Effet Doppler

Définition L'effet Dopper correspond aux modifications de niveau et de hauteur (transposition) perçues lors des déplacements d'une source sonore qui s'approche et/ou s'éloigne de l'auditeur. Cet effet perceptif permet à l'auditeur d'analyser le mouvement et d'en connaître les changements de direction et de distance [Chowning, 1971a].

Fonctionnement Dès lors qu'une relation de mouvement existe entre la source sonore et l'auditeur, l'onde sonore s'allonge ou se comprime. Un signal sonore qui s'approche est perçu de façon plus aiguë qu'il n'est émis à la source, et plus grave lorsqu'il s'éloigne. La vitesse de déplacement de la source s'ajoute à celle de propagation du son. Lorsque les deux vont dans le même sens, les vitesses s'ajoutent et la fréquence perçue augmente ; lorsqu'elles vont en sens contraire, la fréquence diminue. De plus lorsque le changement de sens de la source est brusque, un effet complémentaire de rapprochement-éloignement s'ajoute. Notons c la célérité du son dans l'air et c_s la vitesse de déplacement de la source, c_a la vitesse de déplacement de l'auditeur. Considérons que l'auditeur et la source se déplacent selon un seul et même axe. Soit f_s la fréquence d'une sinusoïde émise par la

source, la fréquence perçue par l'auditeur sera :

$$f_d = f_s \frac{c}{c_s - c_a} \tag{3.19}$$

Le changement de hauteur est réalisé par l'une des méthodes présentée en 3.4, avec un rapport contrôlé par la vitesse relative $c_s - c_a$ entre source et auditeur. Le facteur de changement de hauteur est réalisé par une courbe du type *fig.* 3.17.

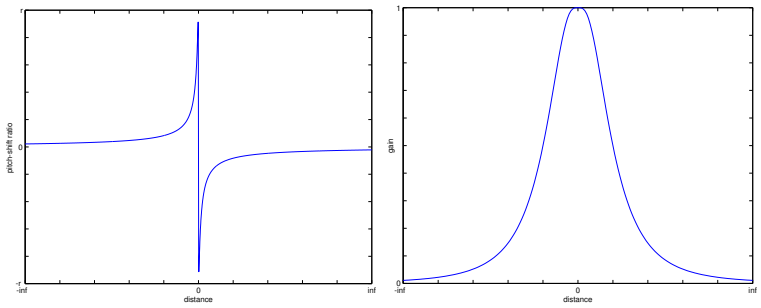


FIG. 3.17 – *Effet Doppler : facteur de changement de hauteur (*fig. *gauche) et gain (*fig. *droite).*

Remarquons qu'un changement de niveau est nécessaire lui aussi (le niveau sonore augmente lorsque la source s'approche, et diminue lorsqu'elle s'éloigne). Un effet connu en psychoacoustique est que lorsque la sonie (intensité sonore perçue) augmente, la hauteur perçue est elle aussi légèrement modifiée, ce qui s'ajoute au changement de hauteur imposé par les différences de vitesses auditeur-source.

3.5.7 Effet Rotary/Leslie

Définition Le rayonnement d'un son émis par un haut-parleur rotatif crée un effet de spatialisation saisissant. L'effet **Leslie**, encore appelé Rotary, consiste à mettre en rotation une enceinte [Smith *et al.*, 2002]. Cet effet rend vivants même les sons sourds et stables en les animant de qualités variant dans le temps.

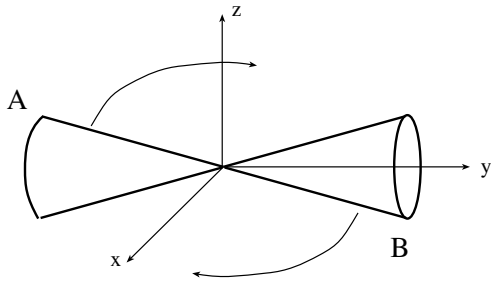


FIG. 3.18 – Système Rotary (à enceintes rotatives) permettant l'effet Leslie.

Fonctionnement Le mécanisme originel d'enceinte rotative fut la Cabine Leslie, qui envoyait un signal entrant dans deux mécanismes rotatifs séparés : un pavillon rotatif pour les hautes fréquences et un écran rotatif (bloquant et débloquant un boomer stationnaire) pour les basses fréquences. Une télécommande pour la vitesse du moteur permettait au musicien d'ajuster la vitesse de rotation. Le pavillon résonant de la Cabine Leslie la rend immédiatement identifiable. Ce système fut à l'origine construit pour enrichir le son d'orgues électriques, tels le Hammond B3, avec lequel il était souvent couplé. Les effets induits par la rotation sont nombreux, mettant en jeu le vibrato

de l'effet Doppler, le filtrage variant dans le temps, les déphasages, les distorsions causées par la turbulence de l'air, les réflexions d'échos des surfaces adjacentes, ainsi que les caractéristiques de transfert des amplificateurs et des enceintes utilisés. Le pavillon est double, le haut-parleur A ayant ses modifications de hauteur et d'intensité dans la direction opposée à celle du haut-parleur B.

Tous ces effets simultanés sont difficiles à simuler de façon convaincante en utilisant le traitement numérique du signal. Une combinaison de modulateurs (pour l'intensité) et de lignes à retard (pour moduler la hauteur) permet de simuler assez bien cet effet. La simulation de l'effet Doppler des deux pavillons opposés est réalisée par deux lignes à retard modulées avec des configurations de vibrato en déphasages de π radians. La caractéristique de directionnalité des haut-parleurs tournant peut être rendue par une modulation en amplitude des signaux en sortie des lignes à retard. Cette modulation est synchrone à la modulation de la ligne à retard, de manière à ce que le mouvement en arrière du pavillon produise une baisse de la hauteur et de la puissance du son. Au point de demi-tour, l'intensité est minimal et la hauteur non modifiée par rapport au son original. Le mouvement inverse, de l'arrière vers l'auditeur, produit les effets inverses : hausse de la hauteur et de l'intensité. Un effet stéréophonique de haut-parleurs tournant est obtenu par un mixage inégal des deux lignes à retard dans les sorties correspondant aux canaux gauche et droit.

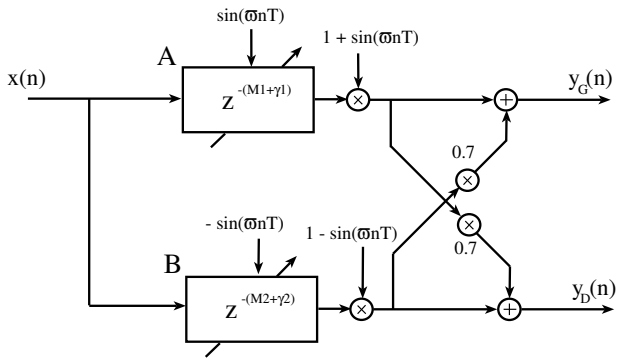


FIG. 3.19 – Diagramme de l'effet Leslie.

3.5.8 Rendu 3D (distance, azimut et élévation) avec casque

Le rendu de distance et d'espace en trois dimensions avec casque (ou en écoute binaurale) consiste à modéliser les filtrages dus aux réflexions des ondes sur la tête, le cou et le torse, avant d'arriver dans les conduits auditifs. Ainsi, lors de l'écoute au casque, le signal envoyé directement dans les conduits auditifs correspondra à celui d'un signal ayant parcouru l'espace dans les mêmes conditions. Pour réaliser cet effet, il faut d'abord parler de localisation et de sa perception, puis de la manière d'externaliser un son qui, à l'écoute au casque, paraît être situé dans la tête. Enfin, nous aborderons les fonctions de transfert relatives à la tête (*HRTF, Head-Related Transfert Functions*). Pour plus de détails sur ces techniques, nous conseillons vivement la lecture de [Rochesso, 2002].

Perception de la localisation Nous rappelons que la position d'une source dans l'espace est donnée en coordonnées sphériques par la distance (rayon), l'azimut et l'élévation (*cf. fig.* 3.12). La localisation d'un son est perçu grâce à trois données du signal sonore stéréophonique : la différence d'intensité interaurale, la différence de temps d'arrivée interaurale, les fonctions de transfert relatives à la tête (*cf.* 1.2.4). La prise en compte des mouvements de la tête a son importance, puisque l'on n'est jamais parfaitement immobile, et que les petits mouvements impliquent de petites variations sur lesquelles le cerveau (système d'analyse différentielle par excellence) se base pour son analyse de localisation.

Codage de l'azimut par les différence interaurales L'amplitude et le temps de retard entre les deux canaux acoustiques permettent de coder l'azimut d'un son, mais pas son élévation ni sa distance (le son est localisé dans la tête de l'auditeur). Lorsque l'on fait dépendre les différences interaurales de temps (ITD) et d'intensité (IID) de la fréquence, on obtient une bien meilleure localisation. La diffraction du son dans l'air implique que ITD soit plus élevé en basses fréquences qu'en hautes fréquences. La limite basse fréquence est donnée par :

$$ITD = \frac{1.5\delta}{c} \sin\theta \tag{3.20}$$

avec θ l'angle d'incidence (rad), δ la distance interaurale (m) et c la célérité du son dans l'air (ms^{-1}). La limite est pour environ 1000 Hz : en dessous, ITD vaut environ -0.38 ms, au-dessus environ -0.26 ms. Pour la différence d'intensité IID, elle décroît de 0 dB à 0 Hz à -2 dB à 1000 Hz, puis ensuite régulièrement jusqu'à -10 dB pour 20 kHz.

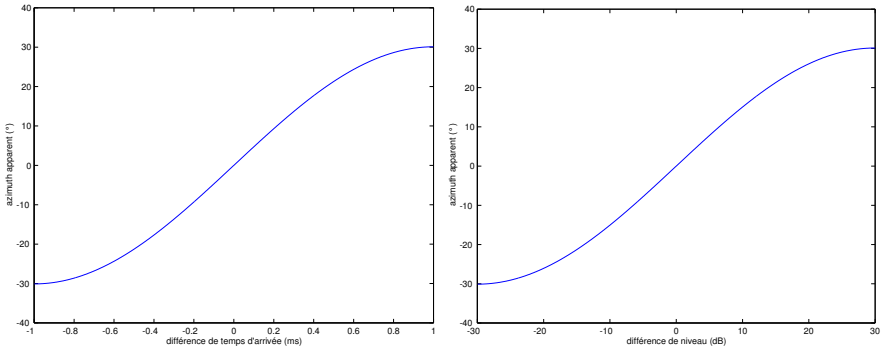


FIG. 3.20 – *azimut perçue en fonction de la différence de temps (gauche) et d'intensité (droite).*

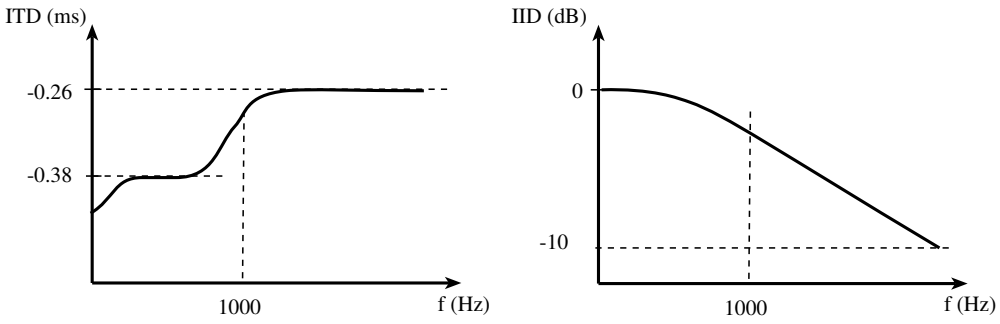


FIG. 3.21 – *Différence de temps (gauche) et d'intensité (droite) interaurale dépendant du temps.*

De plus, les différences interaurales changent avec la distance source-auditeur, du fait que les répartitions en fréquences sont modifiées par le filtrage dans l'air (c'est principalement audible pour les basses fréquences).

Externalisation Lorsque les signaux sonores arrivant aux deux oreilles sont parfaitement corrélés, le son est internalisé (entendu directement dans la boîte crânienne, à égale distance des deux oreilles). Par contre, lorsqu'il sont décorrélés, même partiellement (ce qui est le cas par exemple pour un son réverbéré : les signaux arrivant aux deux oreilles sont la somme de réflexions aux parois suite à des trajets différents), le son peut être externalisé (ceci ne fonctionne pas pour tous les auditeurs). La corrélation entre les deux signaux se mesure à l'aide de l'indice de corrélation :

$$r(\tau) = \lim_{T \to +\infty} \frac{1}{2T\nu} \int_{-T}^{+T} x_L(t)\, x_R(t+\tau) dt \tag{3.21}$$

avec τ la différence de temps entre les deux canaux et ν le facteur de normalisation défini comme suit :

$$\nu = \frac{1}{2T}\sqrt{\int_{-T}^{+T} x_L(t)^2 dt \int_{-T}^{+T} x_R(t)^2 dt} \tag{3.22}$$

Lorsque r vaut 1, les signaux sont parfaitement corrélés ; on les dit cohérents. Lorsque r vaut -1, les signaux sont en opposition de phase. Lorsque r vaut 0, les signaux sont non corrélés.

Il faut savoir que trois méthodes de décorrélation existent : la décorrélation par couple de filtres, la décorrélation par filtres passes-tout à pôles aléatoires et la décorrélation par réseau de délai à réinjection. Nous ne rentrerons cependant pas dans les détails de ces méthodes.

Fonctions de transfert relatives à la tête (HRTF) Comme expliqué précédemment, ces fonctions de transfert correspondent au filtre composé du torse, du cou et de la tête, ainsi qu'aux échos dus aux épaules et aux réflexions sur le cou. Ces fonctions de transfert sont personnelles, varient grandement d'une personne à l'autre. La manière de les calculer précisément pour une personne donnée consiste à introduire des microphones dans les conduits auditifs, et calculer la fonction de transfert réelle entre le signal envoyé sur les haut-parleurs et celui enregistré aux microphones. Il existe aussi des méthodes pour les modéliser [Rochesso, 2002].

3.5.9 Rendu 3D (distance, azimut et élévation) avec haut-parleurs

Le rendu de distance et d'espace en 3D à l'aide de haut-parleurs peut s'obtenir de trois manières : par la reconstruction holophonique (il s'agit de reproduire un champ acoustique 2D ou 3D par 3D-panning ou par procédé Ambisonic), par la technique transaurale (basée sur le binaural, et modifié de façon à fonctionner sur 2 haut-parleurs) ou par une méthode basée sur l'effet de précédence (un modèle de l'effet de salle est utilisé pour ajouter explicitement un délai sur l'une des voies : c'est moins précis en localisation mais aussi moins sensible aux changements de position de l'auditeur). Nous présentons 3 de ces techniques.

Localisation avec plusieurs haut-parleurs Chaque oreille de l'auditeur reçoit deux contributions : une contribution directe, du haut-parleur d'un côté vers l'oreille de ce même côté, et une contribution croisée (*cf. fig.* 3.22) :

$$x_G(t) = x_{GG}(t) + x_{DG}(t) \tag{3.23}$$
$$x_D(t) = x_{DD}(t) + x_{GD}(t) \tag{3.24}$$

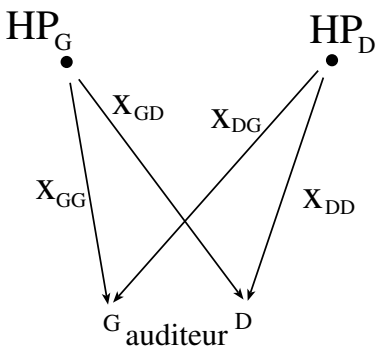


FIG. 3.22 – *Signaux reçus par chaque oreille de l'auditeur sur un système à deux haut-parleurs.*

Si on émet une sinusoïde dans les deux canaux, avec un pur gain A_D à droite et un pur retard τ_D, on entendra deux sinusoïdes distinctes, une à chaque oreille :

$$X_G(\omega) = 1 + A_D A_H e^{-j\omega(\tau_D+\tau_H)} \quad (3.25)$$
$$X_D(\omega) = A_D e^{-j\omega\tau_D} + A_H e^{-j\omega\tau_H} \quad (3.26)$$

avec A_H et τ_H le gain et le délai donnés par la HRTF relative à l'oreille contralatérale à la direction du haut-parleur. Remarquons qu'en basse fréquence, sachant que $A_H \approx 1$, un pur délai au haut-parleur sera perçu comme une pure différence de niveau, et réciproquement. En hautes fréquences, le masquage de la tête n'est plus négligeable, et les valeurs données A_H et τ_H peuvent provoquer un effet en contradiction avec ce que donneraient A_{HP} et τ_{HP} au niveau des haut-parleurs.

3D-panning Le 3D-panning consiste d'abord à utiliser un grand nombre de haut-parleurs situés en demi-sphère autour de l'auditeur. La localisation en azimut est donnée par une panoramisation (*panning*) entre les deux haut-parleurs les plus proches. La localisation en azimut et élévation nécessite quant à elle 3 haut-parleurs. La généralisation se fait mathématiquement par des opérations matricielles [Rochesso, 2002].

Transaural audio Le transaural audio consiste à utiliser un signal binaural, et à le faire passer au travers d'un ensemble de filtres pour corriger les contributions croisées des haut-parleurs, de façon à ce que chaque oreille n'entende que ce qu'elle aurait entendu en binaural. Les filtres utilisés font apparaître deux blocs fonctionnels : un filtre en treillis et un filtre en peigne. Le grand défaut de cette technique est qu'il faut que la tête soit positionnée à un endroit très précis, équidistant des haut-parleurs, et surtout ne doit pas en bouger. De plus, il ne faut pas de réverbération naturelle due à la salle (ce qui, en pratique, est approché en éloignant le plus possible les haut-parleurs des murs).

3.5.10 Directivité

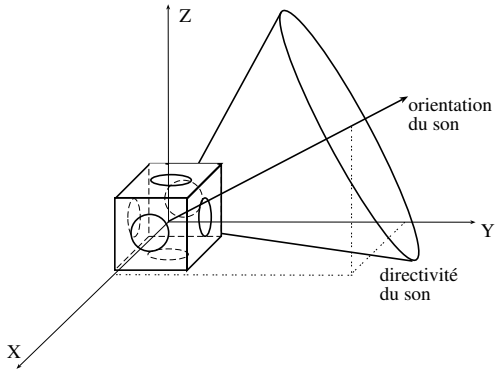


FIG. 3.23 – *Système de reproduction de la directivité et de la directiion d'une source monophonique ("La Timée" ou "LeCube").*

La directivité du rayonnement de sources sonores est propre à chaque type de source. Ainsi, une source ne distribue pas toutes ses fréquences de la même manière dans les différentes directions. Le rayonnement des sources est très rarement pris en compte, si ce n'est lors de l'enregistrement des sources. Mais lors de la restitution, il dépend de la structure vibrante qu'est le haut-parleur, considéré comme une source transparente ponctuelle (onde sphérique) ou plane (onde plane). Cependant, un haut-parleur n'est jamais transparent, aussi il introduit une distorsion à la fois spatiale et fréquentielle.

Le problème consiste à reconstruire le champ acoustique en 3D. Une solution proposée par [Misdariis *et al.*, 2001; Warusfel and Misdariis, 2001] consiste à décomposer la directivité à reproduire sur la base des harmoniques sphériques. Le signal est diffusé sur un système de 6 haut-parleurs disposés sur les faces d'un cube, après traitements différenciés. Par combinaisons linéaires des harmoniques sphériques (modes simples de directivité), on peut reconstruire n'importe quelle directivité pour le système.

3.6 Traitements modifiant le timbre

Les traitements portant sur le timbre existent en très grand nombre. Nous allons en aborder quelques uns, parmi les plus "classiques", à savoir : le chorus, le *flanger*, le *phasing*, la distorsion non-linéaire, les modifications de l'enveloppe spectrale (dont le changement de formants), l'équalisation, le filtrage en peigne, la synthèse croisée, le *morphing*, l'interpolation spectrale, la robotisation et enfin l'effet wha-wha.

3.6.1 Effets sur l'enveloppe

3.6.1.i) Modifications de l'enveloppe spectrale

Modifier l'enveloppe spectrale $\mathcal{E}(f)$ peut se faire de plusieurs manières : soit en la décalant, soit en la dilatant/contractant. Nous ne parlerons pas ici des autres manières de procéder à partir de deux sons, telle l'interpolation spectrale ou la synthèse croisée abordées plus loin (*cf.* 3.6.4).

Décalage de l'enveloppe spectrale Cela consiste à décaler de d Hz la partie en fréquences positives $\mathcal{E}^+(f)$ de l'enveloppe du spectre de magnitude $\mathcal{E}(f)$ vers les basses fréquences ou vers les hautes fréquences. Quelque soit le modèle utilisé pour l'estimation de l'enveloppe spectrale (LPC, cepstre) et pour la modification (modèle additive, vocodeur de phase, modèle soustractif), un décalage de l'enveloppe nécessite des précautions quand aux bornes. En effet, la plage de variation de l'enveloppe décalée est de $[d; d+F_e/2]$. il faut donc appliquer une troncature de cette enveloppe, pour conserver à la fois la symétrie et l'intervalle de variation $[0; F_e/2]$:

$$\mathcal{E}^+(f) = \mathcal{E}^+(f+d)\mathbb{1}_{\mathcal{E}^+(f+d)\in[0;F_e/2]} \tag{3.27}$$

Une fois tronquée cette courbe, on peut recomposer l'enveloppe des fréquences négatives comme étant sa symétrique. Pour le vocodeur de phase, si l'on désire décaler d'un nombre entier de paniers de fréquence, aucune interpolation n'est nécessaire. Par contre, des lorsque l'on veut décaler l'enveloppe d'un nombre réel de paniers de fréquences, on peut interpoler l'enveloppe (par interpolation linéaire, ou par splines). Cela dit, l'oreille n'est pas assez sensible pour entendre de si fines variations sur l'enveloppe, sauf si celle-ci est très spéciale (par exemple un filtre passe-bande très étroit).

Dilatation/contraction de l'enveloppe spectrale Cela consiste à modifier l'échelle de l'enveloppe par un facteur multiplicatif γ constant. Il convient de prendre les mêmes précautions que précédemment. La nouvelle enveloppe est calculée par interpolation de l'enveloppe du spectre original (par interpolation linéaire ou splines).

$$\mathcal{E}^+_\gamma(f) = \mathcal{E}^+(\gamma f)\mathbb{1}_{\mathcal{E}^+(\gamma f)\in[0;F_e/2]} \tag{3.28}$$

Cette transformation permet d'obtenir entre autres un effet "Donald Duck", du fait que les formants ne correspondent plus, après transformation, à la manière dont un être humain aurait produit le son. La dilatation/contraction de l'enveloppe spectrale peuvent s'obtenir de manière involontaire (changement de hauteur ne conservant par l'enveloppe spectrale, *cf.* 3.4.2) ou volontairement comme indiqué ci-dessus, par les modèles additifs, vocodeur de phase ou par changement d'échelle temporelle PSOLA puis ré-échantillonnage.

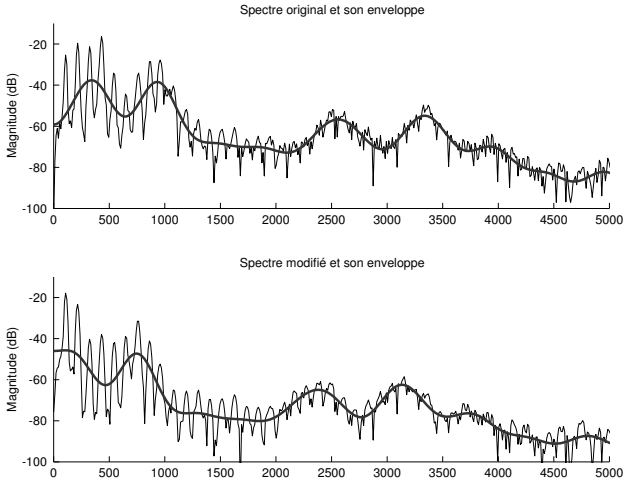


FIG. 3.24 – *Décalage de l'enveloppe spectrale : spectre et son enveloppe avant (*fig. *haut) et après (*fig. *bas) décalage de* $d = -200\ Hz$. *[decEnv-200Hz]*

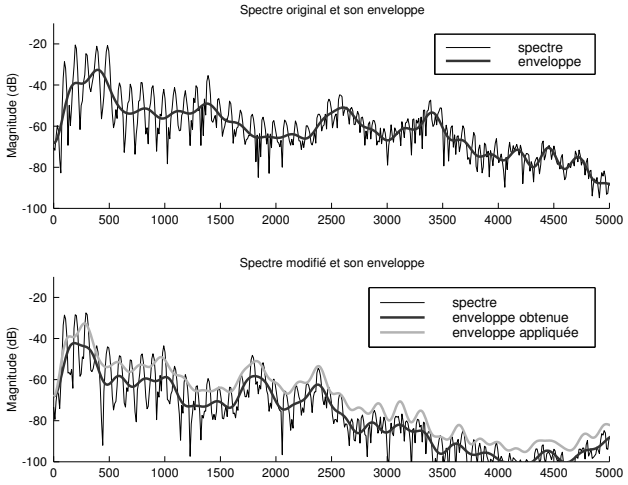


FIG. 3.25 – *Dilatation/contraction de l'enveloppe spectrale : spectre et son enveloppe avant (*fig. *haut) et après (*fig. *bas) étirement d'un facteur* $\gamma_{env} = 0.7$.

3.6.1.ii) Equaliseur

Un équaliseur est un outil de transformation du spectre de magnitude d'un signal. Etant donné un ensemble de filtres, chacun s'appliquant à une bande de fréquences précise, l'équalisation consiste à donner des gains à chaque filtre, afin d'amplifier ou diminuer chaque bande de fréquence. Le spectre de magnitude est décomposé en bandes de fréquences à l'aide de filtres : des filtres à étages aux extrémités, et des filtres en pics entre les deux extrémités. Chaque filtre est définit par son type (passe-bas ou passe-haut, en pic), son ordre (premier ou second), sa fréquence de coupure f_c^i, son gain G_i. Les filtres en pic (ou passe-bande) possèdent un paramètre supplémentaire, la largeur de bande f_b. Des filtres passe-bande couramment utilisés sont les filtres à Q constant : Q est le facteur qualité, obtenu par le rapport $Q = \frac{f_b}{f_c}$. Conserver Q constant revient à augmenter la largeur de bande lorsque la fréquence centrale du filtre augmente. En échelle logarithmique, cela donne des filtres à largeur de bande constante (en $log(Hz)$). Tous ces paramètres sont manipulés à l'aide de potentiomètres (solution matérielle) ou d'une interface graphique faite de potentiomètres (solution logicielle).

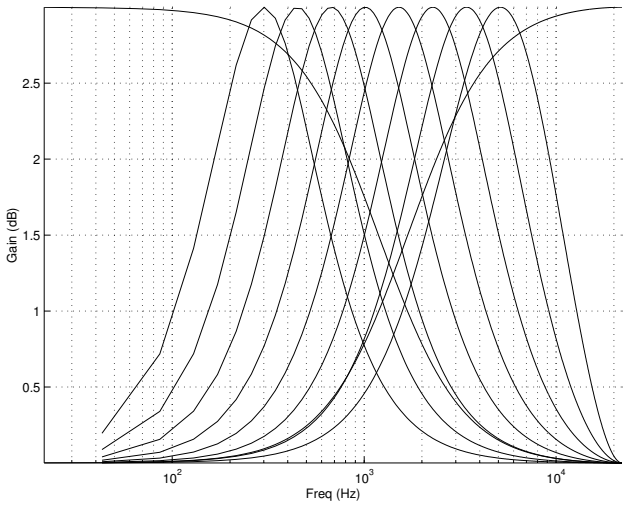


FIG. 3.26 – *Equaliseur : filtres à étage aux extrémités et filtres en pic au milieu.*

Une mise en œuvre alternative consiste à utiliser le vocodeur de phase, et à modifier les paniers de fréquences par groupes afin de respecter le Q constant. Il faut dans cas appliquer une fenêtre à la resynthèse avant l'ajout-superposition pour éviter les effets de repliement (*aliasing*) temporel.

3.6.1.iii) Filtrage en peigne

Le filtrage en peigne consiste à filtrer le son avec une sorte de peigne harmonique, soit pour augmenter, soit pour diminuer les modules du spectre aux multiples d'une pseudo-fondamentale. Le filtre peut être RIF, RII ou universel.

Filtrage à réponse impulsionnelle finie Le filtre en peigne à réponse impulsionnelle finie a été présentée en 2.2.2. Rappellons que a réponse en fréquence de ce filtre ressemble à un peigne et que les trous dans le spectre sont de plus en plus prononcés lorsque le gain du filtre augmente en valeur absolue.

Filtrage à réponse impulsionnelle infinie (RII) Soit $x(n)$ un signal entrant dans une ligne à retard de longueur L. Si l'on réinjecte le signal en sortie de la ligne à retard dans la ligne à retard, avec g_y le gain, et que l'on ajoute ce signal au son d'entrée de la ligne à retard, on obtient un signal :

$$y(n) = \beta\ x(n) + g_y\ y(n - L) \tag{3.29}$$

correspondant au filtrage en peigne à réponse impulsionnelle infinie (RII), de fonction de transfert :

$$H(z) = \frac{\beta}{1 - g_y\ z^{-L}} \tag{3.30}$$

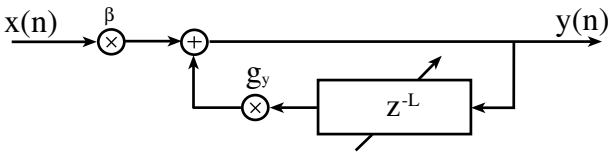


FIG. 3.27 – *Diagramme du filtre RII en peigne.*

La boucle de réinjection implique une réponse impulsionnelle infinie. Le filtre n'est stable que pour $|g_y| \leq 1$. Dans le cas où le gain g_y est positif, le filtre amplifie toutes les fréquences multiples

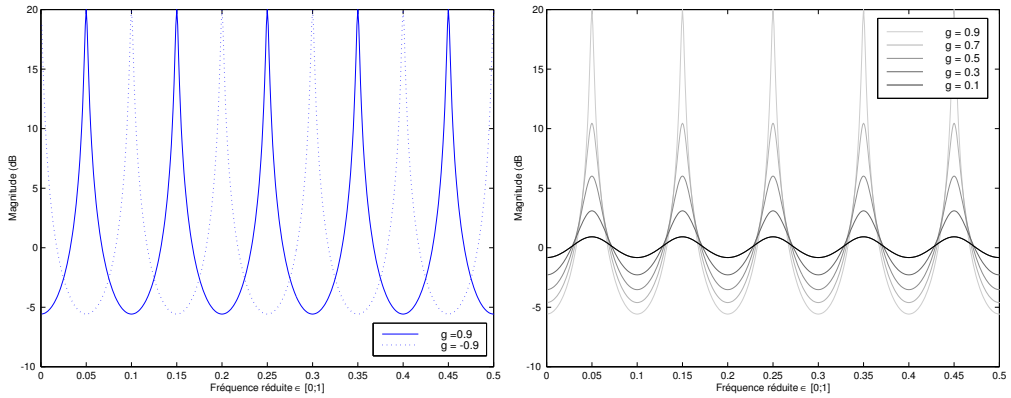


FIG. 3.28 – *Réponse en fréquence d'un filtre en peigne FIR : à gauche pour $g_y = 0.9$ et $g_y = -0.9$, à droite pour $g_y = 0.1$, $g_y = 0.3$, $g_y = 0.5$, $g_y = 0.7$, $g_y = 0.9$.*

de $\frac{1}{\tau} = \frac{F_e}{L}$, et atténue toutes les fréquences entre ces multiples. Le gain varie en amplitude entre $\frac{1}{1-g_y}$ et $\frac{1}{1+g_y}$ (soit en décibels entre $-20\log_{10}(1+g_y)$ et $-20\log_{10}(1-g_y)$ dB). Les pics sont de plus en plus étroits et amplifiés lorsque le gain $|g_y|$ augmente en valeur absolue (*cf.* *fig.* 3.28).

Filtre en peigne universel Il s'agit de la combinaison des filtres RIF et RII. Le signal en sortie d'un filtre en peigne universel est :

$$y(n) = \beta\ x(n) + g_x\ x(n-L) + g_y\ y(n-L) \tag{3.31}$$

Sa fonction de transfert est :

$$H(z) = \frac{\beta + g_x\ z^{-L}}{1 - g_y\ z^{-L}} \tag{3.32}$$

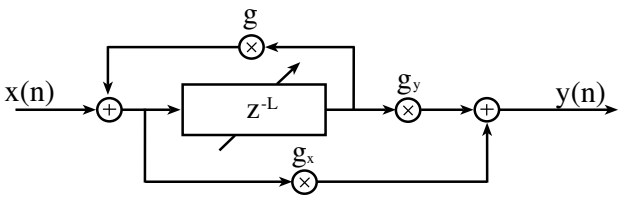


FIG. 3.29 – *Diagramme du filtre en peigne universel.*

Selon les valeurs des coefficients g, g_x et g_y, on obtient différentes configurations (*cf.* 3.2).

Filtre / gains	g	g_x	g_y
RIF	X	0	Y
RII	1	X	0
passe-tout	a	$-a$	1
délai	0	0	1

TAB. 3.2 – *Filtre en peigne RIF, RII et universel en fonction des valeurs des coefficients g, g_x et g_y.*

Intérêt de cet effet Cet effet de filtrage en peigne permet de donner une hauteur à un son (bruit) n'en ayant pas. En effet, l'ajout de pics équi-répartis revient à ajouter des harmoniques ; ce qui est

le plus surprenant, c'est que si on filtre un bruit blanc par un filtre en peigne vers le bas, c'est-à-dire si l'on réalise des trous équi-répartis dans le spectre, une fois encore, on impose une hauteur clairement audible (plus qu'une simple coloration) au son.

3.6.1.iv) Filtre résonant

Le filtrage résonant consiste à utiliser un filtre en pic tels ceux utilisés pour l'équaliseur, mais plus sélectif, de manière à faire amplifier une zone de fréquences particulière. On peut aussi utiliser un filtre en pic très sélectif pour supprimer une fréquence gênant, par exemple un bruit proche de 50 *Hz* provoqué par l'installation électrique.

3.6.1.v) Wha-wha

L'effet wha-wha est utilisé principalement par les guitaristes. Il consiste, par le biais d'une pédale contrôlant un filtre paramétrique (via sa fréquence centrale ou de résonance) passe-bande de petite bande passante. La fréquence de résonance est placée à la fréquence du premier formant de la voyelle [a] ou [u]. Le signal filtré est ensuite mixé avec le signal direct, comme indiqué *fig.* 3.30.

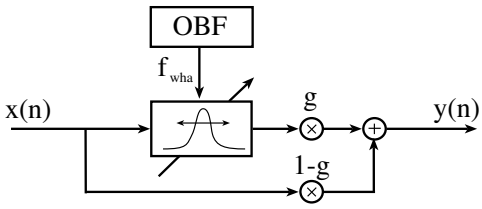


FIG. 3.30 – *Diagramme de l'auto-wha.*

Lorsque la fréquence de résonance est donnée par un oscillateur basse-fréquence (comme indiqué sur le diagramme), il s'agit alors d'un auto-wha, ou effet wha-wha automatique. Un autre type d'effet wha-wha est celui qui se déclenche à l'attaque : lorsque le son s'enrichit en hautes fréquences (ce qui est le cas lors des attaques des notes de guitares, par exemple), on déclenche le mouvement rapide du filtre de la position donnant le son [u] vers le son [a]. Le retour à la position de repos (donnant le son [u]) se fait à une vitesse différente. De même que pour le compresseur ou l'expandeur, on utilise deux temps : le temps de montée ou d'attaque, et le temps de descente, ou de relaxation. On dit dans ce cas que l'effet wha-wha est sensitif ; c'est un effet adaptatif.

3.6.1.vi) Filtrage avec résolution arbitraire

Le filtrage avec résolution arbitraire [Amatriain *et al.*, 2001; Amatriain *et al.*, 2002] consiste à utiliser un modèle temps-fréquence (vocodeur) ou spectral (modèle additif) pour réaliser un filtrage directement dans le domaine fréquentiel. La convolution de la réponse impulsionnelle du filtre avec le signal en temporel correspond à la multiplication du spectre du signal par celui du filtre. Ceci permet de choisir la résolution que l'on désire, au panier de fréquence près pour le vocodeur de phase, et sans limite de précision pour le modèle additif (du moins pour la partie harmonique, la partie résiduelle étant filtrée par TFCT) [Amatriain *et al.*, 2001]. On définit le filtre par sa TFCT (fonction de transfert) :

$$H(f) = \sum_{k=1}^{H} \delta(f_k).g_k \tag{3.33}$$

avec f_k les fréquences des partiels, g_k les gains à appliquer à chaque partiel. Le nouveau spectre est obtenu par produit avec les amplitudes respectives des différents partiels A_k :

$$\tilde{X} = H.X = \sum_{k=1}^{H} A_k.g_k \tag{3.34}$$

Concernant le filtrage par TFCT, la multiplication des TFCT du signal et du filtre équivaut à une convolution circulaire du signal temporel, aussi il y a un effet de repliement temporel du signal. Pour l'éviter, il convient au préalable d'appliquer un bourrage de zéros à $x(n)$ et $h(n)$ après fenêtrage, de façon à doubler la taille des TFCT. On évite ainsi l'effet de repliement temporel. Une solution pratique à ce problème consiste à utiliser un fenêtrage temporel lors de la synthèse par ajout-superposition : l'effet de repliement temporel est alors corrigé.

3.6.2 Effets sur la phase

Les effets sur la phase sont le ***flanger***, le chorus et le ***phaser***, basés sur la modulation de lignes à retard.

3.6.2.i) Flanger

Le ***flanger*** (ou ***flanging***) est un effet de modulation de la fréquence d'un signal sonore [Orfanidis, 1996; Dattoro, 1997]. A l'origine, il provenait de la lecture simultanée d'un même signal sur deux appareils analogiques différents : la très légère différence de vitesse de défilement est à l'origine de cet effet. Numériquement, il s'obtient à l'aide d'une ligne à retard, et est basé sur le même principe que le chorus (*cf.* 3.6.2). L'effet perceptif est assez saisissant, et certains se plaisent à y entendre le son d'un avion à réaction. Le nom de ***flanging*** vient de l'anglais "***flange***", qui signifie "bord", à savoir le bord de la bobine sur laquelle il faut appuyer légèrement pour créer l'effet.

Le ***flanger*** s'obtient en dupliquant le signal d'entrée à l'aide d'une ligne à retard, dont le temps de retard $\tau_{flanging}$ varie avec le temps (c'est un oscillateur à basse fréquence, OBF, ou ***LFO*** pour ***Low Frequency Oscillator***), *cf. fig.* 3.31.

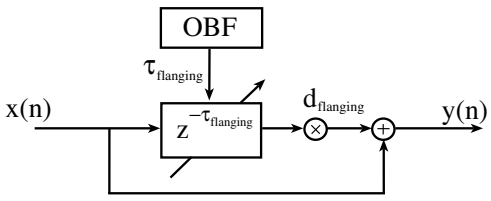


FIG. 3.31 – *Diagramme du* flanger.

Le délai est trop court pour être entendu comme tel (inférieur à 50 millisecondes, alors que l'oreille humaine entend un délai comme tel à partir de 50 à 70 millisecondes). Par contre, le retard crée un effet de filtrage en peigne, en créant un série de nœuds dans le spectre en fréquences du signal. Les nœuds correspondent à des fréquences où le module du spectre est nul et sont espacés régulièrement. Ces nœuds résultent d'interférences destructives entre les deux signaux additionnés, lorsque deux composantes sinusoïdales s'additionnent avec quasiment le même module et en opposition de phase. La profondeur $d_{flanging}$ de l'effet correspond au taux de mixage de l'effet avec le son original. Pour une profondeur nulle, aucun signal n'est ajouté, donc aucun nœud n'apparaît. Pour une profondeur maximale (égale à 1), les nœuds descendent jusqu'à atteindre 0 dans la réponse en fréquence (soit $-\infty$ dB). Le temps de retard varie selon un oscillateur à basses fréquences. Ce peut être un oscillateur sinusoïdal, triangulaire, logarithmique (*cf. fig.* 3.32).
Les paramètres de contrôle de cet effet concernent principalement la forme d'onde. Elle se définit par l'amplitude de la forme d'onde $b_{flanging}$ (ou profondeur de balayage), qui s'additionne au délai

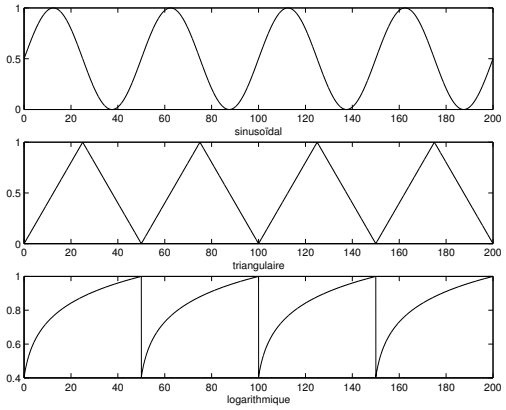


FIG. 3.32 – *Types d'oscillations généralement utilisées pour l'effet de* flanging *: sinusoïdal, triangulaire, logarithmique. Le temps est indiqué en* ms.

fixe $\tau_{flanging}$ pour donner le délai variable qui contrôle la ligne à retard ; par la profondeur ou le taux de mixage $d_{flanging}$ variant entre 0 et 1 et faisant varier l'acuité des nœuds ; par la forme de l'oscillateur basse fréquence ; par sa fréquence, ou plutôt son inverse, le taux $t_{flanging}$ et par le gain de réinjection $G_{flanging}$. Ce gain peut être positif ou négatif (entre -1 et 1), de façon à pouvoir supprimer où ajouter de l'effet.

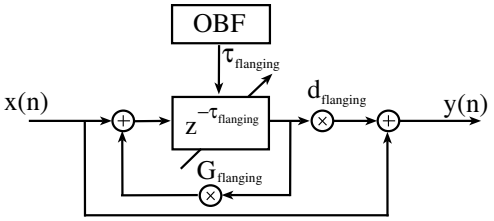


FIG. 3.33 – *Diagramme du* flanger *avec réinjection.*

Pour créer un flanger stéréophonique, on utilise deux ***flangers*** en quadrature de phase (déphasage de $\pi/2$ entre les deux formes de l'oscillateur basses fréquences).

3.6.2.ii) Chorus

Le "chorus" est un choeur, un groupe de chanteur. L'effet numérique correspondant cherche donc à ajouter de l'épaisseur au son, à l'enrichir en donnant la sensation de plusieurs instrumentistes jouant simultanément la même partition. L'utilisation de plusieurs chorus simultanés rend l'effet encore plus intéressant. Il remplace à sa manière les méthodes des luthiers par exemple, qui doublaient les cordes (guitare 12 cordes, mandoline, banjo) afin de créer cet effet chorus.

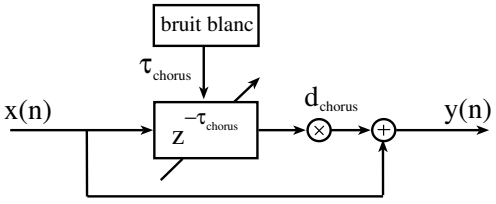


FIG. 3.34 – *Diagramme du chorus.*

L'algorithme utilisé pour l'effet de chorus est très simple, et partagé (à une éventuelle complication près) par le ***flanger*** (*cf.* 3.6.2). Lorsque deux personnes jouent en même temps la même

ligne mélodique, ils ne sont pas tout à fait à l'unisson : ni parfaitement synchrones, ni parfaitement accordés. Le retard entre les deux instruments se modélise avec une ligne à retard. Lorsque le retard de cette ligne varie au cours du temps, on reproduit l'effet de léger désaccord entre le son original et le son traité par transposition du son original [Orfanidis, 1996; Dattoro, 1997]. Le diagramme du chorus est donné *fig.* 3.34.

Pour créer un chorus multiple (par exemple stéréophonique, *cf. fig.* 3.35), il suffit de multiplier les étages de chorus, et de définir pour chaque étage le taux de mixage, la forme de l'oscillateur, les déphasages.

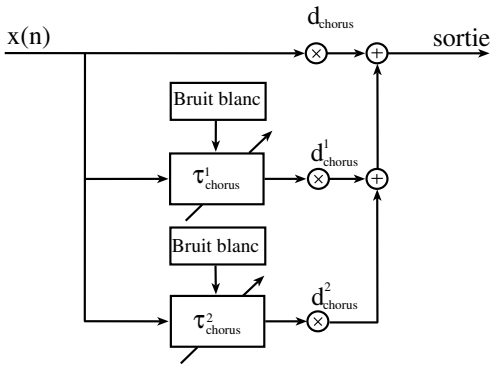


FIG. 3.35 – *Diagramme du chorus stéréophonique.*

Il ne reste plus qu'à mixer le son original et le son traité pour obtenir le chorus. La structure proposée est quasiment la même que celle du ***flanger***. Les différences principales entre le ***flanger*** et le chorus sont l'absence de réinjection pour le chorus et la plage de valeurs possible pour le temps de retard (20 à 30 ms pour le ***flanger***, contre 1 à 10 ms pour le chorus). La transposition est contrôlée par un bruit blanc (au lieu d'un oscillateur dans le cas du ***flanger***).

3.6.2.iii) Phaser

Le terme ***phasing*** vient de la contraction de l'expression ***phase shifting***, signifiant décalage de phase ou déphasage. Cet effet consiste à effectuer un filtrage du spectre du son en créant des nœuds, c'est-à-dire donnant une valeur nulle au module du spectre pour des fréquence données [Bartlett, 1970; Hartmann, 1978; Smith, 1984]. Le ***flanger*** (*cf.* 3.6.2) et le chorus (*cf.* 3.6.2) sont des cas particulier du ***phaser*** où les nœuds forment un peigne harmonique, de fondamentale variable [Orfanidis, 1996; Dattoro, 1997]. Les nœuds que l'on désire créer dans le spectre d'amplitude du son sont le plus souvent mis en œuvre à l'aide de filtres passe-tout (*cf. fig.* 3.36).

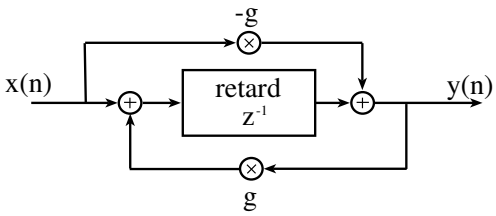


FIG. 3.36 – *Diagramme du filtre passe-tout.*

Ces filtres ne modifient en rien le spectre d'amplitude ; par contre, ils introduisent des déphasages (modifications dans le spectre de phase). Le gain est unitaire pour toute les fréquences, mais la réponse en phase est non-linéaire. De ce fait, lorsque l'on va additionner le son original avec celui contenant des retards de phase, on va obtenir un son dont le spectre d'amplitude contiendra des trous aux noeuds désirés, par addition de composantes en opposition de phase.

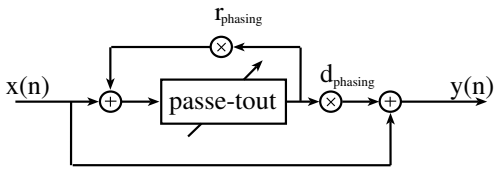


FIG. 3.37 – *Diagramme du* phasing.

Tout comme pour le chorus et le ***flanger***, on définit la profondeur de l'effet $d_{phasing}$ comme étant le taux de mixage de l'effet avec le son original (entre 0 et 100 %). La plage de balayage $b_{phasing}$ correspond à la largeur en fréquence (Hz) sur laquelle chaque nœud peut se déplacer autour d'une valeur fixe (parfois donnée par l'utilisateur). Le gain de réinjection $r_{phasing}$ correspond encore à la même chose que pour les effets précédemment décrits. La fréquence de balayage $t_{phasing}$ correspond à la fréquence à laquelle les nœuds se déplacent (nombre de déplacement aller-retour par seconde).

Un ***phaser*** stéréophonique se réalise de même qu'un chorus ou un ***flanger*** stéréophonique, à l'aide de deux filtres passe-tout, et un mixage de gain croisé entre les deux lignes (*cf. fig.* 3.38).

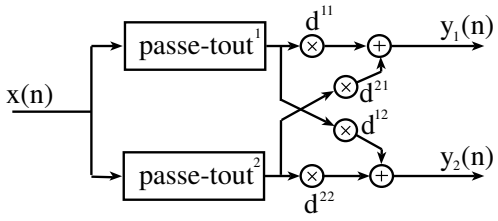


FIG. 3.38 – *Diagramme du* phasing *stéréophonique.*

3.6.3 Effets sur le spectre et sa structure

3.6.3.i) Décalage du spectre, avec ou sans duplication

Le décalage du spectre consiste à prendre un spectre de magnitude $\rho(f)$, et à le décaler en fréquences $\rho_\delta(f) = \rho(f+\delta)$ de δ Hz. Pour de très faibles décalages, il permet de modifier uniquement le spectre (et donc le timbre) : lorsque la fréquence fondamentale n'est pas trop différente de l'écart entre deux partiels successifs, pour peu qu'il soient bien équi-répartis, la hauteur entendue est celle de l'écart entre les harmoniques. Soient $f_{a,i} = if_0$ les fréquences des harmoniques analysées. Pour δ petit, le son composé des partiels de fréquences $f_{s,i} = if_0 + \delta$ aura pour hauteur f_0. Par contre, dès que δ atteint quelques Hertz, ce n'est plus vrai, le son devient inharmonique, dissonant, le timbre est modifiée et la hauteur n'est plus perçue. C'est encore plus sensible pour les modulations en anneau, du fait que deux composantes proches à la place de chaque harmonique provoquent des battements.

Lorsque l'on dispose d'une représentation additive du signal, en décaler le spectre consiste à en décaler toutes les composantes : l'enveloppe spectrale du résidu, celle du transitoire, et chacune des harmoniques. Cela se réalise très aisément. Pour un signal traité avec le vocodeur de phase, c'est un peu plus délicat car le décalage seul des magnitudes ne suffit pas, il faut aussi recalculer les phases. Lorsqu'on reste dans le domaine temporel, le modulateur à bande latérale unique (BLU) permet de réaliser ce décalage aisément, à ceci près qu'il y a repliement du spectre, ce que l'on peut éviter avec le modèle additif et avec le vocodeur de phase. Avec le modulateur en anneau, on réalise une copie du spectre, et chacune des deux copies est décalée en sens inverse, l'une vers les fréquences positives, l'autre vers les fréquences négatives.

Dans tous les cas, on fera attention à la plage de validité des fréquences (de même que pour les décalages et étirements/contractions de l'enveloppe spectrale) que l'on sait respecter avec le

vocodeur de phase et le modèle additif : les fréquences négatives obtenues doivent être éliminées, de même que celles supérieures à la demi-fréquence d'échantillonnage.

3.6.3.ii) Vibrato

Le vibrato est une lente modulation de fréquence, généralement comprise entre 4 et 8 Hz. C'est une spécificité de la voix chantée, un de ses attributs perceptifs. En effet, un vibrato est perçu comme un mode de jeu et non comme une modulation de fréquence : le système auditif le traduit comme attribut perceptif. La plupart des instruments utilisent le vibrato dans leurs modes de jeu (vents, cuivres, cordes frottées ou pincées comme la guitare). Il s'accompagne généralement d'un trémolo (modulation d'amplitude) de même fréquence et de profondeur de l'ordre de 3 dB.

Quelle que soit la méthode de transposition utilisée (PSOLA, vocodeur de phase, modèle spectral, avec ou sans conservation des formants), le facteur de transposition est donné par :

$$\gamma_{hauteur}(n) = 1 + d_{vib}.\sin(2\pi\ f_{vib}n) \tag{3.35}$$

avec d_{vib} la profondeur du vibrato, de l'ordre de 10 cent (soit 10/200 ton), et f_{vib} la fréquence de la modulation.

Il a été montré que le vibrato, en produisant un balayage de l'enveloppe spectrale, permet au système auditif de reconnaître plus facilement les formants d'un son. Aussi, il est préférable d'utiliser une méthode de changement de hauteur qui conserve les formants, de manière à reproduire fidèlement cette mise en relief, ce parcours de l'enveloppe spectrale.

3.6.4 Effets sur le spectre et l'enveloppe

3.6.4.i) Distorsion non-linéaire

La distorsion non-linéaire d'un signal audio ajoute des harmoniques au signal ; c'est un effet indésirable dans une installation haute-fidélité, et très désirable dans des situations musicales. La distorsion s'obtient en faisant circuler le signal à travers une fonction de conformation non-linéaire :

$$y(n) = f(x(n)) \qquad \text{avec } f \text{ non-linéaire} \tag{3.36}$$

La fonction de conformation f, non linéaire, fait apparaître dans le spectre des composantes qui n'y existaient pas.

Si le signal $x(n)$ est harmonique, il peut être décomposé en une somme de sinusoïdes $x(n) = \sum_{i=-p}^{p} e^{j\omega_i n}$. Si la fonction f se développe en série de Taylor-MacLaurin autour de 0 :

$$f(u) = f(0) + u\ f'(0) + \frac{u^2\ f''(0)}{2!} + \frac{u^3\ f'''(0)}{3!} + \ldots = \sum_{i=0}^{\infty} \frac{u^i}{i!}\frac{d^i f}{du^i} \tag{3.37}$$

alors on remarque que le signal $x(n)$ passé à travers la fonction f apparaît comme une somme pondérée de ses puissances entières, on obtient donc :

$$y(n) = \sum_{i=0}^{\infty} \alpha_i (x(n))^i \tag{3.38}$$

avec $(x(n))^i = \sum_{-p \le j(k) \le p} e^{j(\omega_{j(0)} + \omega_{j(1)} + \ldots + \omega_{j(i-1)})}$.

Pour un signal composé de sinusoïdes réelles, les termes ω_i apparaissent par paires de signes opposés. Par suite, le signal de sortie $y(n)$ contient des composantes sinusoïdales dont les fréquences sont toutes les sommes et les différences des multiples des fréquences originales ω_i. Ainsi, un couple de sinusoïdes $x(n) = \cos(\omega_1 n + \Phi_1) + \cos(\omega_2 n + \Phi_2)$ donnera en sortie un signal qui en général contiendra les fréquences ω_1, ω_2, $(\omega_1 + \omega_2)$, $(\omega_1 - \omega_2)$, 2 ω_1, 2 ω_2, etc. Si $x(n)$ est périodique, alors $y(n)$ est également périodique mais de contenu spectral différent.

Une manière de mettre en œuvre cette distorsion non linéaire consiste à utiliser les **polynômes de Chebyshev** [Arfib, 1979]. Ces polynômes T_i possèdent la propriété suivante :

$$T_i(\cos\theta) = \cos(i\theta) \tag{3.39}$$

et se calculent par récurrence selon la formulation :

$$T_0(x) = 1 \tag{3.40}$$
$$T_1(x) = x \tag{3.41}$$
$$T_{n+1}(x) - 2x\,T_n(x) + T_{n-1}(x) = 0 \tag{3.42}$$

La propriété (3.39) est très intéressante, car elle permet de doser le degré de distorsion obtenu par le système et de maîtriser le spectre résultant. En choisissant pour $f(u)$ une somme pondérée de polynômes de Chebyshev, on fait apparaître dans le signal $f(\cos(\omega n))$ une somme pondérée de certaines fréquences multiples de ω. On peut ainsi contrôler le nombre et les amplitudes des harmoniques qui apparaissent après transformation (*cf.* [Arfib, 1979], [Brun, 1979]). L'un des intérêts de ces fonctions est que le signal de sortie est limité en fréquences : aucun repliement du spectre n'est possible.

Puisque la fonction $f(u)$ est non-linéaire, la distorsion générée par son application au signal $x(n)$ dépend fortement de l'amplitude maximale du signal : pour un signal de très faible niveau, $f(u)$ peut être considérée comme linéaire (si son terme constant est nul), et la distorsion devient faible. Dans le cas contraire, le niveau du son de sortie n'augmente pas linéairement avec le niveau d'entrée, mais le spectre s'enrichit. Cette sensibilité de la distorsion à l'amplitude peut modéliser les caractéristiques d'instruments acoustiques. En effet, on peut jouer d'un instrument acoustique de façon plus "dure", par exemple en pinçant une corde de guitare avec force, en soufflant de façon stridante dans un saxophone, ou en frappant sur un tambour avec intensité, ce qui enrichit le spectre en hautes fréquences. Avec la distorsion non-linéaire, nous pouvons simuler cet effet en faisant passer le signal dont l'amplitude varie à travers une fonction non-linéaire, obtenant un signal de sortie dont le spectre varie.

Des études (*cf.* [Brun, 1979]) ont été menées pour corriger l'intensité du son de sortie, de manière à la normaliser. Ainsi, on peut s'assurer qu'en plus de l'enrichissement spectral produit par la distorsion, l'intensité de sortie est identique à l'intensité d'entrée.

A partir du moment où l'on utilise des fonctions autres que les polynômes de Chebyshev, plus rien n'assure que le son produit soit harmonique lorsque le son d'entrée l'est. Ainsi, des combinaisons inharmoniques de partiels ainsi que de structures formantiques peuvent être obtenues (*cf.* [Arfib, 1979]). La fonction de conformation peut être donnée graphiquement, évoluer dans le temps. Dans ces cas, il est nécessaire de s'assurer qu'aucun repliement n'a lieu, à l'aide d'un filtrage passe-bas à la demi-fréquence d'échantillonnage.

3.6.4.ii) Conformation spectrale

Principe Il s'agit d'appliquer une transformation au spectre, par exemple dans le modèle du vocodeur de phase. Les phases sont calculées selon le modèle du vocodeur de phase, sans chercher à conserver l'enveloppe.

Si la courbe de conformation est de la forme $y = ax$ cette transformation est un étirement linéaire du spectre d'amplitude, c'est-à-dire une sorte de transposition sans conservation de l'enveloppe spectrale, mis à part que la largeur de bande des pics correspondant aux harmoniques n'est plus respectée. Si la courbe est de la forme $y = x + b$, cette transformation est un décalage spectral. Par contre, si le spectre de magnitude est modifié de manière non linéaire, le résultat sonore plus difficilement compréhensible.

Remarque : il est possible de n'appliquer la conformation qu'à l'enveloppe, en utilisant le modèle source-filtre. Nous n'en avons pas trouvé d'indication dans la littérature. Néanmoins, nous le proposons dans les effets adaptatifs, *cf.* sec. 5.6.1.

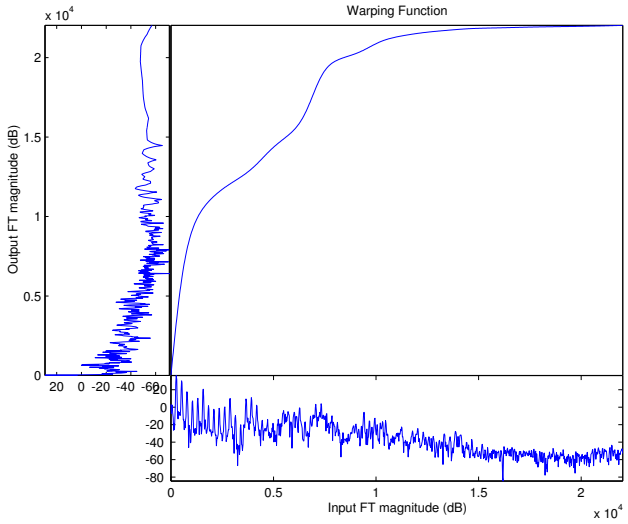


FIG. 3.39 – *Conformation spectrale : spectre original, fonction de conformation et spectre transformé.*

Un cas particulier de la conformation spectrale est le changement d'échelle fréquentielle dépendant du partiel réalisé avec le modèle additif [Amatriain *et al.*, 2001]. En réalisant une transposition dépendant du partiel :

$$f_k = f_k.\gamma^{(k-1)} \tag{3.43}$$

, on introduit une légère inharmonicité, comme celle des sons d'instruments à cordes (tel que le piano). Cet effet est adaptatif, du fait que la valeur du facteur de transposition dépende du numéro du partiel.

3.6.4.iii) Mutation, synthèse croisée, *morphing*, interpolation spectrale

Les effets de synthèse croisée, de ***morphing*** ou de mutation, et d'interpolation spectrale semblent correspondre à la même chose, et les auteurs utilisent parfois l'un pour l'autre. Cependant, il existe des différences, que nous allons préciser.

Mutation La mutation correspond à la création d'un son à partir des spectres d'amplitude ρ_1, ρ_2 et de phase Φ_1, Φ_2 de deux sons : le spectre d'amplitude de synthèse ρ_s et celui de phase Φ_s sont donnés arbitrairement, et le son est reconstruit par FFT inverse. Le modèle sous-jacent est le modèle temps-fréquence (vocodeur de phase). La transformée de Fourier résultant est donnée simplement par :

$$Y(f) = \rho_s(f).e^{j\Phi_s(f)} \tag{3.44}$$

La représentation de Fourier à court-terme obtenue n'est pas forcément valide. De ce fait, pour imposer une représentation valide, on calcule la transformée de Fourier inverse, puis on lui applique la transformée de Fourier : les spectres d'amplitude et de phase sont alors valides mais différents[2].

$$Y(f) = \mathcal{F}\left(\mathcal{F}^{-1}\left(\rho_s(f).e^{j\Phi_s(f)}\right)\right) \tag{3.45}$$

On peut par exemple prendre la magnitude du premier son et la phase du second (*cf. fig.* 3.40).

$$Y(f) = \mathcal{F}\left(\mathcal{F}^{-1}\left(\rho_1(f).e^{j\Phi_2(f)}\right)\right) \tag{3.46}$$

[2] Il est aussi possible d'utiliser l'algorithme de [Griffin and Lim, 1984] afin d'obtenir des spectres d'amplitude et de phase valides pour un spectre d'amplitude donné, par la méthode des moindres carrés

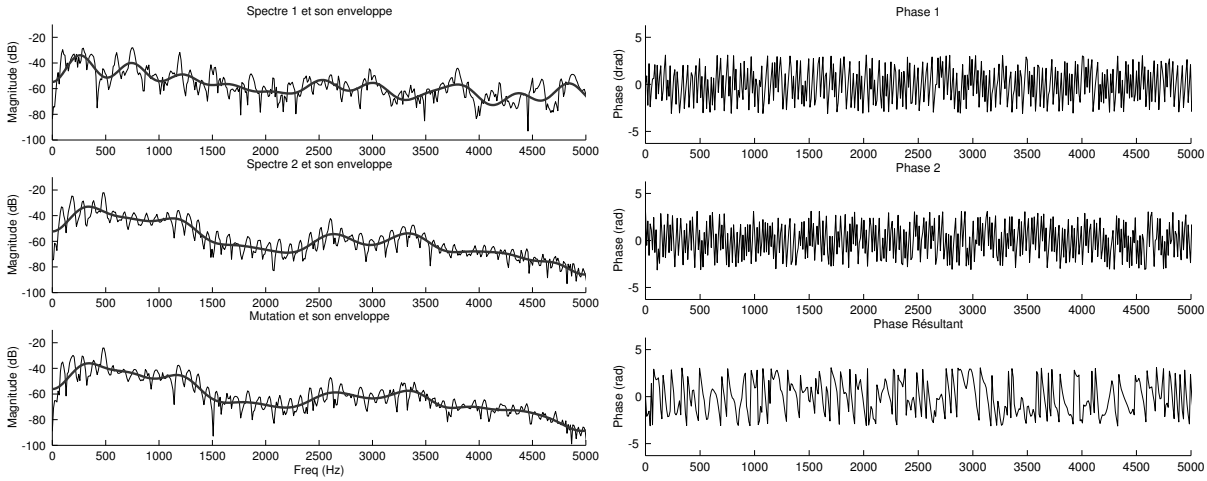


FIG. 3.40 – *Mutation : le son résultant est obtenu par resynthèse à partir du spectre d'amplitude du son 1 (en haut à gauche) et du spectre de phase du son 2 (au milieu à droite). Le spectre de phase résultant (en bas à droite) après transformation de Fourier de la transformée de Fourier inverse est différent de celui du son 2, du fait que la représentation de Fourier imposée n'était pas valide.*

Synthèse croisée La synthèse croisée ou **effet vocodeur** (*vocoding*) consiste à prendre la source d'un son $e_1(n)$ et à lui appliquer le filtre du second $f_2(n)$. Les deux sons analysés sont représentés fréquentiellement comme le produit d'une excitation E avec un filtre H : $X_i(f) = E_i(f).H_i(f)$. On utilise ici le modèle source-filtre. Notons que E est la TFCT de l'excitation et $H(f) = \exp(C(f))$ est l'enveloppe spectrale obtenue à partir du cepstre ou de la LPC. La transformée de Fourier du son résultant est donnée par :

$$Y(f) = E_1(f).H_2(f) \tag{3.47}$$

Cet effet et permet par exemple de faire parler un instrument de musique (principe des "vocodeurs" en musique électronique, de la *talk-box* : le son de la guitare est blanchi et passe par la cavité buccale où il est filtré). La hauteur provient alors du signal d'excitation. Un exemple sonore est donné Piste n°57-CD2 ♫♪ : l'enveloppe spectrale est extraite de la Piste n°17 ♫♪ et la source de la Piste n°20 ♫♪. La synchronisation entre les deux sons (au niveau rythmique) explique la réussite de l'effet musical.

Morphing Le *morphing* quant à lui est défini différemment selon les auteurs. En effet, l'analogie avec le *morphing* visuel n'est pas aisé. Doit-on considérer un son qui progressivement, dans son évolution temporelle, glisse d'un timbre vers un autre de manière continue (une seule note), discrète (plusieurs notes, chacune différente de ses voisines) ou encore un timbre hybride, entre les deux ? Les premiers *morphings* ont été réalisés par Chowning, Rush et Grey dès les années 70.

Considérons que l'effet de *morphing* consiste à créer un son hybride à partir de plusieurs sons (2 ou plus)[3]. Par exemple, on peut créer un timbre situé entre le piano et la flûte, à partir d'une analyse additive des deux sons. Le timbre hybride est alors obtenu en interpolant les données spectrales des deux analyses : pour deux sons de même hauteur, les amplitudes et fréquences des pics ne sont pas identiques (le piano voit ses partiels dévier du multiple de la fondamentale au fur et à mesure que l'on monte dans les harmoniques, par exemple, ce qui n'est pas le cas de la flûte). Le son de synthèse est obtenu par ajout de la partie harmonique, obtenue par moyenne ou interpolation entre ces valeurs, avec la partie stochastique, obtenue par filtrage d'un bruit blanc à l'aide d'un filtre dont l'enveloppe spectrale est la moyenne des deux enveloppes des parties stochastiques des sons

[3]auquel cas, on peut répondre aux autres définitions dès que l'on sait créer ces hybrides, par des évolutions temporelles des coefficients d'hybridation.

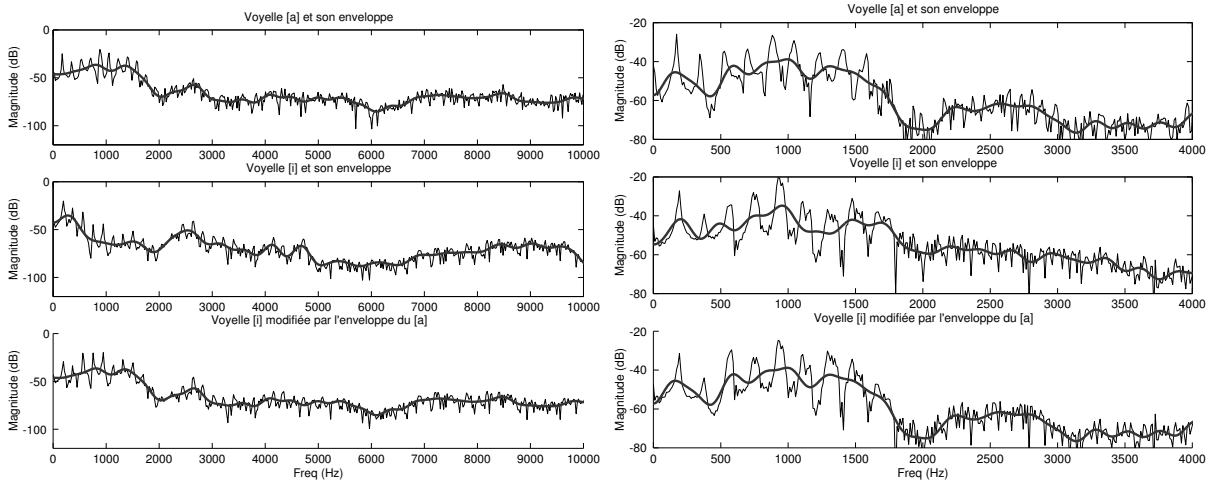


FIG. 3.41 – *Synthèse croisée : application de l'enveloppe spectrale du [a] (*fig. *haut) à la voyelle [i] (*fig. *milieu) : l'enveloppe du spectre obtenu est identique à celle du [a] (*fig. *bas). Les figures de droite sont un zoom sur la partie basse fréquence, et montrent que la hauteur d'origine est conservée : seule l'enveloppe spectrale est modifiée.*

analysés. Le *morphing* est donc bien différent du fondu enchaîné, pour lequel on perçoit deux sons (de fait de l'incohérence vibratoire, il n'y a pas de fusion).

Interpolation spectrale Un résultat proche peut s'obtenir par interpolation spectrale : les deux sons analysés sont représentés fréquentiellement comme le produit d'une excitation E avec un filtre H (modèle source-filtre). Le son de synthèse est obtenu par l'interpolation à la fois sur les excitations et sur les enveloppes spectrales :

$$Y(f) = [e_1 E_1(f) + e_2 E_2(f)] \,.\, [c_1 H_1(f) + c_2 H_2(f)] \tag{3.48}$$

L'interpolation spectrale se différencie du simple mixage, qui s'écrit dans le domaine fréquentiel :

$$Y(f) = k_1 E_1(f) H_1(f) + k_2 E_2(f) H_2(f) \tag{3.49}$$

Notons enfin que l'interpolation spectrale peut se faire sur M sons :

$$Y(f) = \left[\sum_{i=1}^{M} e_i E_i(f)\right] . \left[\sum_{i=1}^{M} c_i H_i(f)\right] \tag{3.50}$$

D'autre part, remarquons que lorsque $e_1 = 0$ et $s_2 = 0$ (de même lorsque $e_2 = 0$ et $s_1 = 0$), l'interpolation spectrale n'est ni plus ni moins qu'une synthèse croisée ; la transformée de Fourier résultant est donnée par :

$$Y(f) = (e_1.c_2) E_1(f).H_2(f) \tag{3.51}$$

ce qui correspond, à un facteur multiplicatif près, à la transformée de Fourier d'une interpolation spectrale, comme indiqué équation (3.47). La synthèse croisée est donc un cas particulier d'interpolation spectrale.

3.6.4.iv) Chuchotement, *whisperization, hoarseness*

La voix humaine, lorsqu'elle est chuchotée, contient une bien plus grande proportion de bruit que de signal harmonique, comparée à une voix parlée ou chantée normalement. Le moyen d'obtenir cet effet est assez simple : il faut augmenter le niveau de la partie résiduelle du signal. Une première

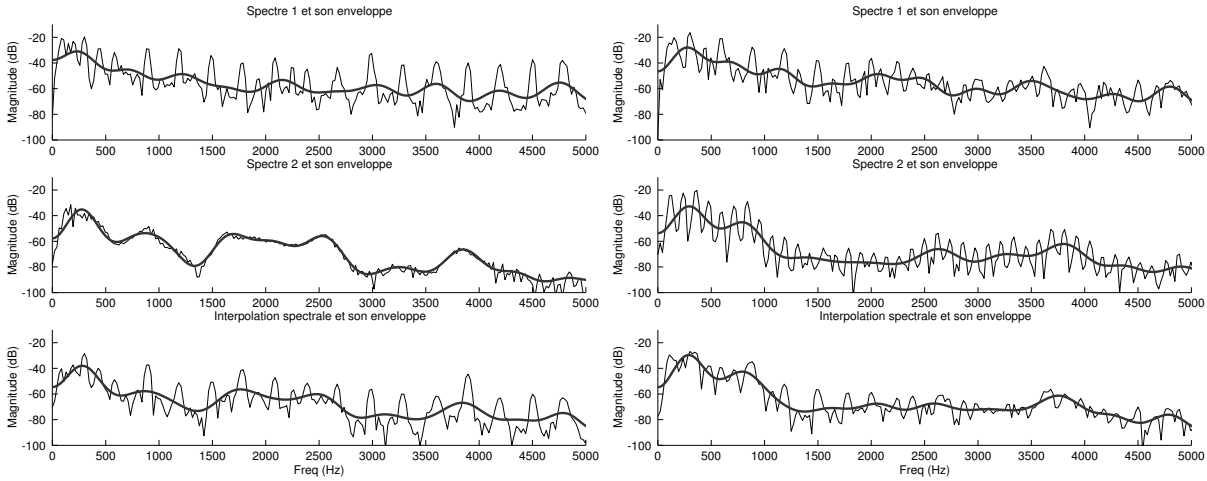


FIG. 3.42 – *Interpolations spectrales (à deux instants différents) : spectre d'amplitude et enveloppe du son 1 (*fig. *haut), du son 2 (*fig. *milieu) et du son interpolé, avec* $e_1 = 0.6$, $e_2 = 0.4$, $s_1 = 0.1$ *et* $s_2 = 0.9$ *d'après (3.48).*

manière de faire consiste à bruiter l'information sur les fréquences présentes dans le signal (appelé ***whisperization***, mis en œuvre avec le vocodeur de phase). Une second manière consiste à rehausser le niveau du bruit, après séparation des composantes sinusoïdale et bruitée (appelé ***hoarseness***, mis en œuvre avec le modèle additif).

Fonctionnement par vocodeur de phase (*whisperization*) La première méthode proposée consiste à utiliser une représentation du son de type vocodeur de phase, et à rendre aléatoire soit la phase, soit le module de chaque panier de fréquence de la Transformée de Fourier courante du vocodeur de phase (*cf. fig.* 3.43). L'utilisation de petits grains (32 ou 64 échantillons) est recommandée. Lorsqu'on resynthétise le son, on obtient une voix plus ou moins chuchotée, selon la taille du grain sur laquelle on travaille. En effet, pour de grand grains (4096, 8192 échantillons), l'information de magnitude non modifiée impose un déroulement de phase proche du déroulement original, avec un grande précision fréquentielle du fait de la taille de la TFCT identique à la taille du grain : la voix est donc très proche de la voix originale. Pour un petit grain au contraire (64, 128 échantillons), la précision fréquentielle est tellement faible que le son est très chahuté, et la voix semble chuchoté.

Fonctionnement par modèle additif (*hoarseness*) Disposant d'une représentation du son par le modèle additif, il suffit d'appliquer un facteur de gain inférieur à 1 (en linéaire) à la partie harmonique, ce qui proportionnellement va rehausser le résidu.

3.6.4.v) Autres effets : *overdrive*, *fuzz*, *exciter*, *enhancer*, débruitage, décliquage

Pour terminer, nous présentons brièvement d'autres effets sur le timbre présentés dans [Dutilleux and Zoelzer, 2002]. La distorsion est basée sur l'écrêtage doux symétrique (***soft symetric clipping***). Lorsque la fonction d'écrêtage est très douce, l'effet est appelé ***overdrive*** (la distorsion est moindre, et le son plus "chaud"). Si l'on utilise une loi asymétrique, l'effet est légèrement différent et s'appelle ***fuzz*** : la distorsion est bien plus prononcée. Il a été très utilisé pour la guitare électrique, notamment par Jimi Hendrix. L'***exciter*** est un effet qui ajoute de la brillance au son, par un subtil jeu sur les phases du signal. L'***enhancer*** est un effet qui ajoute de la brillance et un peu de distorsion au son. Il combine une équalisation avec des traitements non linéaires. L'équalisation est effectuée en fonction de connaissance psychoacoustiques, et une petite distorsion est ajoutée (de manière à peine distinguable). Le débruitage (***denoising***) consiste à diminuer le bruit de fond en augmentant le

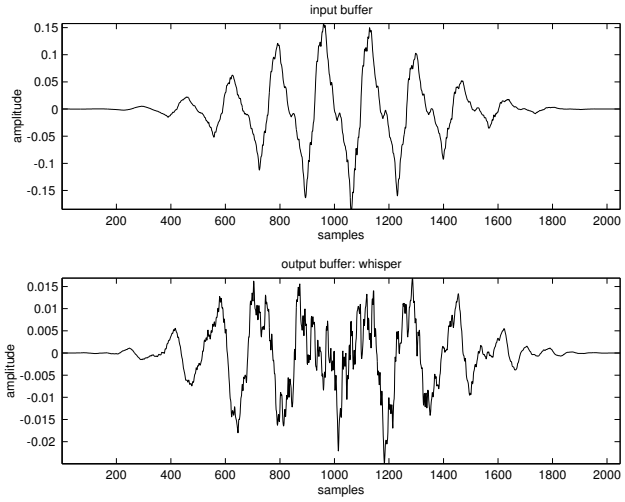


FIG. 3.43 – *Grain de 2048 échantillons avant (en haut) et après chuchotement (par PV, en bas) : on voit clairement que les phases sont rendues aléatoires.*

rapport signal sur bruit. Le décliquage (***de-clicking***) consiste à supprimer les clics des enregistrements sur disque vinyle, et fait appel à des méthodes de traitement du signal stochastique [Cappé, 1993].

3.7 Traitements modifiant plusieurs paramètres à la fois

Une hypothèse de base à la représentation des sons par leurs indices perceptifs est que ces dimensions sont indépendantes. En réalité, ceci n'est pas parfaitement vérifié. En effet, augmenter la sonie d'un son implique une modification du timbre et de la hauteur perçue. Modifier le spectre (la répartition en fréquence et l'amplitude des harmoniques) implique une modification du timbre mais aussi de la hauteur perçue. Un filtrage peut modifier à la fois la perception du timbre mais aussi la localisation. Ces quelques exemples montrent les faiblesses de cette représentation.

Des traitements sonores peuvent modifier volontairement plusieurs paramètres perceptifs à la fois. Nous en présentons ici quelques uns, à savoir la robotisation, la transposition sans conservation de durée ni ne formants et le brassage.

3.7.1 Robotisation

La robotisation s'applique à des sons de voix parlée ou chantée. Elle transforme une voix humaine en voix métallique, robotique, tout en conservant l'articulation. Le traitement correspondant consiste à remplacer la voix par un train d'onde. La fréquence du train d'onde détermine la hauteur de la voix robotique. La manière d'obtenir le train d'onde est simple : on modifie les phases d'un grain de longueur L_w en les remplaçant par à 0 pour chaque FFT. Ceci revient à caler chaque sinusoïde de façon à ce que la phase soit nulle au centre du grain (on obtient alors un cosinus). On obtient alors un pic au centre du grain (*cf.* 3.44). Le grain obtenu est ensuite fenêtré (par exemple par une fenêtre de Hanning modifiée), ce qui minimise l'amplitude des pics secondaires sur les côtés du pic central, s'ils existent. L'information sur le contenu fréquentiel (les formants) proprement dit est donc conservé du fait que l'on ne modifie pas le spectre de magnitude, mais pas l'information de déroulement de phase puisqu'on modifie le spectre de phase. Entre deux fenêtres, le signal original est remplacé par des 0 (***zero-padding*** ou bourrage de zéros). La hauteur de la voix de robot est déterminée par le pas $R_S = 1/F_0^{robot}$ entre deux fenêtres.

Effet de la taille du grain et/ou de la fenêtre L_w En changeant la taille de la fenêtre L_w, on sélectionne une plus ou moins grande partie du signal de départ : le grain de synthèse contient plus ou moins de pics. La voix est donc plus ou moins robotique selon la taille du grain : pour un seul pic, la voix est robotique ; pour plusieurs pics, la voix est moins robotique, et ressemble plus à la voix originale.

Une bonne taille de grain est entre 256 et 512 échantillons (voix robotique très proche de l'originale, sauf concernant la hauteur), mais peut varier entre 64 (*cf. fig.* 3.44 gauche) (voix robotique, son très granuleux) et 1024 échantillons (*cf. fig.* 3.44 droite : voix robotique avec double hauteur). On remarque que plus L_w augmente, plus des pics secondaires apparaissent, de période la période du signal original. Ceci provoque dans le son un effet de deuxième hauteur : la hauteur originale du son apparaît dans la voix de robot. Le son résultant est aussi intéressant, mais ne correspond plus à un robot tel qu'on s'y attend. On dirait plutôt un combinaison de deux effets. Il est donc conseillé de prendre de petits grains lorsqu'on ne désire pas conserver la hauteur originale.

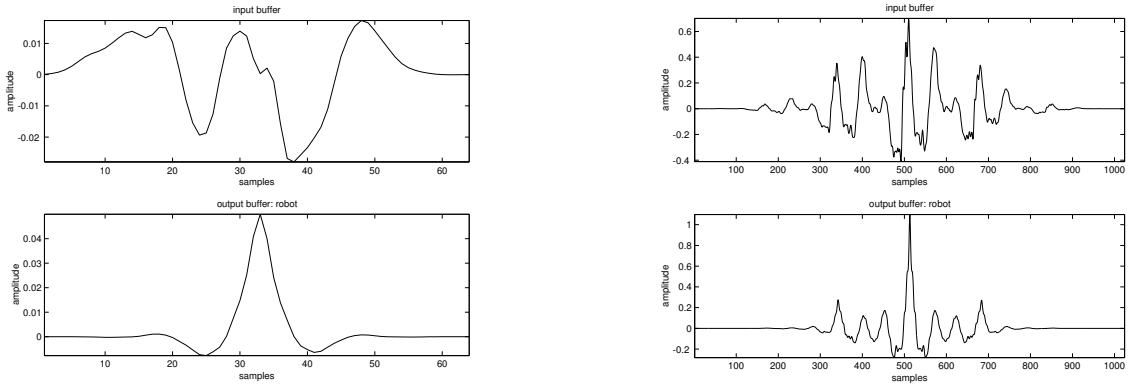


FIG. 3.44 – *Effet de la taille du grain sur la robotisation.*
*A gauche : robotisation appliquée à un grain de 64 échantillons : grain d'origine (*fig. *supérieure) et après traitement (*fig. *inférieure). Un pic central (et un seul) est créé au centre du grain.*
*A droite : robotisation appliquée à un grain de 1024 échantillons : grain d'origine (*fig. *supérieure) et après traitement (*fig. *inférieure). Plusieurs pics sont créés, l'un au centre du grain (principal) et les autres de part et d'autre.*

Intérêt de cet effet L'effet de robotisation est très utilisé dès lors que l'on veut donner une connotation "machine parlante" : les films de science-fiction y font largement appel lorsqu'ils donnent une voix à des machines s'adressant à des humains.

3.7.2 ré-échantillonnage

Le ré-échantillonnage d'un signal consiste à modifier à la fois la durée et la hauteur d'un son. On peut soit modifier la fréquence d'échantillonnage de lecture par rapport à cette d'écriture du son, soit interpoler la forme d'onde pour ajouter ou supprimer des valeurs. L'effet obtenu est alors une transposition sans conservation de la durée ni des formants (ou une dilatation/contraction sans conservation de la hauteur).

Prenons l'exemple analogique des magnétophones à bande. De manière à pouvoir caler un morceau ou trouver une zone temporelle particulière, une fonction de lecture à vitesse variable existe. C'est ce que réalise un DJ (*disk jockey*) lorsqu'il cale ses morceaux à partir d'un disque vinyle : il peut lire à l'endroit ou à l'envers le son enregistré sur le support vinyle, et ce à différentes vitesses.

Le fait de lire le son à des vitesses différentes correspond, en numérique, à avoir un signal échantillonné à une fréquence F_e^1 *Hz* et à le lire à la fréquence d'échantillonnage F_e^2 *Hz*. La durée du signal n'est pas conservée, puisqu'il faut soit F_e^1, soit F_e^2 échantillons par seconde pour produire

le son. D'autre part, une sinusoïde de fréquence $F_e^1/10$ a 10 périodes par seconde lorsqu'elle est échantillonnée à F_e^1 Hz alors qu'elle en a F_e^2 Hz lorsqu'elle est échantillonnée à F_e^2 Hz (transposition d'un facteur $\frac{F_e^2}{F_e^1}$).

3.7.3 Brassage (*time shuffling*), granulation

Le brassage consiste à considérer le signal comme un ensemble de grains et à en modifier l'ordonnancement. Ainsi, on reconstruit un signal par ajout-superposition, sans se soucier des phases. C'est la base des méthodes de synthèse granulaire[Roads, 1999; Roads, 2002]. Cet effet ne respecte pas l'axe temporel du son, ni le timbre et l'amplitude, du fait que l'ajout-superposition sans conservation des phases peut créer des trous dans le spectre.

3.7.4 Changement de genre

Une voix d'homme est différent d'une voix de femme par le spectre et par le fondamental utilisé pour parler. Changer le genre d'une voix consiste alors à changer la voix d'un homme en celle d'une femme, ou le contraire (c'est un cas particulier du *morphing*, *cf.* p. 91).

L'ambitus de fréquence de l'homme est différent de celui de la femme. Modifier cet ambitus par transposition avec préservation des formants est une première étape (par exemple, monter le son d'une octave pour passer d'une voix d'homme à une voix de femme, ou baisser le son d'un octave pour passer d'une voix de femme à une voix d'homme).

Ensuite, il a été montré que les formants (notamment le premier) d'une même voyelle varient lorsque la fondamentale évoluent. Une manière de modéliser cela consiste à effectuer un décalage de l'enveloppe spectrale à la suite de la transposition. Pour passer d'une voix d'homme à une voix de femme, celui-ci se fait uniquement pour les fréquences au dessus de 100 Hz. En dessous de 100 Hz, on ne décale pas l'enveloppe spectrale. Entre 100 et 500 Hz, on décale linéairement de 0 à 50 Hz l'enveloppe spectrale. Au-dessus de 500 Hz, on ne décale pas plus l'enveloppe. Pour passer d'une voix de femme à une voix d'homme, le décalage se fait inversement, dans la même plage $[100; 500]Hz$. Du fait que le décalage du spectre dépend de la fondamentale, cet effet est adaptatif.

Nous allons maintenant aborder le contrôle des effets audionumériques dans un premier temps, puis préciser quelques notions sur le geste instrumental afin de posséder le vocabulaire utilisé quand on parle du contrôle gestuel. Dans un troisième temps, nous parlerons de la mise en œuvre des effets audionumériques, en temps-réel et hors temps-réel.

3.8 Contrôle et mapping

Le problème du contrôle est important en traitement sonore. Tout d'abord, le type de contrôle que l'on veut sur l'effet déterminera s'il faut utiliser du temps réel ou non. Ainsi, un contrôle donné par une courbe peut se faire en temps différé, alors qu'un contrôle donné par un transducteur gestuel est plus adapté pour des traitements en temps réel.

Le *mapping* est la mise en correspondance entre des paramètres en entrée (les coordonnées d'un transducteur gestuel, des valeurs, des courbes) et les contrôles de l'effet (les paramètres qui le définissent). Différentes stratégies de mise en correspondance existent, notamment le *mapping* explicite (sa formulation mathématique est donnée via des équations) ou implicite (le modèle de correspondance est non-linéaire, et seul son principe est décrit, par exemple les réseaux de neurones).

3.8.1 Contrôle avec interfaces graphiques

Le contrôle par interfaces graphiques est le plus simple, très fréquemment utilisé du fait que tout ordinateur dispose au moins d'un écran, d'une unité centrale, d'un clavier et d'une souris. L'écran est alors utilisé pour présenter les paramètres de contrôle de l'effet (ex : *plug-in*, *patch Max/MSP*,

jMax ou *Pure Data*), et la souris et ses boutons pour sélectionner (activer) un paramètre, et le faire varier. Des courbes et indicateurs (jauges, potentiomètres, valeur numérique) permettent un retour visuel à l'utilisateur. Voici quelques points clefs à respecter lors de la réalisation d'interfaces graphiques utilisateurs[4] (IGU, ou *GUI, graphical User Interfaces*), d'après [Zoelzer, 2002] :

- la visibilité : elle permet à l'utilisateur de voir ce qu'il peut faire de tel ou tel outil ;
- la transparence : l'utilisateur ne voit pas ce qui lui a été volontairement caché (les calculs entrepris par l'ordinateur), mais il a accès à une visualisation de la tâche réalisée, selon l'image mentale qu'il s'est faite du processus ;
- prévisibilité : le système doit réaliser la tâche à laquelle l'utilisateur naïf s'attend, à partir de ses connaissances et à l'aide de métaphores (par exemple avec la représentation d'un potentiomètre, d'un clavier) ;
- consistance : l'application doit être prévisible dans tout contexte du programme et d'une application à l'autre ;
- intégrité : l'interface doit protéger les données précieuses, même lorsque l'utilisateur effectue une mauvaise manipulation ; il doit être possible d'annuler une action, et de revenir en arrière ;
- concision : l'interface doit être concise, tant au niveau du contrôle (raccourcis clavier, menu déroulants, valeurs par défaut, etc) qu'au niveau de l'affichage écran via la fenêtre de l'application (tout particulièrement l'information importante qui doit être claire et précise) ;
- apparence à l'écran : l'apparence doit être soignée, claire et ordonnée ; brillance, contraste, couleurs, textures, clignotements doivent être utilisés principalement pour leur sens, avant les raisons esthétiques ;
- adaptabilité : l'utilisateur doit pouvoir, sans avoir à programmer, réaliser l'interface correspondant à ses besoins et son niveau de connaissances ;
- guide : tout utilisateur peut avoir parfois besoin de réponses à certaines questions ; si l'interface ne se comprend pas d'elle-même, une manuel en ligne ou une aide contextuelle peut être utile.

3.8.2 Contrôle gestuel

L'utilisation du geste comme moyen de contrôle se fait à l'aide de transducteurs gestuels : la souris, le clavier, mais aussi les claviers midi, le joystick, les capteurs électromagnétiques de position, de vitesse, d'accélération, d'orientation (dont le *radio baton* et le *flock of birds*), les capteurs de pression, le volant de jeu et les pédales associées, le gant de réalité virtuelle, les *racks* de potentiomètres, les capteurs à ultrasons, les capteurs de percussion, etc. Les données transitent entre le transducteur et l'ordinateur peut être au format MIDI (c'est le cas pour *Max/MSP*, *jMax*, *Pure Data*), ou série, USB, selon le transducteur utilisé.

3.8.3 Contrôle automatisé

Le contrôle automatisé ou automatique peut se faire de deux manières. La première consiste à utiliser une interface graphique pour faire apparaître la valeur des contrôles de l'effet, et les modifier en temps réel à l'aide de la souris. Les séquenceurs midi et audio, tel **ProTools**, **Cubase**, **Logic Audio**, **Digital Performer**, permettent généralement de réaliser cette tâche. Une autre manière de faire est d'automatiser le contrôle par des paramètres extraits du son. C'est le principe des effets audionumériques adaptatifs, qui sera développé dans la partie II du mémoire.

3.8.4 Le geste instrumental

Le **geste** est l'« ensemble des comportements corporels associés à notre activité musculaire »[Cadoz, 1999]. Le geste **musical** est un geste dont l'effet est de produire un son musical. Cet effet peut être volontaire ou involontaire, tout dépend à la fois de l'état d'esprit de celui qui effectue le geste

[4] interfaces d'applications musicales ou non.

et de celui qui écoute le son produit (qui n'est pas forcément le même individu). La typologie du geste instrumental (*cf. fig.* 3.45) qui va être donnée provient entièrement du chapitre "***Musique, geste, technologie***" de C. Cadoz de l'ouvrage "***Les Nouveaux gestes de la musique***" [Cadoz, 1999], et est représenté sous la forme d'un diagramme heuristique [Buzan and Buzan, 2003].
Le geste musical comporte deux types de gestes :

- le geste instrumental, dont l'intention est de produire ou de moduler l'énergie acoustique (dont la destination finale est nos tympans) ;
- le geste non instrumental qui, même s'il peut produire ou moduler de cette énergie, ne s'applique pas physiquement sur un instrument produisant l'énergie acoustique.

Les gestes instrumentaux peuvent ensuite se diviser en trois catégories :

- les gestes d'**excitation**, mettant en vibration une structure produisant l'énergie acoustique ;
- les gestes de **modification**, modulant l'énergie acoustique en modifiant quelques chose dans la structure vibrante ;
- les gestes de **sélection**.

Les gestes d'excitation sont au nombre de trois : l'excitation instantanée (par exemple le pincement d'une corde, la frappe d'une percussion, d'un verre, d'une barre en métal), l'excitation continue (une succession d'excitations instantanées, tel le raclement) et l'excitation entretenue (de l'énergie est apportée continuellement, par le frottement de l'archet pour les instruments à cordes, par le souffle pour les instruments à vent).

Les gestes de modification sont structurels ou paramétriques. Un exemple de geste de modification structurelle est le changement de jeux d'orgues en cours de jeu. Les modifications paramétriques peuvent être continues, telle la modification de la longueur de la structure vibrante : glissando en déplaçant un doigt le long d'une corde, en changeant la longueur d'un tuyau tel le piston du trombone, ou en changeant la longueur d'une plaque vibrante par glissement sur une structure fixe. Elles peuvent encore être discrètes, par exemple par l'utilisation de frettes sur la guitare pour discrétiser les longueurs de vibrations de cordes possibles.

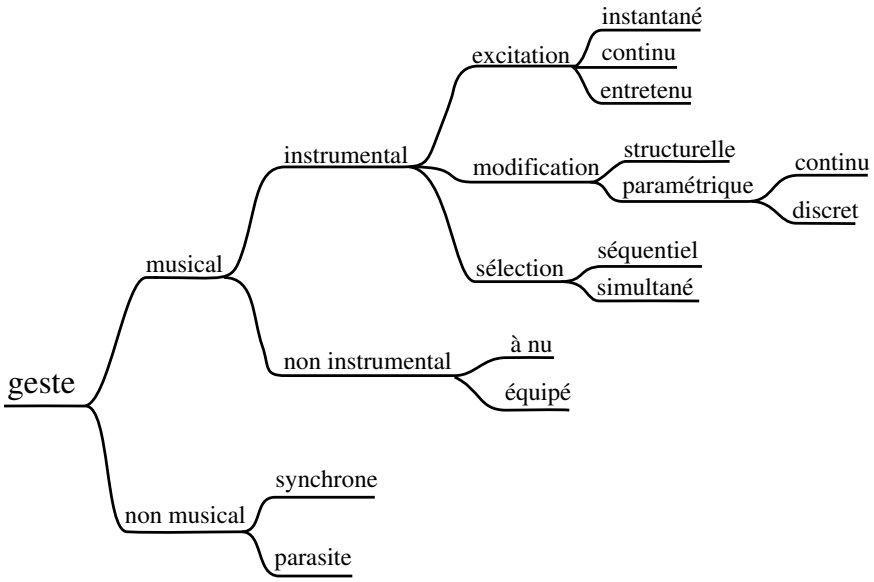


FIG. 3.45 – *Récapitulatif de la typologie des gestes par diagramme heuristique.*

Les gestes de sélection sont séquentiels ou simultanés, ou les deux à la fois. Un geste de sélection séquentiel est une suite de sélections uniques (par exemple, une ligne mélodique sur un instrument), tandis qu'un geste de sélection simultané correspond à plusieurs gestes de sélection effectués au même instant. C'est le cas d'un accord joué sur un clavier ou une guitare.

Le geste non instrumental correspond à un geste communicationnel dont l'énergie consommée par

le phénomène sonore n'est pas de source humaine. Cela signifie qu'il y a deux circuits en dépendance énergétique : un circuit humain et un circuit extérieur sur lequel il interagit. Ces gestes peuvent être des gestes à mains nues (le geste qui accompagne la parole, le langage des signes) ou à mains équipées (baguette du chef d'orchestre, stylo pour l'écriture). Notons que ces gestes peuvent être de modification ou de sélection, et peuvent être communs aux gestes instrumentaux de modification ou de sélection (par exemple les mouvements amples d'un interprète en jeu, [Wanderley and Depalle, 1999; Wanderley, 2001]).

Pour conclure, il sera très intéressant d'avoir à l'esprit cette typologie lorsque nous présenterons un par un les effets adaptatifs et leur contrôle (*cf.* chap. 5 et 6). Nous allons maintenant présenter les différentes techniques de mise en œuvre temps-réel et non temps-réel des effets audionumériques.

3.9 Mises en œuvre

3.9.1 Mise en œuvre en temps-réel

Nous avons vu que la plupart des effets audionumériques peuvent se mettre en œuvre en temps-réel. Ceci peut se faire soit par calcul du signal traité échantillon après échantillon, auquel cas la latence est très faible (quelques échantillons), soit par calcul du signal traité bloc par bloc, auquel cas la latence est plus importante (de la taille du bloc).

Mise en œuvre échantillon par échantillon

La mise en œuvre échantillon par échantillon consiste à prendre un échantillon en entrée, et à donner l'échantillon traité dans un temps moyen inférieur ou égale au temps d'acquisition de l'échantillon suivant. Cette mise en œuvre nécessite une optimisation des calculs du traitement.

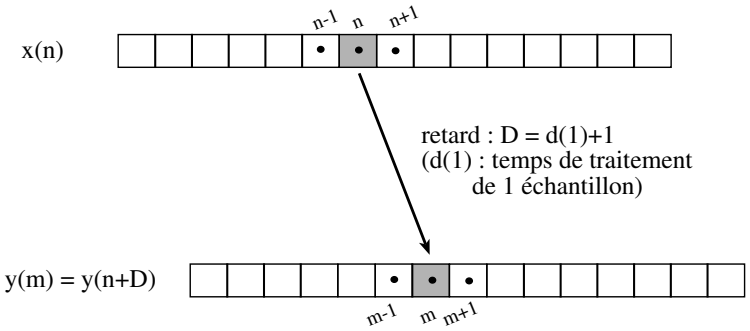


FIG. 3.46 – *Mise en œuvre échantillon par échantillon : le retard D est fonction du temps de calcul du traitement d'un échantillon $d(1)$.*

L'avantage principal de cette mise en œuvre est qu'il se prête parfaitement au temps-réel : le délai entre entrée et sortie est au plus de quelques échantillons. Les inconvénients inhérents à la mise en œuvre échantillon par échantillon sont :

- les traitements doivent être rapides, réalisables sur un cycle d'échantillonnage ;
- les traitements doivent être simples, réalisables échantillon par échantillon (domaine temporel) ;
- des codes de calcul bien optimisés sont peu lisibles (portions de langage assembleur et astuces d'optimisation).

Mise en œuvre par blocs

La mise en œuvre par blocs consiste à découper le signal à traiter en blocs, et à traiter successivement chaque bloc. Ces blocs peuvent être des fenêtres d'analyse (vocodeur de phase, analyse

additive), ou un ensemble de N fenêtres d'analyse. Ces blocs peuvent se superposer (*cf.* la TFCT).

Le délai entre le bloc entrant et le bloc traité est au minimum de la durée du bloc (en nombre d'échantillon) : il faut attendre d'avoir reçu les T_{bloc} échantillons avant de les traiter, et le traitement peut nécessiter un temps supérieur à une période d'échantillonnage ($1/44100$ s).

Cette technique permet de mieux répartir la charge de calcul que la technique échantillon par échantillon (*cf.* sec. 3.9.1) : le traitement de certains échantillons peut nécessiter plus de temps que pour d'autres échantillons ; il peut aussi y avoir une étape d'initialisation commune au traitement de tous les échantillons du bloc, pour des traitements spécifiques.

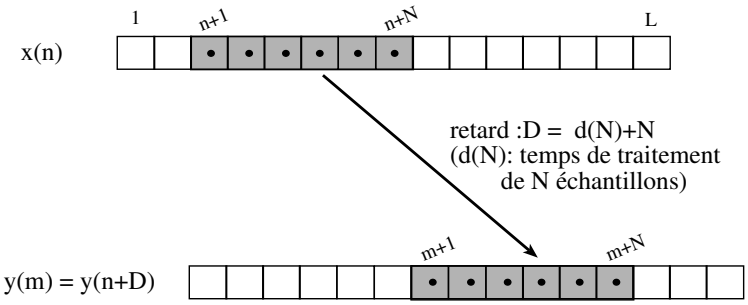


FIG. 3.47 – *Mise en œuvre bloc par bloc : le retard D est fonction du temps de calcul du traitement de N échantillons $d(N)$.*

Les avantages principaux de cette mise en œuvre sont :
- la mise en œuvre est temps-réel ou quasiment (faibles retard, de 20 à 30 ms) pour de petits blocs ;
- les traitements complexes (notamment spectraux) sont rendus possibles ;
- en cas de lecture-écriture dans des fichiers, c'est plus rapide que la mise en œuvre échantillon par échantillon.

Les inconvénients inhérents à la mise en œuvre par bloc sont :
- le temps de retard de l'ordre de 20 ms pour des blocs (TFCT) de 2048 échantillons à 44 100Hz : ce n'est pas vraiment temps-réel, le délai peut poser problème pour certaines applications. Pour des blocs plus grands, c'est encore pire ;
- la mise en œuvre des applications récursives (telles les filtres récursifs) n'est pas directe, car il faut initialiser les coefficients du filtre pour chaque bloc.

3.9.2 Mise en œuvre en temps différé

Mise en œuvre par blocs

La mise en œuvre par blocs se prête aussi au temps-réel, dès lors que les blocs sont assez petits. Elle est souvent utilisée pour des méthodes utilisant la TFCT et la transformée inverse...

Mise en œuvre vectorielle

Cette mise en œuvre est liée au traitement en temps différé. Le seul cas où elle peut être utilisée en temps réel est celui d'applications spécifiques utilisant de grands délais entre un son et sa version traitée. Dans cette approche, le son numérisé est considéré comme un seul vecteur. Le traitement consiste en une succession d'unités de traitement prenant un vecteur en entrée, et donnant un vecteur en sortie. Ce type de traitement est naturel dans des langages tel que *Matlab*.

Les avantages principaux de cette mise en œuvre sont :
- chaque procédure de traitement est indépendante des autres, et peut être testée indépendamment ;

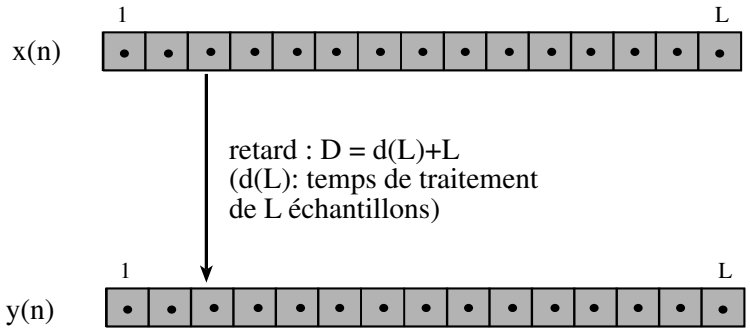


FIG. 3.48 – *Mise en œuvre par vecteur : le retard D est fonction du temps de calcul du traitement de tous les échantillons $d(L)$.*

– le temps de traitement peut être très court dès lors que les procédures de traitement sont optimisées ;
– un effet est constitué d'une succession de procédures.

Les inconvénients inhérents à la mise en œuvre vectorielle sont :

– le coût en terme de mémoire : il faut suffisamment d'espace mémoire pour stocker tous les vecteurs, y compris les variables intermédiaires ;
– cette mise en œuvre n'est pas utilisable en temps réel, puisque le son d'entrée est supposé entièrement connu avant le traitement (sauf dans le cas évoqué en introduction).

Chapitre 4

Descripteurs du son

L'homme se distingue de l'animal par plusieurs traits remarquables. Il paie des impôts, écoute du rock'n'roll, rase les poils de son visage et fait cuire une bonne partie de ses aliments.
Philippe Meyer [Meyer, 1995]

Une description qui dépasse dix mots n'est plus visible.
Jules Renard [Renard, 2003]

Sommaire

Les descripteurs du son correspondent à des paramètres qui le décrivent. On les appelle aussi paramètres descriptifs, traits, caractéristiques.

4.1 Notre but : accéder à des représentations du son

On désire accéder à des courbes de contrôles qui décrivent le son. Ces courbes, à quelques transformations près (*cf.* chap. 6), seront utilisées comme contrôle d'un effet, comme expliqué précédemment. Dans les systèmes de reconnaissance d'image, l'extraction d'information se fait par plusieurs étapes d'analyse. La première étape d'analyse est très proche du signal, et consiste à

rechercher les ruptures (d'intensité, de couleur) . A partir de ces points de rupture, on cherche les arêtes et de proche en proche (deuxième étape d'analyse), puis les formes qu'elles décrivent (deuxième étape d'analyse). La démarche est donc en plusieurs étapes, en partant de descripteurs de bas-niveau, de signal, pour arriver à des descripteurs perceptifs, de haut-niveau.

Pour le son, la démarche est similaire : on va d'abord extraire des descripteurs continus à partir du signal, puis les analyser pour en extraire une information de plus haut niveau. On distingue cependant plusieurs descripteurs de haut-niveau : ils peuvent être des descripteurs psychoacoustiques, des descripteurs de signal ou des méta-descripteurs. L'ensemble de ces descripteurs permet de construire des espaces de représentation des sons. Les espaces de représentation basés sur des descripteurs perceptifs sont appelés espaces perceptifs. Différentes applications effectuent des analyses et extractions de descripteurs, notamment les applications multimédia de codage, de compression, de classification automatique, de reconnaissance, de transposition automatique, de segmentation automatique de signaux. Toutes ces applications cherchent à extraire l'information pertinente pour bien effectuer leur tâche. La pertinence de l'ensemble de descripteurs utilisés est vérifié par comparaison avec ce que l'être humain sait faire. Cette capacité est basée sur sa perception, si bien qu'on utilise des modèles de la perception auditive.

Les paramètres que l'on extrait du son peuvent se décrire de différentes manières, selon le point de vue où l'on se place et selon ce que l'on attend de l'information véhiculée par le paramètre. Pour un contrôle automatique d'effet audionumérique, tout descripteur peut bien sonner, selon le son que l'on traite, l'effet que l'on y applique, le *mapping* que l'on applique. Il n'y a pas de restriction à priori sur tel ou tel descripteur que nous allons présenter par la suite. Cela dit, nous porterons une attention particulière aux descripteurs faisant sens à la perception auditive.

Nous allons d'abord présenter les grandes catégories de descripteurs et espaces perceptifs pour mettre en relief le contexte, puis décrire les modèles sous-jacent à l'extraction de descripteurs, selon la problématique (avec classification des descripteurs utilisés). Nous présenterons alors en détail les descripteurs de bas niveau, puis les descripteurs de haut niveau, ainsi que les autres descripteurs dérivés des précédents. Enfin, nous parlerons du problème de l'échantillonnage des courbes des descripteurs et de la qualité du calcul.

4.2 Catégories de descripteurs du son

Tant que l'on ne pouvait réaliser d'analyse de signaux (d'abord analogiques puis numériques depuis les années 1960), la perception du signal ne pouvait pas être quantitativement mise en relation avec signal, par des méthodes de calcul. Il y avait alors la représentation du son par les partitions pour les musiciens d'une part, et la connaissance des modèles physiques des instruments d'autre part. Ceci a permis notamment d'établir le lien entre la taille d'une structure vibrante (colonne d'air ou corde) et la hauteur fondamentale de la note produite. Cependant, seuls les sons dont on connaissait le mode de production et dont on savait faire l'analyse physique pouvaient être décrits par l'analyse, ou sinon par des qualificatifs concernant leur perception [von Helmholtz, 1954]. Une fois l'analyse du signal rendue possible, les chercheurs ont cherché à établir les liens entre le signal et la perception que l'on en a. De là est venu l'essor de la psychophysique de l'audition et de la psychoacoustique, avec la description de modèles de la perception et de la méthode de calcul des descripteurs perceptifs. Nous allons maintenant présenter en premier les descripteurs de signal, extraits par analyse temporelle et fréquentielle, puis les descripteurs perceptifs.

4.2.1 Descripteurs de signal

Les descripteurs de signal portent sur l'analyse temporelle et fréquentielle, éventuellement combinées. Ils peuvent être très proches ou très éloignés des descripteurs perceptifs.

Descripteurs temporels

Les descripteurs temporels concernent principalement l'amplitude du signal et les bornes de segmentation des notes :
- l'amplitude ;
- le trémolo s'il existe ;
- pour une note, l'attaque, la partie stable, et la décroissance (le début et la fin pour chacune des trois) ;
- pour un bruit ou un son électroacoustique : le début et la fin.

Descripteurs spectraux

Les descripteurs spectraux sont issus d'une analyse additive du signal, laquelle fournit une décomposition en une partie déterministe (sinusoïdale, harmonique) et une partie stochastique (résidu). Les descripteurs que l'on peut extraire portent donc sur chacune des deux composantes, et se répartissent en 3 catégories [Herrera *et al.*, 1999a; Herrera *et al.*, 1999b] : les paramètres de base de l'analyse, les descripteurs quasi-instantanés qu'on en extrait et enfin les attributs des segments du signal segmenté. Les paramètres de base de l'analyse additive sont :
- les fréquences des partiels ;
- les amplitudes des partiels ;
- le spectre du résidu.

Les descripteurs quasi-instantanés que l'on peut en extraire sont :
- la fréquence fondamentale pour les spectres harmoniques ;
- l'amplitude de la composante déterministe ;
- l'amplitude du résidu ;
- la forme ou enveloppe spectrale de tout le spectre ;
- la forme ou enveloppe spectrale de la composante déterministe ;
- la forme ou enveloppe spectrale du résidu ;
- le degré d'harmonicité (la distorsion harmonique) ;
- la mesure de la quantité d'information non sinusoïdale (caractère bruité ou *noisiness*) ;
- le centroïde spectral ;
- la balance des harmoniques paires/impaires ;
- la fréquence de coupure du spectre de la composante déterministe ;
- la fréquence de coupure du spectre de la composante résiduelle ;
- la pente spectrale.

Une fois que l'on a extrait les paramètres quasi-instantanés, leurs dérivées, les différences d'une trame à l'autre, on peut segmenter le signal d'après ces descripteurs, et ensuite extraire les descripteurs et les paramètres statistiques des régions (valeur moyenne, variance, coefficient d'asymétrie, coefficient d'aplatissement). Les segments se caractérisent par exemple selon :
- le nom de la note jouée ;
- le vibrato ;
- la présence et la description d'un ornement ;
- un mode de jeu ;
- une texture (d'après l'analyse des variations de fréquence des harmoniques).

Structure hiérarchique des descripteurs

On caractérise la dynamique des notes instrumentales par l'attaque, la première décroissance, la portion stable et la décroissance (relâchement). Pour cela, on utilise la variation moyenne de chaque attribut instantané, tel que le variation moyenne de fréquence fondamentale, la variation d'amplitude moyenne, la variation de la variation d'enveloppe spectrale. Pour caractériser la portion stable des sons, on extrait des valeurs moyennes de chaque attribut local et on mesure les autres

attributs globaux tel que les taux et profondeur du vibrato, variant dans le temps. La structure de données est alors hiérarchique, à plusieurs niveaux, allant des descripteurs quasi-instantanés aux descripteurs de segments. Des analyses ont été réalisée [Strawn, 1985; Strawn, 1987] pour quantifier les variations d'intensité, de fréquence fondamentale lors des transitions entre notes.

4.2.2 Descripteurs perceptifs du son

Il existe différents types de paramètres perceptifs du son. Dans la synthèse par modèles physiques, certains paramètres physiques sont directement reliés à des paramètres perceptifs du son. Des paramètres de signal tels que le centre de gravité spectrale (ou centroïde) ont aussi un lien direct avec des paramètres perceptifs du son. Des méta-paramètres sont utilisés en synthèse ou effets pour contrôler un grand nombre de paramètres. Enfin, les paramètres psychoacoustiques sont les descripteurs privilégiés de la perception humaine du son (*cf. fig.* 4.1).

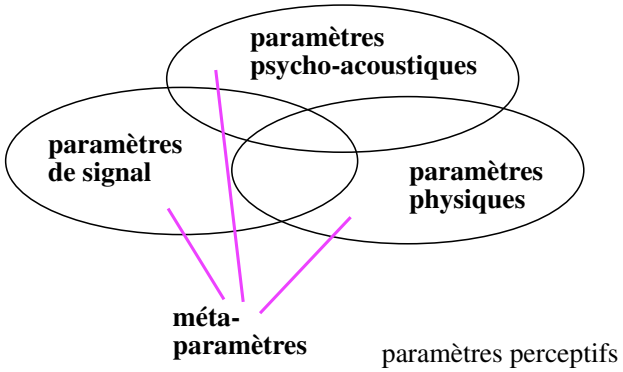


FIG. 4.1 – *Descripteurs perceptifs, répartis entre descripteurs psychoacoustiques, descripteurs physiques, descripteurs de signal et méta-descripteurs.*

Descripteurs physiques liés à la perception

Des paramètres physiques de l'objet sonore peuvent être reliés à la perception : le volume de l'objet, le matériau, l'élasticité, etc. [McAdams and Bigand, 1994]. Dans un article sur l'exploration du timbre par analyse-resynthèse [Risset and Wessel, 1999], les auteurs expliquent que l'un des avantages des modèles physiques est que les paramètres de contrôle sont plus simples et plus intuitifs que certains paramètres de contrôle des modèles de signal — tel que l'index de modulation de la distorsion non-linéaire, par exemple — parce qu'ils sont corrélés à une réalité physique que nous sommes habitués à évaluer. Ces paramètres sont des descripteurs de bas-niveau dans le sens qu'ils sont très proches du modèle, mais aussi de haut-niveau lorsqu'ils ont une grande signification perceptive. On peut changer leurs valeurs — par normalisation, translation, changement d'échelle, combinaison linéaire — pour obtenir des paramètres de haut-niveau, mieux liés à la perception.

Descripteurs de signal liés à la perception

Certains paramètres de signal donnent des indications sur la perception. Par exemple, la mesure de l'énergie par RMS (*root mean square*), même si elle est différente de la sonie — le descripteur psychoacoustique de l'intensité sonore perçue — permet une bonne évaluation de l'évolution de l'énergie du signal, qui peut être proche de l'évolution de la sonie. La fréquence fondamentale d'un son harmonique est souvent assimilée à la hauteur tonale perçue (*pitch*) — le descripteur psychoacoustique correspondant — du fait que leurs valeurs sont très proches. La fréquence fondamentale peut donc remplacer le hauteur perçue dès lors que le *mapping* n'est pas trop sensible aux faibles variations. Un dernier exemple est le centre de gravité spectral (CGS ou centroïde), directement

corrélé au descripteur psychoacoustique du timbre appelé brillance [McAdams *et al.*, 1995]. Ces descripteurs ne doivent pas remplacer leurs contreparties psychoacoustiques si l'on veut contrôler et modifier un son avec des contrôles temps-réels sur les paramètres psychoacoustiques.

Méta-descripteurs

On peut vouloir manipuler un ensemble de paramètres du même niveau d'une manière globale, grâce à un petit sous-ensemble de paramètres, par exemple une courbe de N points que l'on manipule à l'aide de la souris (2 contrôles). Ces paramètres de haut-niveau relatifs au son et au geste sont plus liés au modèle de synthèse que ne le sont les paramètres physiques du modèle ; on les appelle **méta-paramètres**. Ces méta-paramètres peuvent être décrits à l'aide d'une courbe, que l'on peut décaler (selon les deux axes), déformer ; ce peuvent aussi être des paramètres descriptifs de la courbe — tels que les parties de la courbe [Jensen, 1999a], l'ordre d'approximation polynomiale pour des courbes splines, par exemple. Quand on manipule les méta-paramètres à l'aide de profils prédéfinis, on peut les considérer comme une configuration de *mapping* particulière. Dans ce cas, lorsque le profil est modifié, le *mapping* est lui aussi modifié. Pour qualifier ces méta-paramètres comme paramètres perceptifs du son, leur manipulation doit avoir un effet perceptif sur le son.

Descripteurs psychoacoustiques

Les attributs de la perception auditive humaine sont appelés **paramètres psychoacoustiques**. Ils peuvent être extraits d'une analyse temps-fréquence du signal. Certains correspondent à la caractérisation de l'instrument de musique ou de l'objet sonore à un micro-niveau — hauteur tonale, sonie, timbre, espace auditif — d'autres correspondent à la performance musicale à un niveau plus global — par exemple le vibrato, les modes de jeu tels que le *legato*, le rythme ou le timbre. Parmi ceux que la littérature en psychoacoustique propose comme descripteurs du son, nous gardons un petit sous-ensemble de paramètres, déjà utilisé dans d'autres études. Ce sous-ensemble donnera des accès à l'expressivité. Il définit un espace à quatre dimensions : la hauteur, la sonie, le timbre et l'espace. Le timbre est encore considéré comme ayant plusieurs dimensions, parmi lesquelles on utilise souvent la brillance et le logarithme du temps d'attaque.

On peut aussi utiliser plus de descripteurs du timbre, tels qu'un indicateur de rugosité, le rapport signal-sur-bruit ; pour des sons quasi-harmoniques ou harmoniques, on peut aussi utiliser le vibrato, la position des formants, la synchronie des partiels lors de l'attaque, le *jitter* et le *shimmer* des harmoniques — les deux étant à la limite entre paramètres de signal et paramètres psychoacoustiques, mais ayant un sens dans la caractérisation de sources — l'harmonicité des partiels, de même que le rapport entre puissances des harmoniques paires et puissances des harmoniques impaires. Cet ensemble de paramètres utilisé pour décrire le timbre comporte des redondances et n'est pas constitué de dimensions indépendantes, mais correspond à une description efficace du signal et de sa perception, aidant à l'analyser aussi bien qu'à le synthétiser.

Pour ces quatre types de paramètres (physique, de signal, méta et psychoacoustiques), on veut tenir compte de leur évolution du son. Par exemple, la dérivée de la sonie permet de différencier les portions d'attaque et de décroissance de sons ayant le même timbre ; cela permet de conserver ou d'enlever l'axe temporel [Drame *et al.*, 1998].

4.3 Applications utilisant des descripteurs

Nous présentons maintenant un ensemble de domaines d'applications faisant appel à des descripteurs. Il s'agit de la transformation sonore, de l'analyse-synthèse couplée avec le contrôle gestuel, de la définition de formats de données représentant le son, utilisés pour le codage, de la classification et de la recherche de sons dans une base, de la segmentation de signaux sonores, et enfin de la transcription automatique de partitions. En présentant ces applications, nous pourrons mieux concevoir

à quoi les descripteurs servent habituellement, et nous pourrons choisir lesquels nous seront utiles pour les effets adaptatifs.

4.3.1 Transformation sonore

Afin d'effectuer des transformations sonores de haute qualité, il peut être nécessaire d'extraire certains descripteurs du son, et de les respecter lors de transformations. La dimension temporelle, lorsqu'elle est modifiée par un simple ré-échantillonnage, ne respecte pas les fréquences. lorsqu'elle est effectuée par vocodeur de phase ou modèle additif, les modulations d'amplitude et de fréquences ainsi que l'enveloppe spectrale ne sont pas préservés. Aussi, il faut faire usage d'une bonne connaissance du signal pour en extraire toutes ces informations, les faire disparaître du signal de départ (par exemple, supprimer le vibrato, supprimer le trémolo) avant de modifier l'échelle temporelle, puis les appliquer au signal dilaté ou contracté. Il en va de même pour les attaques, si l'on veut les préserver lors du changement d'échelle temporelle.

FIG. 4.2 – *Diagramme du schéma Analyse–Transformation–Synthèse.*

D'autre part, les méthodes utilisent fréquemment des descripteurs temporels pour fonctionner correctement. Ainsi, la méthode PSOLA et l'analyse additive on besoin de connaître la fréquence fondamentale pour bien fonctionner. PSOLA l'utilise pour bien synchroniser les grains de synthèse, l'analyse additive l'utilise pour bien choisir la taille du grain d'analyse avant bourrage de zéros, afin que la TFCT soit la plus précise. De même, la fréquence fondamentale est utilisée dans la méthode du cepstre pour déterminer précisément la quéfrence de coupure (*cf.* sec. 2.4.4). Le calcul de l'amplitude par RMS est optimal pour un nombre entier P de périodes. On le voit donc, l'analyse gagne beaucoup à utiliser des descripteurs pour bien se paramétrer.

4.3.2 Analyse, synthèse et contrôle gestuel

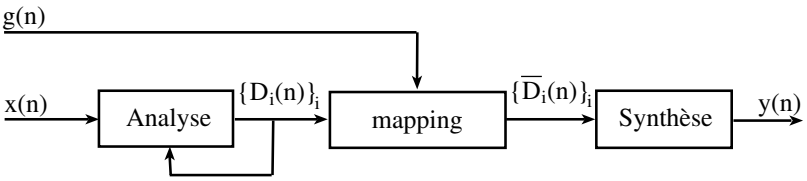


FIG. 4.3 – *Diagramme du contrôle gestuel de modèles de synthèse.*

Dans le cadre de la synthèse sonore numérique, nous présentons la chaîne de traitement de l'information allant du geste physique au son numérique. Ensuite, nous présentons des espaces de contrôle de la hauteur et du timbre, décrivons leurs limites et proposons des solutions pour y pallier.

Depuis l'avènement du temps-réel, on cherche à piloter au mieux des modèles de synthèse sonore numériques à partir de gestes. Dans ce but, on construit une chaîne de traitement de l'information (*cf. fig.* 9). Dans cette chaîne, on fait apparaître d'un côté ce qui a trait à l'interprétation et à l'intention (à gauche, le geste), et de l'autre côté ce qui a trait à l'instrument de synthèse. Le premier se perçoit comme un geste, une intention, le second comme un son. On comprend le son en l'intégrant depuis la perception : à un certain niveau d'abstraction (deuxième niveau), on a une représentation (psychoacoustique) bien plus parlante que le signal lui-même (premier niveau), avant ce qui concerne la cognition (mémoire à plus long terme et cognition, troisième niveau).

type de trame	type de données	méthode de synthèse appropriée
1FQ0	estimation de F_0	oscillateur par table d'onde à cette fréquence
1TRC	lignes spectrales	synthèse additive
1STF	trame TFCT	synthèse par FFT^{-1}
1TDS	échantillons (domaine temporel)	lecture des échantillons
1LPC	coefficients de LPC	modèle source filtre

TAB. 4.1 – *Descriptif des données du format SDIF.*

Lorsqu'on joue d'un instrument que l'on découvre, on cherche à comprendre le son par l'intention, par le jeu que l'on peut en avoir. Deux intentions coexistent, l'une dans le geste, l'autre dans l'espace perceptif. Les indices perceptifs ont ceci d'intéressant qu'ils peuvent être mis en lien directement avec le geste. Ceci permet de jouer directement sur le son par le geste, et de fusionner le *feedback* auditif (la perception que l'on a du son) avec l'espace du geste.

Espace de contrôle du timbre (de sons instrumentaux)

L'espace de timbre correspond à un espace de contrôle du son très utilisé en musique contemporaine, et encore plus en musique électroacoustique. Le premier exemple d'espace de timbre est donné par Grey en [Grey, 1975] : pour une hauteur donnée, il utilise un espace à deux dimensions : brillance (descripteur du timbre) et sonie. Wessel [Wessel, 1979] fait aussi appel à des espaces de timbre pour proposer des trajectoires musicales. Beauchamp [Beauchamp, 1982], effectue de la synthèse contrôlée à partir de la sonie et de la brillance : il utilise une synthèse additive avec une base réduite d'harmoniques, obtenue par algorithmes génétiques. McAdams, Cunibile proposent un contrôle du timbre par le logarithme du temps d'attaque, le flux spectral et la brillance [McAdams and Cunibile, 1992]. Métois contrôle la hauteur, la sonie, et le timbre (via la brillance) dans un espace d'état [Métois, 1996]. Wessel, Drame and Wright [Drame *et al.*, 1998] contrôlent le son à partir de la hauteur, la sonie et la brillance ; les paramètres de la synthèse additive sont calculés par réseaux de neurones à partir des paramètres de contrôle. Jehan et Schoner contrôlent la hauteur, la sonie et le timbre (via la brillance), les données formant des agrégats pour un meilleur contrôle des différentes parties d'un son [Jehan and Schoner, 2001a; Jehan and Schoner, 2001b]. On peut aussi créer un espace de contrôle de la couleur sonore [Slawson, 1985], exploré par exemple en utilisant un synthétiseur de voix chantée contrôlant la hauteur et les formants dans le triangle vocalique [Kessous, 2003; Arfib *et al.*, 2002b]. Un modèle de timbre instrumental complet a été proposé [Jensen, 1999b] prenant notamment en compte les variations des harmoniques ou partiels en fréquence et en module autour d'une valeur ou courbe moyenne. Cependant, son contrôle gestuel n'a pas encore été clairement exploité.

4.3.3 Format de données de représentation spectrale

Les représentations spectrales du son sont utilisés dans de nombreuses applications, et formalisées dans plusieurs standards ou formats de fichier, dont MPEG-4, SDIF, et aussi MPEG-7. Le format de fichier **SDIF** (pour *Sound Description Interchange Format*) est un format de description de haut niveau du signal, où des informations spectrales sont données, telles que les lignes spectrales, l'enveloppe spectrale (via les coefficients de LPC). Le qualificatif "haut-niveau" est à prendre ici dans le sens "haut niveau de signal" et non pas perceptif. Le tableau *tab.* 4.1 décrit les types de trames inclues dans le fichier, les types de données auxquelles cela correspond, et la méthode de resynthèse adéquate. Ce format de fichier est utilisé dans de nombreux logiciels (au CNMAT, *Additive* à l'IRCAM, *Max/MSP*). Un lien a même été établi entre le format SDIF et le format MPEG-4 [Wright and Scheirer, 1999; Wright *et al.*, 1999].

4.3.4 Codage : le standard MPEG-7

L'acronyme **MPEG** signifie *Moving Picture Experts Group*. Ce groupe de l'organisme ISO/IEC est en charge du développement de standards de représentations codées de signaux vidéo et audio-numériques. Etabli en 1988, le groupe a produit les standards :

- MPEG-1, standard que supportent le Vidéo CD et le MP3 (*MPEG-1 Layer 3*) ;
- MPEG-2, standard sur lequel se basent la télévision numérique et le DVD ;
- MPEG-4, standard multimédia pour le Web fixe et mobile, en fin de développement ;
- MPEG-7, standard pour la description et la recherche de contenu audio et visuel, en fin de développement ;
- MPEG-21, nouveau standard multimédia en cours de formalisation depuis juin 2000.

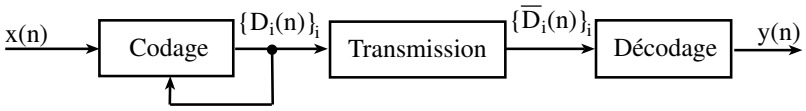


FIG. 4.4 – *Diagramme du codage-décodage de signal sonore.*

Le standard MPEG-7 correspond parfaitement au genre de formalisation qui nous intéresse, puisqu'il doit prendre en compte les caractéristiques (descripteurs) permettant de différencier les sons. Ce standard permettra la compression son et image avec une haute définition, jusqu'ici non encore atteinte. Il utilise des ensembles de descripteurs du timbre comprenant à la fois des descripteurs de signaux harmoniques et des descripteurs de signaux percussifs [Peeters *et al.*, 2000]. Les descripteurs de signaux harmoniques utilisés sont extraits des études de [Krumhansl, 1989; McAdams and Cunibile, 1992] :

- lat : logarithme du temps d'attaque ; il s'agit du logarithme en base décimale de la durée entre l'instant où le signal commence et le minimum entre l'instant où le signal est maximum et l'instant où le signal atteint sa portion stable (*sustain*)

$$lat = \log_{10}\left(\min\left(t_{max}, t_{stable}\right) - t_{debut}\right)$$

- hsc : centroïde spectral harmonique (en échelle linéaire) ;
- hss dispersion harmonique spectrale (*harmonic spectral spread*) : moyenne de la déviation standard pondérée de l'amplitude des pics harmoniques du spectre, normalisés par le hsc, sur la durée totale du son ;
- hsv : variation spectrale harmonique, définie comme la moyenne sur la durée du son du complément de la corrélation normalisée entre l'amplitude (échelle linéaire) des pics harmoniques du spectre sur deux trames adjacentes ;
- hsd : moyenne entre l'amplitude des pics harmoniques du spectre et l'enveloppe spectrale globale.

Les descripteurs de signaux percussifs sont extraits des études de [Lakatos, 2000] :

- lat : logarithme du temps d'attaque ; il 'agit du logarithme en base décimale de la durée entre l'instant où le signal commence et l'instant où le signal est maximum ;
- tc : centroïde temporel ; moyenne pondérée de l'énergie du temps moyen du signal
- sc : centroïde spectral ; moyenne pondérée de l'amplitude des composants du spectre de puissance

4.3.5 Classification, recherche de sons

Les programmes de classification automatique et de recherche d'extraits sonores dans des bases de sons se basent sur une comparaison du contenu de la référence que l'on cherche avec le contenu des sons de la base. On peut disposer du son de référence, ou d'une version chantonnée (*query by humming*), démarche très développée depuis une dizaine d'années). A partir de celui-ci, on extrait des descripteurs, qui vont être utilisés pour la classification ou pour la recherche.

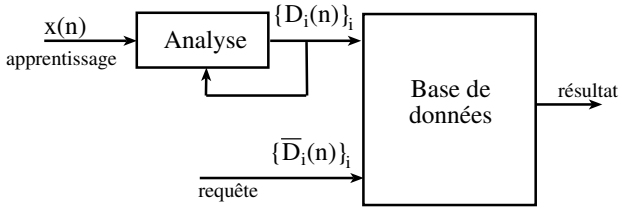


FIG. 4.5 – *Diagramme de la classification et recherche de sons dans une base.*

Les méthodes de classification automatique d'extraits sonores s'appuient aussi sur des représentations perceptives [Scheirer and Slaney, 1997; Martin and Kim, 1998; Wold *et al.*, 1996]. Chaque auteur utilise un ensemble de descripteurs qu'il éprouve sur une base de sons, ce qui lui permet d'évaluer la qualité de la classification, en comparaison de celle qu'il réalise lui-même à l'écoute des sons. Quelques exemples dont nous nous sommes inspiré pour définir notre ensemble de descripteurs sont donnés.

Scheirer et Slaney [Scheirer and Slaney, 1997] proposent ainsi un discriminateur parole-musique pour des applications d'archivage radiophoniques. Ce discriminateur utilise l'énergie de modulation à 4 *Hz*, caractéristique de la parole, le pourcentage de trames temporelles de basse énergie, le point de roulement du spectre, le centre de gravité spectral, le flux spectral, le taux de passage par zéros, la magnitude du résidu de resynthèse cepstrale, et la pulsation rythmique.

Martin et Kim [Martin and Kim, 1998] effectuent l'identification d'instruments de musique par taxinomie hiérarchique, basée sur des descripteurs acoustiques et un modèle d'audition. Les descripteurs utilisés ne sont pas tous perceptifs, mais amènent une classification à l'aide de modèles statistiques et selon des critères perceptifs (reconnaissance d'une voix, d'un instrument). Ainsi, ils utilisent le corrélogramme *log-lag* (logarithme du temps) comme représentation perceptive, obtenue par autocorrélations à court-terme appliquées aux sorties d'un banc de filtres *gammatone*. De cette représentation, ils extraient : la hauteur, le vibrato et le *jitter*, l'enveloppe spectrale, le centroïde spectral, l'intensité, le trémolo, l'asynchronie des harmoniques aux attaques, l'inharmonicité des partiels, la balance paire/impaire, les moyennes et variances de certains descripteurs ainsi que des ratio entre valeurs moyennes durant l'attaque sur valeur après l'attaque. L'ensemble comprend au total 31 descripteurs.

Fraser et Fujinaga [Fraser and Fujinaga, 1999] proposent un système de reconnaissance de timbre en temps-réel analysant le transitoire d'attaque de sons d'instruments acoustiques. L'évolution dynamique du spectre est quantifiée à l'aide des vitesses de l'intégrale, du centroïde, de la variance, du coefficient d'asymétrie du spectre et de la vitesse du *pitch* estimé. La moyenne et la variance de ces cinq paramètres sont aussi utilisées. Ce système augmente significativement la reconnaissance par rapport à un système ne s'appuyant que sur la portion stable des notes (entre 10 et 20 %).

Afin de vérifier la représentativité des descripteurs du standard MPEG-7 dans le cadre de recherche d'extraits sonores, deux expériences ont été menées, l'une à l'**IRCAM** avec le **Studio-On-Line** [IRCAM, 2000] (recherche un extrait sonore à partir de ses critères perceptifs), l'autre à Barcelone au **Music Technology Group** sous forme d'une extension au programme **SMS** [Serra, 1996].

4.3.6 Segmentation de flux sonores

Enfin, la segmentation de flux audionumériques en atomes ou éléments unitaires (notes, grains, etc.) fait elle aussi appel à des descripteurs de signal et à des descripteurs perceptifs [Rossignol *et al.*, 1998b; Tzanetakis and Cook, 1999]. Il s'agit de segmenter le signal en portions temporelles ayant des propriétés les rendant identifiables en tant qu'unités temporelles (uniformes ou d'évolution quantifiée).

Cette segmentation peut se faire avec ou sans contrainte, c'est-à-dire avec une connaissance du

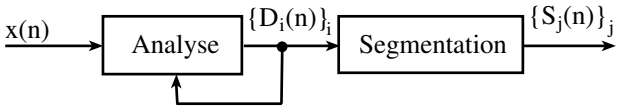


FIG. 4.6 – *Diagramme des systèmes de segmentation de flux sonores.*

signal. Les contraintes peuvent porter sur le type de son (harmonique, inharmonique, bruit), sur sa qualification en terme de flux (monophonique, polyphonique), en terme de trajets (déplacements pour un signal multicanal), sur la rapidité d'évolution du signal, qui détermine la qualité d'analyse, en terme de mode de jeu (différentes articulations entre notes existent, ce qui pose le problème de leur reconnaissance et de leur différentiation, à l'aide notamment de l'articulation en fréquence mais aussi en amplitude). Les contraintes sont un atout pour la segmentation dès lors qu'elles permettent de mieux régler les paramètres de l'analyse, mais elle sont aussi une limitation pour l'automaticité de la segmentation, dès lors qu'on ne sait pas qualifier le son avant analyse.

Les attributs des segments sont qualitatifs et quantitatifs : ainsi, on utilise une étiquette ou un indicateur pour l'attribut qualitatif, et une valeur pour l'attribut quantitatif. Les attributs qualitatifs sont de type :

- harmonicité : harmonique, inharmonique, bruité ;
- type de source : monophonique, polyphonique ;
- avec ou sans vibrato (mode de jeu) ;
- avec ou sans trémolo (mode de jeu) ;
- articulation : legato, pizzicato, glissando, etc. (mode de jeu) ;
- la rugosité.

Les attributs quantitatifs sont soit des données statistiques d'un descripteur sur le segment (ce descripteur n'étant pas uniquement l'un de ceux utilisés pour la segmentation), soit les paramètres d'un mode de jeu :

- somme sur le segment, ou moyenne sur le segment (moment d'ordre 1) d'un descripteur ;
- variance sur le segment (moment d'ordre 2) d'un descripteur ;
- coefficient d'asymétrie sur le segment (moment d'ordre 3) d'un descripteur ;
- coefficient d'aplatissement sur le segment (moment d'ordre 4) d'un descripteur ;
- la phase à l'origine, la fréquence moyenne et l'amplitude moyenne du vibrato ;
- la phase à l'origine, la fréquence moyenne et l'amplitude moyenne du trémolo ;

Ces deux ensembles ne sont pas exhaustifs.

4.3.7 Transcription automatique

La transcription automatique de partition consiste à pouvoir analyser un extrait musical de manière à en écrire la partition, et ceci avec le moins d'intervention humaine possible dans le réglage de l'analyse. Cela signifie qu'il faut pouvoir estimer la fréquence fondamentale d'un son pur, par exemple par la méthode de l'histogramme [Schrœder, 1968], par autocorrélation ou autocorrélation étroite [Brown and Zhang, 1991; Brown, 1992], avec un banc de filtres à Q constant [Brown, 1991], par analyse de la TFCT [Gibiat and Jardin, 1992], par une analyse additive avec pistage des partiels amélioré par utilisation de chaînes de Markov cachées [Doval and Rodet, 1991; Doval, 1994] ou d'une procédure comparative dans les deux sens de propagation (*two-way mismatch procedure*) [Maher and Beauchamp, 1994]. Ensuite, il faut pouvoir extraire les fréquences fondamentales d'une mixture de sons purs, par exemple d'un duo de deux instruments ayant des tessitures sans intersection [Maher, 1990]. L'information de flux auditifs peut servir à construire un modèle perceptif [Dixon, 1996] permettant la transcription de sons d'un instrument polyphonique (guitare, piano). Des modèles avec contraintes permettent une séparation assez précise des partiels découlant de notes distinctes en calibrant un banc de filtres à Q constant [Mani and Nawab, 1995]. Des modèles statistiques tels que les modèles bayésiens en polyphonique [Walmsley *et al.*, 1999] ont été utilisés. D'autres modèles

basés sur les ondelettes ont été développés [Bobrek and Koch, 1998] : leur résolution fréquentielle est quasi identique à celle de l'oreille humaine, et la faible bande de transition permet une bonne détection des *onset*.

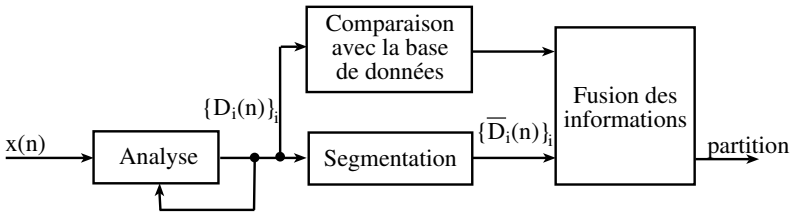


FIG. 4.7 – *Diagramme des systèmes de transcription automatique de partition.*

Sans être exhaustive, cette liste d'illustration des méthodes utilisées montre que les systèmes proposés à ce jour sont nombreux, et tendent vers la complexification pour prendre en compte les attributs perceptifs du son afin de pouvoir séparer les sources[Martin and Kim, 1998] : harmonicité, multiplicité de sources, vibrato, etc. L'utilisation de descripteurs et de méthodes de segmentation tend à devenir la règle pour ces systèmes de transcription automatique, même si le problème est encore loin d'être résolu : les meilleurs systèmes transcrivent deux voix monophoniques ou une voix d'instrument polyphonique. On est encore loin de la transcription d'un ensemble (quatuor, octet, orchestre symphonique) !

4.4 Ensemble des descripteurs utilisés pour les effets adaptatifs

Nous présentons maintenant les programmes utilisés pour l'extraction des descripteurs, ainsi que l'ensemble des descripteurs que nous utilisons.

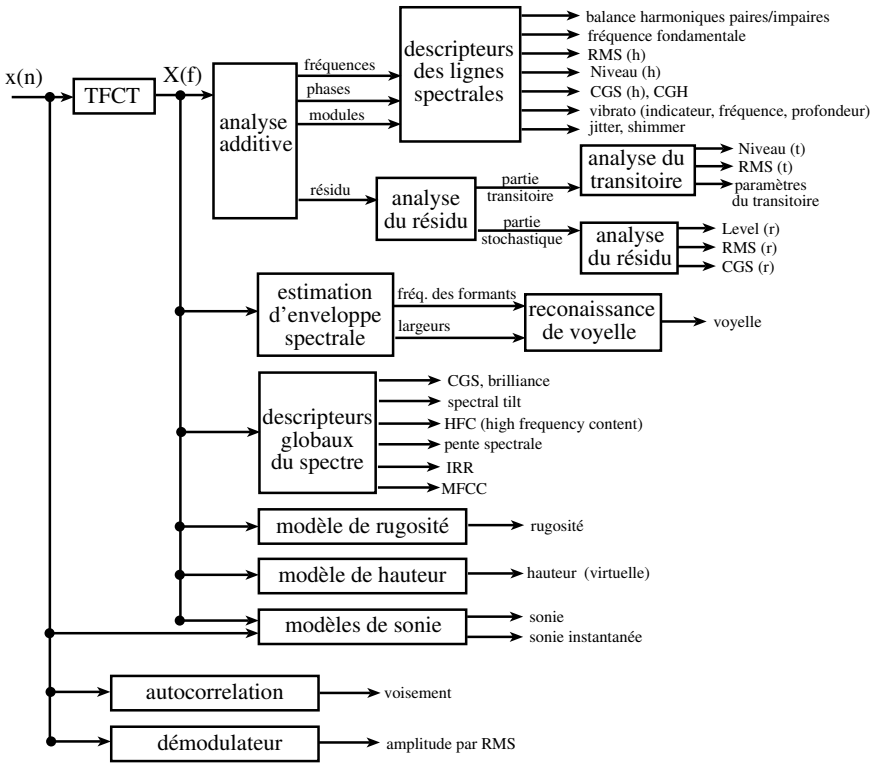


FIG. 4.8 – *Exemple d'ensemble de descripteurs extraits du son.*

4.4.1 Programmes, algorithmes, méthodes

Concernant les programmes utilisés, une partie a été développée par nous-même, l'autre partie provient de l'industrie et de centres universitaires ou de création musicale ; leurs résultats sont intégrés sous forme de courbes à notre programme principal :

- paramètres de bas niveau (RMS, centre de gravité spectrale, indice de voisement, HFC, fonction d'observation de Masri) : routines Maltab (en partie extraites du livre DAFx [Zoelzer, 2002]
- paramètres d'analyse additive : programme *Additive* de l'IRCAM, version *Matlab* du programme SMS, développée dans le cadre du livre [Zoelzer, 2002], *toolbox* `sine-model` en *Matlab*, analyse additive guidée graphiquement (développement par moi-même) ;
- MPEG4 : en *Matlab* (développement par moi-même), d'après la littérature et les basés sur l'analyse additive ;
- segmentation : d'après [Rossignol, 2000], indices de voisement, fonction de rupture de Masri, écrit en Matlab par moi-même ;
- classification : mise en œuvre avec *Matlab* par moi-même : pourcentage de trames de basses énergies [Scheirer and Slaney, 1997], taux de passage par zéros, flux spectral [Scheirer and Slaney, 1997; Lu *et al.*, 2001; Rossignol, 2000] ; magnitude du résidu de la resynthèse spectral, modulation à 4 Hz, point de roulement spectral [Scheirer and Slaney, 1997], ratio de haut taux de passage par zéros [Lu *et al.*, 2001]
- psychoacoustique : hauteur tonale d'après [Terhardt, 1979] mise en œuvre en *Matlab* par moi-même, sonie d'après [Zwicker, 1977], mise en œuvre par Alain Marchioni de l'équipe APIM selon la procédure de Fastl (ISO 532 B) ; paramètres (nombreux) du programme *PsySound* 2 développé par Densil Cabrera [Cabrera, 1999b; Cabrera, 1999a; Cabrera, 2000] pour la recherche et l'enseignement dans les domaines liés à la musique (ce programme effectue l'analyse de paramètres psychoacoustiques tels que la hauteur, la sonie, la brillance ainsi que beaucoup d'autres paramètres selon plusieurs méthodes).

4.4.2 Ensemble des descripteurs utilisés

L'ensemble de descripteurs que nous utilisons, décrit partiellement *fig.* 4.8, est composé des paramètres de bas niveau et de haut niveau utilisés pour la description d'espaces de timbre, de codage ou de classification. Il n'est pas exhaustif, et comme nous le verrons au fur et à mesure de la présentation des descripteurs, des redondances existent. Cependant, ces redondances ne sont pas un problème pour nous, dès lors que nous ne cherchons pas un modèle de signal parfait pour de l'analyse-resynthèse, mais plutôt des courbes de contrôle servant de contrôles variants dans le temps. A cet ensemble de descripteurs, nous ajoutons les dérivées, les intégrales et les moments statistiques, qui peuvent aussi être considérés comme partie intégrante du *mapping*.

4.4.3 Classifications

Nous considérons cinq points d'accès à ces paramètres qui, d'un signal échantillonné (par exemple à $44,1\ kHz$) nous fait connaître certaines propriétés.

Le niveau de description

On considère plusieurs niveaux de description d'un signal, allant du signal lui-même jusqu'à la perception que l'on a de ce signal. On appelle généralement **bas niveau** tout ce qui concerne le signal lui-même, ne nécessitant pas de modèle du son, et **haut niveau** ce qui concerne la perception, dont l'extraction nécessite l'utilisation d'un modèle (que ce soit un modèle d'audition ou un modèle de signal). Ainsi, le RMS est un descripteur de bas niveau, alors que la hauteur perçue est un descripteur de haut niveau. C'est selon cette description bas niveau–haut niveau que nous présentons et définissons par la suite, en sec. 4.5 et 4.6, les descripteurs que nous avons utilisés.

Je propose ici de raffiner cette échelle de niveau en 5 niveaux distincts. Parmi les descripteurs de bas niveau, certains sont **directement calculables** (RMS, CGS, niveau 1), d'autres **indirectement calculables**, à l'aide de méthodes complexes (tels les modules et fréquences des partiels extraits d'une analyse additive), que je qualifierai de deuxième niveau signal (niveau 2). Quant aux paramètres de haut niveau, le plus haut niveau correspond aux paramètres **psychoacoustiques**, contenant les descripteurs de type **partition** (hauteur, sonie, etc., niveau 4) et les descripteurs de type **perception du jeu** (rugosité, vibrato, etc., niveau 5). Entre le niveau 2 et le niveau 4, le niveau 3 correspond à des paramètres faisant sens à la perception, pas tout à fait de niveau 2, pas non plus psychoacoustiques (par exemples le *jitter*, descripteur du timbre).

La méthode d'acquisition

Différents méthodes d'acquisition et de calcul pouvant être utilisées, nous proposons de les passer en revue pour mieux les appréhender. Un descripteur peut être obtenu par **calcul direct**, c'est le cas du RMS. Il peut aussi être la dérivée, l'intégrale, la valeur absolue d'un autre paramètre, la moyenne ou la variance sur le fenêtre de temps : c'est alors un descripteur obtenu par **calcul indirect**. On pourrait le voir comme un *mapping* d'un autre descripteur, mais dès lors que ce descripteur a un sens descriptif du signal, on le considérera comme tel, et non comme transformation d'un autre paramètre.

Le temps d'intégration

Chaque paramètre est calculé et n'a de sens que pour un intervalle de temps donné. Du fait de l'incertitude temps-fréquence (similaire au principe d'incertitude de Heisenberg), aucun paramètre ou descripteur ne peut être considéré comme instantané, mais plutôt obtenu à partir d'un nombre d'échantillons plus ou moins grands (quasi-instantané). Remarquons que cette assertion est fausse pour les échantillons eux-mêmes, puisque chaque échantillon a une valeur pour un temps donné.

Nous considérons donc les paramètres **quasi-instantanés**, correspondant aux paramètres extraits d'une trame temporelle de 512 à 2048 échantillons, soit obtenus pour 12 ms et 46 ms de signal. A l'extrême opposé se trouvent les descripteurs **à très long terme**, qui décrivent une séquence sonore de longue durée par des termes faisant appel à l'analyse musicale : style, tonalité, structure, etc. Au niveau inférieur, les descripteurs **à long terme** informent sur la segmentation du signal : note, présence de vibrato et ses caractéristiques, par exemple. Encore en dessous se trouvent les descripteurs **à moyen terme** : ce peuvent être la moyenne ou la variance d'un paramètre sur N valeurs d'une fenêtre glissante, un coefficient d'aplatissement ou de symétrie, la pulsation (nécessitant des fenêtres de calcul plus grands que les descripteurs quasi-instantanés), la sonie évoluant dans le temps, etc.

Le type

On peut considérer que deux types principaux de descripteurs existent : les **indicateurs**, pouvant prendre un nombre fini et petit de valeurs, par exemple la présence ou non d'un vibrato (1 ou 0), la voyelle reconnue dans une séquence de voix ([a], [e], [i], [o], [u]). Les **valeurs continues** quant à elles peuvent prendre toutes les valeurs imaginables dans un segment donné, par exemple le $RMS \in [0;1]$, $f_0 \in [20;20000]$ Hz. Enfin, les **probabilités** ou **quasi-probabilités** sont des valeurs continues servant, selon des seuils fixés, d'indicateurs. L'indice de voisement peut-être considéré comme tel, du fait que sa valeur indique si oui ou non, un signal de parole est voisé, selon un seuil donné par l'utilisateur. Les indicateurs sont par définition de plus haut-niveau que les descripteurs continus.

La causalité

Selon que l'on fait appel seulement à des échantillons passés (**causal**) ou à des échantillons passés et à venir (**anti-causal**), le calcul des descripteurs appartient à une classe différente. Par exemple, lorsque j'attribue à une trame temporelle la valeur du RMS correspondant, l'attribuer au temps du dernier échantillon rend ce calcul causal (il est alors réalisable en temps réel), alors que l'attribuer au temps de l'échantillon central de la trame rend ce calcul anti-causal. Le temps-réel ne peut en effet utiliser que des paramètres causaux, aussi cette distinction impose des choix selon la mise en œuvre effectuée.

Temps-fréquence

Un descripteur calculé directement à partir de la forme d'onde sera dit **à calcul temporel**, alors qu'un descripteur calculé à partir de la transformée de Fourier de la trame temporelle sera dit **à calcul fréquentiel**. C'est cette classification que l'on utilise par la suite, par soucis de lisibilité.

4.5 Descripteurs de bas niveau

Les descripteurs de bas niveau sont des descripteurs que nous calculons à partir du signal, soit dans le domaine temporel, soit à l'aide d'une représentation temps-fréquence.

4.5.1 Descripteurs temporels

Les descripteurs temporels concernent principalement l'amplitude et l'énergie du signal, ainsi que le taux de passage par zéros et ses variations. L'utilisation du modèle "sinusoïdes + résidu", ou du modèle "sinusoïdes + transitoire + résidu", permet d'extraire l'amplitude et l'énergie de chaque composante. De plus, à partir d'un calcul d'autocorrélation sur une fenêtre glissante, on peut extraire d'autres descripteurs, tels que l'indice de voisement.

Niveau d'énergie $E_N(x,k)$

L'énergie d'un signal peut se mesurer de plusieurs manières : soit de manière physique, soit de manière perceptive. Dans le premier cas, on mesure l'énergie présente de manière objective dans le signal, alors que dans le second (*cf.* sonie, en 4.6.3) on préfère une mesure de l'énergie telle qu'elle est subjectivement faite par la perception humaine. Connaissant le signal d'entrée $x(t)$ discrétisé, on définit l'énergie $E_N(x,k)$ dans le domaine temporel par :

$$E_N(x,k) = \sum_{n=-N/2}^{N/2} x^2(n+k)w^2(n+k) \tag{4.1}$$

avec N pair, la taille de la fenêtre w impaire $(2N+1)$. L'énergie mesurée par RMS correspond à un filtrage RIF des carrés des échantillons. Elle peut aussi se calculer dans le domaine fréquentiel, d'après l'égalité de Parseval, selon l'expression :

$$E_N(x,k) = \sum_{l=0}^{N-1} X^2(\frac{l}{N}) \tag{4.2}$$

L'énergie peut encore se calculer de manière plus complexe en utilisant l'opérateur de Teager, pour des signaux harmoniques [Kaiser, 1990; Kaiser, 1993].

Taux de trames de basses énergies $T_{BE}(x,k)$

Dans des applications de discrimination parole/musique [Lu *et al.*, 2001; Scheirer and Slaney, 1997], le taux de trames de basses énergies est utilisé car les distributions de probabilités de ce descripteur sont bien différentes : loi normale centrée sur 0.3 pour la parole et loi normale centrée sur une valeur négative (en fait, seules les valeurs positives nous intéressent, ce qui implique une courbe décroissante pour la partie positive) pour la musique. Le taux de basse énergie $T_{BE}(x,k)$ aura donc tendance à valoir de manière prépondérante 0 pour de la musique et une valeur non nulle pour de la parole. Sa formule est :

$$T_{BE}(x,k) = \frac{1}{2N_{1s}} \sum_{n=0}^{N_{1s}-1} (\text{sign}(0.5 < E > -E_N(x,k)) + 1) \tag{4.3}$$

avec N_{1s} le nombre de trames temporelles représentant 1 seconde de son, et

$$< E >= \frac{1}{N_{1s}} \sum_{n=0}^{N_{1s}-1} E_N(x,n) \tag{4.4}$$

la moyenne de l'énergie. Un exemple est donné pour un extrait de voix parlée *fig.* 4.9.

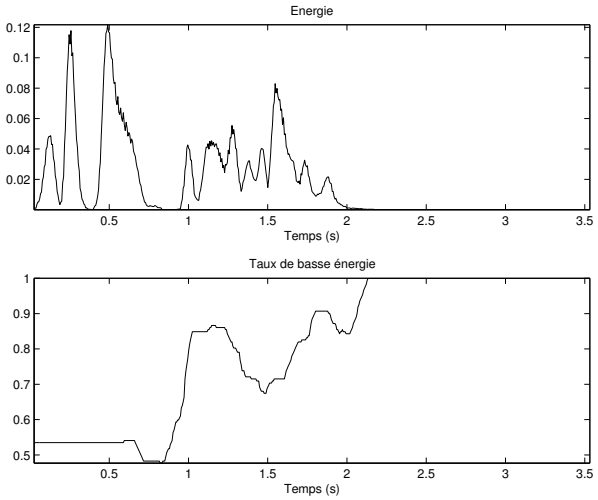


FIG. 4.9 – *Taux de trames de basses énergies de la Piste n°16* ♫♪.

Amplitude $A_N(x,k)$

L'estimation de l'amplitude du signal, et donc de son enveloppe dans le temps se fait à partir de la mesure de l'énergie du signal, en en prenant la racine carrée :

$$A_N(x,k) = \frac{\sqrt{\sum_{n=-N/2}^{N/2} x^2(n+k) w_A^2(n)}}{\sqrt{\sum_{n=-N/2}^{N/2} w_A^2(n)}} \tag{4.5}$$

avec $w_A(n)$ une fenêtre d'analyse symétrique (de Hanning, par exemple). Cette amplitude correspond à un filtrage RIF des carrés des échantillons (tout comme l'énergie), auquel on applique une racine carrée. Lorsque la fenêtre de taille N est trop courte, l'enveloppe suit des variations très rapides qui ne correspondent pas vraiment à l'enveloppe. En effet, si la taille de la fenêtre n'est pas un multiple de la période fondamentale, on ne prend pas une période complète pour le calcul

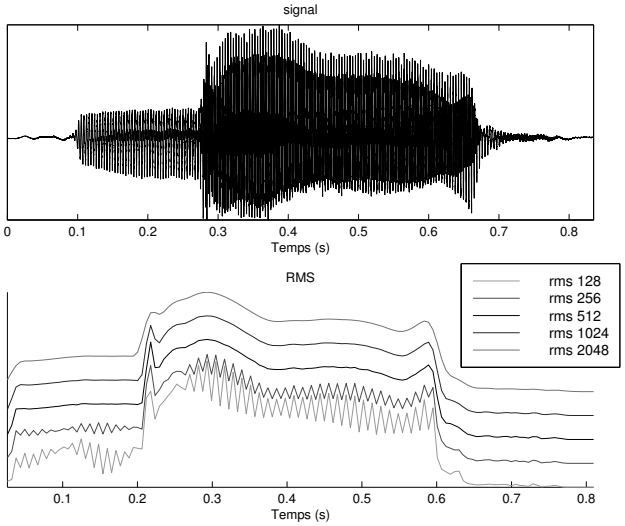


FIG. 4.10 – *Courbes d'amplitude par RMS pour des fenêtres de taille allant de 128 à 2048 échantillons de la Piste n°12 ♫♪.*

de l'énergie, ce qui implique des variations d'une fenêtre à l'autre, et ceci même pour un signal parfaitement harmonique à puissance constante.
Au contraire, si la fenêtre est trop grande, le lissage dû au filtrage est tel que l'on ne détecte plus les modulations tels que le trémolo, qui pourtant correspond à une modulation sinusoïdale lente de l'enveloppe. On comprend ainsi l'intérêt de bien choisir la taille de la fenêtre. Typiquement, pour un signal échantillonné à $44,1\ kHz$, des fenêtres de 512 échantillons sont bien adaptées.

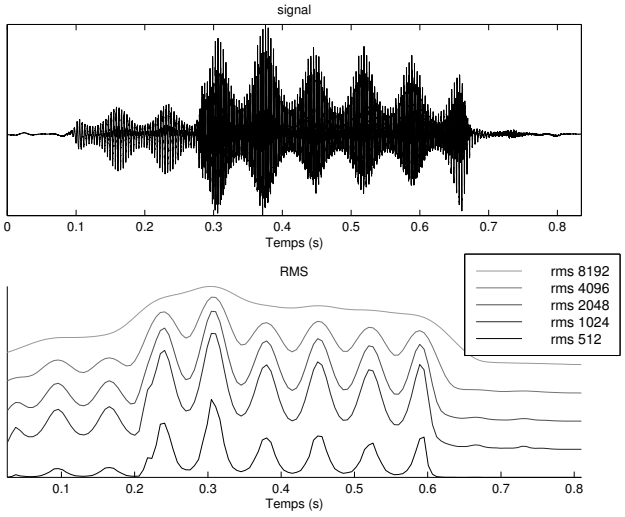


FIG. 4.11 – *Courbes d'amplitude par RMS d'un signal avec trémolo ($F_0 = 260\ Hz$, $F_{trem} = 16\ Hz$) pour des fenêtres de taille allant de 512 à 8192 échantillons de la Piste n°12 ♫♪.*

Remarquons que pour une grande fenêtre (4096 ou 8192 échantillons), le lissage est tel que l'on récupère une bonne estimation de l'amplitude du signal *fig.* 4.10 avant le trémolo. Ceci peut être utilisé pour réaliser un changement de trémolo, dès lors que la profondeur du trémolo n'est pas trop grande (*cf.* sec. 5.2.7). L'estimation de l'enveloppe par RMS peut aussi se faire dans le domaine fréquentiel, dans le cas où on doit l'insérer dans un boucle de traitement FFT/IFFT. Il existe une autre définition de l'enveloppe d'un signal. Idéalement, on peut considérer que c'est la magnitude du signal analytique $x^+(n) = x(n) + j\hat{x}(n)$ avec $\hat{x}(n)$ la transformée de Hilbert de la partie réelle

$x(n)$. Ce descripteur est intéressant pour notre étude. Cependant, il faut savoir que cette définition de l'amplitude ne correspond pas à la perception : l'enveloppe est supposée varier lentement dans le temps et ne contenir aucune information fréquentielle. Aussi, la magnitude du signal analytique conserve des oscillations qui suivent la fréquence fondamentale (sauf pour une sinusoïde pure), ce qui ne correspond pas à la définition d'un détecteur d'amplitude.

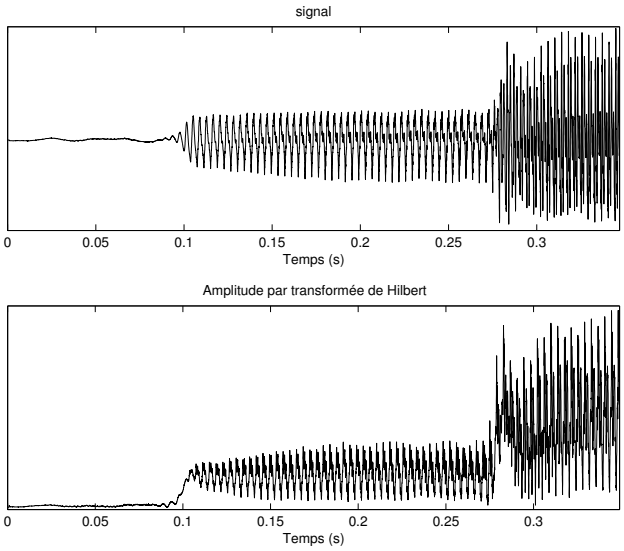


FIG. 4.12 – *Courbes d'amplitude (par fonction de Hilbert) de la Piste n°12* ♫♪.

Amplitudes de la composante déterministe et de la composante stochastique

On peut envisager de calculer séparément les amplitudes des trois composantes (sinusoïdes + transitoire + résidu) ou seulement des deux composantes (harmoniques + résidu, *cf. fig.* 4.14). Dans le modèle additif, l'amplitude de la composante sinusoïdale (déterministe) est définie comme la somme des amplitudes de toutes les harmoniques [Serra, 1996; Amatriain *et al.*, 2001], exprimée en dB :

$$A_s(x,n) = 20\log_{10}\left(\sum_{i=1}^{H} a_i(x,n)\right) \tag{4.6}$$

L'amplitude de la composante résiduelle correspondant à l'énergie du résidu, exprimé en dB [Serra, 1996; Amatriain *et al.*, 2001] :

$$A_r = 20\log_{10}\left(\sum_{n=0}^{M-1} |x_R(n)|\right) = 20\log_{10}\left(\sum_{k=0}^{N-1} |X_R(k)|\right) \tag{4.7}$$

On peut aussi les calculer à partir des signaux resynthétisés en utilisant le RMS (*cf. fig.* 4.14).

Niveaux SPL (SPL_A, SPL_B, SPL_C) Le niveau SPL correspond au niveau de pression du son (*SPL : Sound Pressure Level*) : c'est une mesure objective donnée en décibels (*cf. fig.* 4.15 gauche). Il existe trois pondérations pour le calcul du niveau SPL, en plus du SPL non pondéré. Chaque pondération atténue certaines bandes de fréquences plutôt que d'autres, et dépendent d'une courbe d'isosonie à un niveau donné en dB. Ainsi, la pondération A (*cf. fig.* 4.15 droite) atténue fortement les basses fréquences et atténue un peu les hautes fréquences ; elle est liée à l'isosonie à 40 *phones*. La pondération B (*cf. fig.* 4.16 gauche) atténue beaucoup moins les basses fréquences que la pondération A, et un peu plus les hautes fréquences ; elle est liée à l'isosonie à 70 *phones*. La pondération C

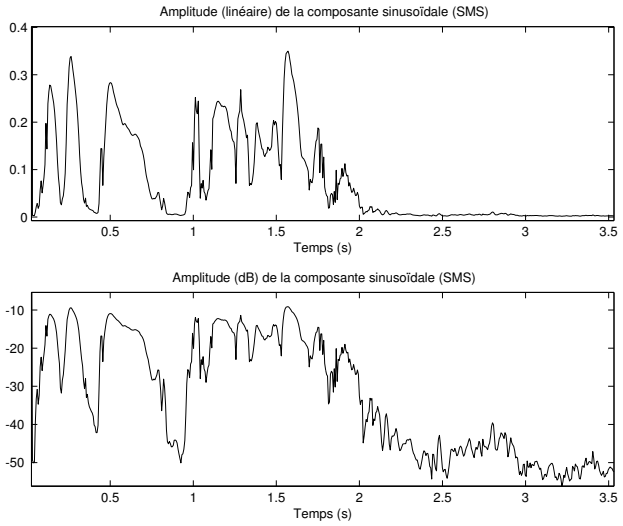


FIG. 4.13 – *Amplitude A_s de la composante sinusoïdale (analyse additive).*

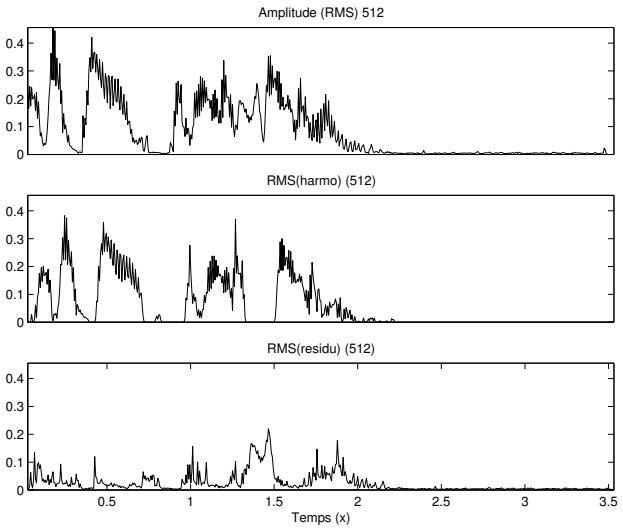


FIG. 4.14 – *Amplitudes de la partie harmonique et de la partie résiduelle calculée par RMS (fenêtres de 512 échantillons) ; séparation par analyse additive Piste n°16 ♫♪.*

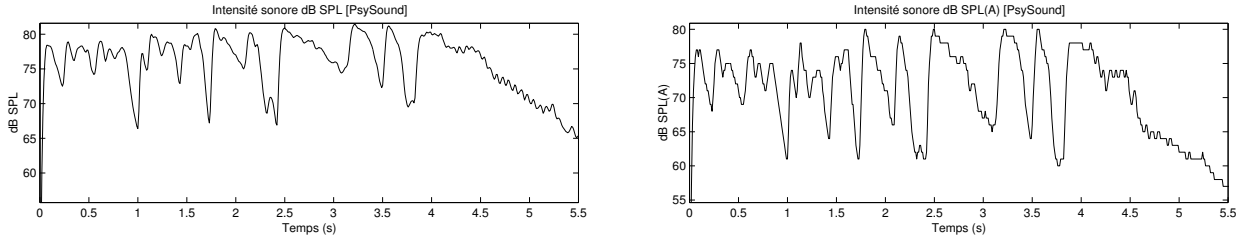


FIG. 4.15 – *Intensité sonore SPL de la Piste n°20 ♫♪, sans pondération (à gauche) et avec pondération A (à droite).*

(*cf. fig.* 4.16 droite) quant à elle est plus proche d'une pondération linéaire, mais toujours avec une petite atténuation dans les basses et dans les hautes fréquences ; elle est liée à l'isosonie à 100 *phones*.

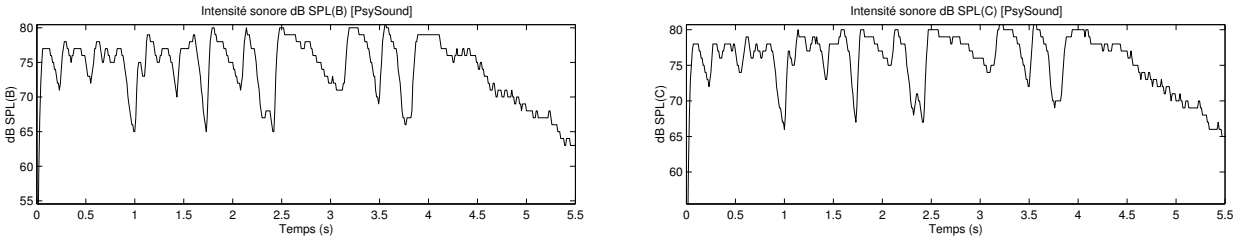


FIG. 4.16 – *Intensité sonore de la Piste n°20 ♫♪ avec pondération B (ou* SPL(B)*) à gauche, et avec pondération C (ou* SPL(C)*) à droite.*

Fonction d'autocorrélation $C_N(y,k)$

La fonction d'autocorrélation n'est pas utilisée usuellement comme descripteur du signal sonore (classification, segmentation). Ceci dit, elle est utilisée pour extraire des descripteurs (indice de voisement, balance des harmoniques paires/impaires), et peut servir comme descripteur de dimension N lorsque l'effet nécessite un tel descripteur (par exemple la conformation du spectre, *cf.* sec. 5.6.3, la conformation de l'enveloppe spectrale *cf.* sec. 5.6.1). La fonction d'autocorrélation se définit comme suit :

$$C_N(y,k) = \sum_{n=0}^{N-1} y(n)y(k-n) \tag{4.8}$$

avec $y(n)$ le grain $x(n)$ fenêtré par une fenêtre $w_A(n)$ symétrique. On peut calculer la fonction d'autocorrélation de manière plus efficace (en temps de calcul) à l'aide de Y, la transformée de Fourier à court-terme de y :

$$C_N(y,k) = \Re\left(FT^{-1}\left(|Y|\right)\right) \tag{4.9}$$

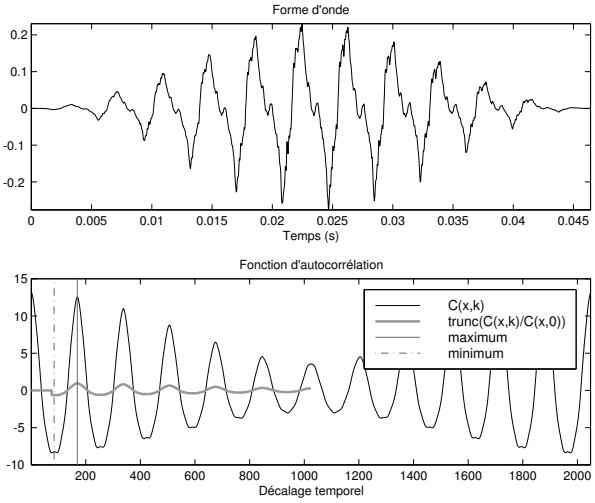


FIG. 4.17 – *Fonction d'autocorrélation pour un grain du son Piste n°12 ♫♪.*

Indice de voisement Vois$_N(x,k)$

L'indice de voisement pour des signaux de voix parlée est obtenu à partir de l'autocorrélation. Le signal $x(n)$ étant réel, la fonction d'aurocorrélation est paire. D'autre part, son maximum est en 0. L'autocorrélation d'un morceau de signal consiste à dire, pour différents décalages, dans quelle mesure le signal et sa version décalée se ressemblent (sont corrélés). Ainsi, lorsque le décalage correspond exactement à la période fondamentale p (le PGCD des périodes des harmoniques, puisque la

fréquence fondamentale est le PPCM des fréquences des harmoniques), on observe un pic de la fonction d'autocorrélation. Ce pic $C_x(p)$ correspond à la puissance des harmoniques. Pour un décalage nul, l'autocorrélation possède un pic (son maximum) qui correspond à la puissance du signal total (harmoniques + résidu). Ainsi, en calculant la fonction d'autocorrélation, puis en calculant le rapport de son second maximum local sur le maximum en 0, on obtient l'indice de voisement.

$$\text{Vois}_N(x,k) = \frac{C(x,p)}{C(x,0)} \tag{4.10}$$

Pour trouver le premier pic en $p \neq 0$, on utilise le fait que la courbe est symétrique pour en oublier la seconde moitié, puis le fait que le maximum est obtenu en 0 et le minimum (négatif) juste après pour supprimer cette première portion de la courbe d'autocorrélation.

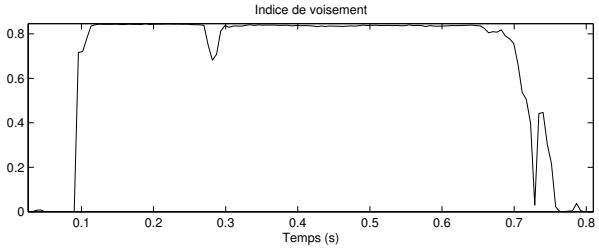


FIG. 4.18 – *Indice de voisement* $\text{Vois}_N(x,k)$ *de la Piste n°18* ♫♪.

Taux de passage par zéro ZCR(x,k)

Le taux de passage par zéros (ou ZCR pour *zero crossing rate*) correspond à la proportion de passage par zéros de la forme d'onde, dans le domaine temporel. Cette mesure est corrélée au centroïde. Soit $zer(x,N)$ la fonction qui compte le nombre de passages par 0 du grain $x(n)$ de taille N. Le taux de passage par zéros est alors :

$$\text{ZCR}(x,k) = \frac{\text{zer}(x,N)}{N} \tag{4.11}$$

On l'utilise pour la différenciation parole/musique [Lu *et al.*, 2001]. Il est corrélé au centre de gravité spectral, et [Kedem, 1986] le considère comme mesure de la fréquence dominante du signal.

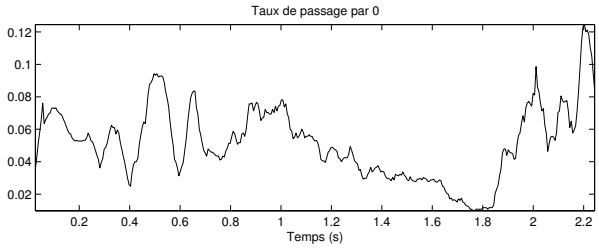


FIG. 4.19 – *Taux de passage par zéros* $ZCR(x,k)$ *de la Piste n°3* ♫♪.

Variations du taux de passage par zéro HZCRR$_N(x,k)$

[Lu *et al.*, 2001] ont montré que les variations du ZCR sont plus discriminantes que le ZCR lui-même. Aussi, ils définissent les hautes variations du ZCR, définies par :

$$\text{HzCRR}_N(x,k) = \frac{1}{2N_{1s}} \sum_{n=0}^{N_{1s}-1} (\text{sign}(ZCR(n) - 1.5 avZCR) + 1) \tag{4.12}$$

avec N_{1s} le nombre de trames temporelles représentant 1 seconde de son, et

$$avZCR \quad = \quad \frac{1}{N_{1s}} \sum_{n=0}^{N_{1s}-1} ZCR(n)$$

la moyenne du taux de passage par zéros sur une fenêtre de 1 seconde. En effet, les signaux de paroles comportent généralement des alternances entre passages voisés et passages non voisés, alors que les signaux musicaux n'ont pas ce genre de structure. Ce descripteur est utile surtout pour discriminer automatiquement de longs signaux où la parole et la musique sont mixés, par exemple de enregistrements d'émissions radiophoniques.

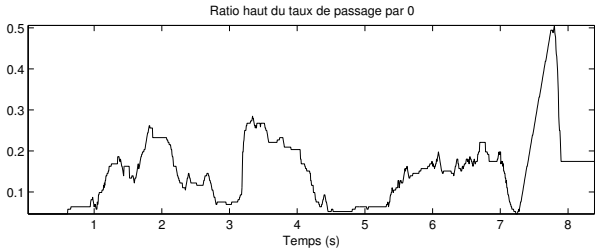


FIG. 4.20 – *Variations du taux de passage par zéros de la Piste n°3 ♫♪.*

4.5.2 Descripteurs fréquentiels

En utilisant un modèle spectral, tel que le modèle de raies spectrales "sinusoïdes + bruit" [Serra, 1996; Verma *et al.*, 1997] pour décrire un son, on peut extraire un bien plus grand nombre de descripteurs que dans le domaine temporel, tels que : la fréquence et le module de chaque harmonique, l'harmonicité et le synchronisme des partiels ; l'enveloppe spectrale du spectre, de la composante harmonique et de la composante résiduelle ; le centre de gravité spectrale et le roulement spectral ; la balance entre les harmoniques paires et impaires ; la pente spectrale ; le contenu en hautes fréquences ; le flux spectral et la différence spectrale, etc.

Amplitudes $a_i(n)$ et fréquences $f_i(n)$ des harmoniques

Lors de la recherche de pics dans le TFCT que l'on utilise pour séparer la partie harmonique du bruit (dans le cas des signaux quasi-harmoniques ou harmoniques) ainsi que pour calculer la hauteur perçue par un algorithme prenant en compte les phénomènes perceptifs de l'homme (tel l'algorithme de Terhardt [Terhardt, 1979]), on doit calculer le trajet des harmoniques jusqu'à un ordre suffisant. Le choix du nombre d'harmoniques et les critères selon lesquels on choisit sont laissés aux programmes plus experts, tels SMS [Serra and Smith, 1990] ou SAS [Desainte-Catherine and Marchand, 1999]. Toujours est-il que ces paramètres (amplitude a_i et fréquence f_i des harmoniques ou partiels d'ordre supérieur) sont utiles à différents calculs d'extraction d'indices.

Fréquence fondamentale F_0

La fréquence de la fondamentale, notée f_0, correspond à une notion de hauteur des sons harmoniques. Tout son peut se décomposer en une somme de sinusoïdes dont l'amplitude varie plus ou moins lentement dans le temps. Dans le cas où le son est harmonique ou quasi-harmonique, la décomposition du son donnera des sinusoïdes prépondérantes dont les fréquences seront toutes

multiples, plus ou moins exactement entiers, d'une fréquence que l'on qualifie alors de fréquence fondamentale. La fréquence fondamentale n'est donc pas égale à celle du premier harmonique si le son n'est pas parfaitement harmonique. En notant f_i les fréquences des harmoniques et a_i leurs amplitudes respectives, on donne la définition suivante pour la fondamentale F_0 :

$$F_0(n) = \sum_{i=1}^{H} \frac{f_i(n)}{i} \frac{a_i(n)}{\sum_{i=1}^{H} a_i(n)} \tag{4.13}$$

d'après [Amatriain *et al.*, 2001]. Bien que pouvant différer de la fréquence fondamentale de quelques comas, on emploie souvent (par abus de langage) autant "pitch" que "hauteur" pour parler de la fréquence fondamentale.

Inharmonicité des harmoniques $H_l(n)$

L'indice d'inharmonicité [Rossignol, 2000] de chaque harmonique est donné par :

$$H_l(n) = \left| \frac{F_l(n) - lF_0(n)}{lF_0(n)} \right| \tag{4.14}$$

Son utilité vient du fait que durant les transitions, la fréquence fondamentale n'a plus vraiment de sens, et les partiels deviennent inharmoniques. On peut alors en mesure l'inharmonicité.

Synchronisme des harmoniques

L'idée d'utiliser le synchronisme des harmoniques provient de ce que le mode de production des harmoniques implique un synchronisme à l'attaque seulement ou tout au long d'une note. Pour les instruments à vent et les bois, les harmoniques restent synchrones. Pour les cordes, leurs modes propres impliquent une désynchronisation progressive des harmoniques, jusqu'à la prochaine note ou le prochain coup d'archet. Dans ses travaux (*cf.* [Dubnov and Tishby, 1996]), Dubnov pose les hypothèses de non gaussianneté et de linéarité des variations de fréquence de la partie harmonique. Il utilise pour cela deux mesures statistiques, le coefficient d'asymétrie et le coefficient d'aplatissement (*skewness* et *kurtosis*, *cf.* sec. 4.7.1).

Déviation des fréquences et des amplitudes des harmoniques (Jitter$_l(n)$, Shimmer$_l(n)$)

D'après Jensen [Jensen, 2001], le *shimmer* correspond au bruit sur l'enveloppe d'amplitude, et le *jitter* au bruit sur la fréquence. Ils servent à modéliser les attaques, *sustain* et relâchement des notes, et sont supposés suivre une distribution gaussienne.

$$\text{Shimmer}_l(n) = \frac{a_k(n) - \hat{a}_l(n)}{\hat{a}_l(n)} \tag{4.15}$$

avec $\hat{a}$ l'amplitude "propre", moyenne, et :

$$\text{Jitter}_l(n) = \frac{f_l(n) - \hat{f}_l(n)}{\hat{j}_l(n)} \tag{4.16}$$

avec $\hat{f}$ l'amplitude moyenne.

Enveloppe spectrale $\mathcal{S}_{env}(x, k)$

Nous avons rappelé dans la premier partie les méthodes d'estimation de l'enveloppe spectrale $\mathcal{S}_{env}(x, k)$ du signal $x(n)$ autour du temps k à l'aide de la LPC (*cf.* sec. 2.4.4) et du cepstre (*cf.* sec. 2.4.4). Avec le modèle additif, on peut donner deux enveloppes spectrale :

– l'enveloppe spectrale de la composante sinusoïdale : enveloppe décrite par les amplitudes et fréquences des harmoniques, ainsi que son approximation [Serra, 1996; Amatriain *et al.*, 2001] :

$$\mathcal{S}^{s}_{env}(x,k) = \{(f_1(k), a_1(k)), (f_2(k), a_2(k)), ..., (f_H(k), a_H(k))\} \tag{4.17}$$

– l'enveloppe spectrale du résidu : approximation du spectre de magnitude du résidu :

$$\mathcal{S}^{r}_{env}(x,k) = \{e_1, e_2, ..., e_q, ..., e_{N/M}\} = \max_k \{|X_R(qM+k)|\} \tag{4.18}$$

Norme du résidu de la resynthèse cepstrale $\mathcal{N}_{env}(x,k)$

Si l'on utilise le cepstre réel pour estimer l'enveloppe spectrale, on a une estimation de l'enveloppe médiane. Cette enveloppe médiane est plus proche du spectre de magnitude pour les sons bruités que pour les sons harmoniques, du fait que les harmoniques sont éloignées, vers le haut pour les pics, vers le bas pour les creux, de l'enveloppe. La norme du résidu de la resynthèse cepstrale est donnée par :

$$\mathcal{N}_{env}(x,k) = \|\mathcal{S}_{env}(x,k) - |X(f)|\|_2 \tag{4.19}$$

On remarque que si la norme est calculée avec le spectre de magnitude et son enveloppe, on retrouve une bonne approximation du RMS (*cf. fig.* 4.21 en haut). Si la norme est calculée avec le spectre en décibels et l'enveloppe en décibels, on a par contre une courbe assez différente, moins corrélée avec le RMS (*cf. fig.* 4.21 en bas).

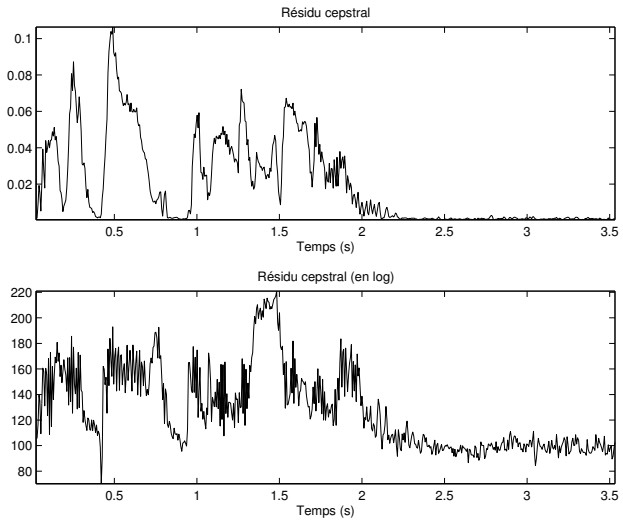


FIG. 4.21 – *Norme du résidu de la resynthèse spectrale (magnitude et dB), Piste n°16 ♫♪.*

Centre de gravité spectrale $\text{cgs}_i(x,k)$ et balance grave/aigu $b_{g/a}(x,k)$

Le centroïde cgs_i est le centre de gravité de la distribution d'énergie du spectre de magnitude de la trame (encore appelé centre de gravité spectral). On peut le considérer comme le point d'équilibre du spectre. Il peut se définir de deux manières, selon si l'on utilise une représentation Temps-Fréquence basée sur une analyse de Fourier à court-terme ou un modèle additif. En notant $X(f)$ la transformée de Fourier à court-terme de $x(n)$, on obtient son expression :

$$\text{cgs}_1(x,k) = \frac{\sum_{l=0}^{N/2} \frac{l}{N}.X(l)}{\sum_{l=0}^{N/2} X(l)} \tag{4.20}$$

Si par contre on utilise une représentation additive avec $a_k, k = 1, ..., H$ les amplitudes des sinusoïdes, le centroïde est alors défini par :

$$\text{cgs}_2(x,k) = \frac{\sum_{l=1}^{H} l.a_l}{\sum_{l=1}^{H} a_l} \tag{4.21}$$

Une troisième définition donnée par [Beauchamp, 1982] utilise une constante c_0 ajoutée au dénominateur de la fraction pour forcer la valeur du centroïde à aller vers 0 lorsque l'amplitude du signal tend vers 0 :

$$\text{cgs}_3(x,k) = \frac{\sum_{l=1}^{N/2} \frac{l-1}{N} X(l)}{c_0 + \sum_{n=1}^{N/2} X(l)} \tag{4.22}$$

La balance grave/aigu $b_{g/a}$ est une approximation du centroïde obtenue à partir d'un calcul de RMS sur le signal fenêtré $y(t)$ et sur sa dérivée :

$$b_{g/a}(x,k) = cgs_4(x,t) = \frac{RMS\left(\frac{dx(n)}{dn}\right)}{2\pi RMS(x(n))} \tag{4.23}$$

Ceci se montre en utilisant l'opérateur de Teager [Kaiser, 1990; Kaiser, 1993; Sussman and Khars, 1996]. En effet, pour un signal harmonique s'exprimant comme une somme de sinusoïdes $x(n) = \sum_{k=0}^{N/2-1} a_k \sin(\Omega_k n)$ avec $\Omega_k = \frac{(k-1)*F_e}{N}$, sa dérivée est $\frac{dx(n)}{dn} = \sum_{k=0}^{N/2-1} a_k \Omega_k \cos(\Omega_k n)$. Le calcul de la dérivée fait donc apparaître le facteur k. Le facteur 2π au dénominateur corrige quant à lui le facteur 2π introduit au numérateur par la dérivée du RMS.

Pour chacune de ces quatre définitions, on peut utiliser le carré (soit du module, soit du signal) au lieu de la valeur elle-même, ce qui donne pour la première définition :

$$cgs_5(x,k) = \frac{\sum_{l=0}^{N/2} \frac{l}{N}.X^2(l)}{\sum_{k=0}^{N/2} X^2(l)} \tag{4.24}$$

Roulement spectral $\mathcal{R}_{spec}(x,k)$

Le roulement spectral ou point de roulement du spectre $\mathcal{R}_{spec}(x,k)$ (*spectral RollOff*) correspond aux 95 pourcentile de la distribution en puissance spectrale. C'est une mesure du coefficient d'asymétrie utilisée pour la discrimination parole / musique [Scheirer and Slaney, 1997].

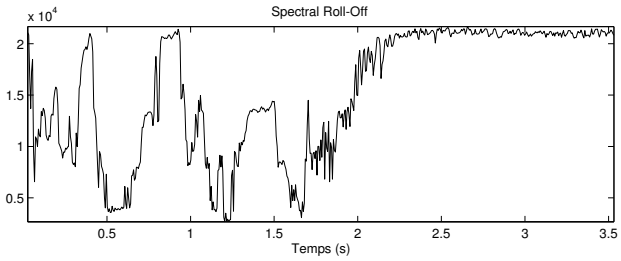


FIG. 4.22 – *Point de roulement spectral $\mathcal{R}_{spec}(x,k)$ de la Piste n°16* ♫♪.

Ce descripteur est assez bien corrélé avec le centre de gravité spectral calculé par TFCT, donné en (4.20), comme on peut le voir *fig.* 4.23. Cependant, les différences entre les courbes des deux descripteurs laissent penser qu'il est intéressant de conserver les deux descripteurs, afin de permettre des contrôles aux comportements à la fois proches en terme d'évolution et différents dans le détail (cf. sec. 4.9).

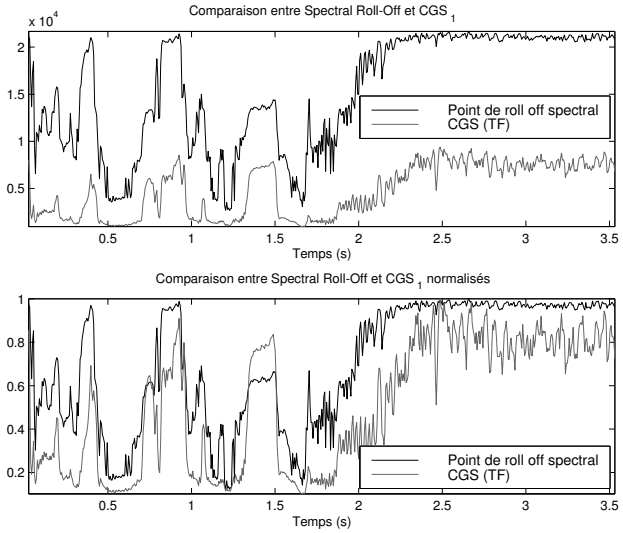


FIG. 4.23 – *Comparaison point de roulement spectral et CGS de la Piste n°16 ♫♪.*

Balance harmoniques paires/impaires $b_{p/i}$

La balance des harmoniques paires/impaires peut se définir comme la puissance des harmoniques paires sur la puissance de toutes les harmoniques, ou comme la puissance des harmoniques paires sur la puissance des harmoniques impaires. Nous utilisons la première définition. En utilisant la fonction d'autocorrélation lorsque le son est bien harmonique, on a la puissance de toutes les harmoniques au second maximum $C(x,p)$ et la puissance des harmoniques paires $C(x,p/2)$ à la demi-période $p/2$.

$$b_{p/i}(x,k) = \frac{C(x,p/2)}{C(x,p)} \tag{4.25}$$

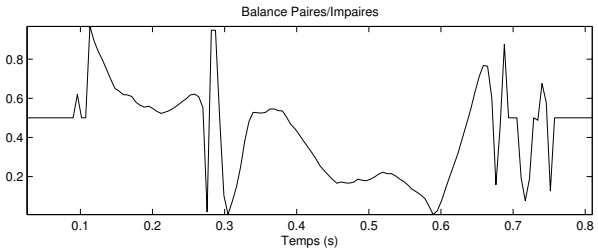


FIG. 4.24 – *Balance des harmoniques paires/impaires $b_{p/i}(x,k)$ de la Piste n°12 ♫♪. On remarque que pour les parties non voisées ainsi qu'au passage entre le [l] et le [o], l'indicateur n'est plus fiable, du fait que le son ne soit plus harmonique, ou alors fortement bruité.*

Contenu en hautes fréquences $HFC(x,k)$

Le contenu en hautes fréquences, ou *HFC* (*High Frequency Content*) est défini par la formule :

$$HFC(x,k) = \sum_{l=1} N/2l|X(l)|^2 \tag{4.26}$$

avec $X(f)$ la TFCT de $x(n)$.

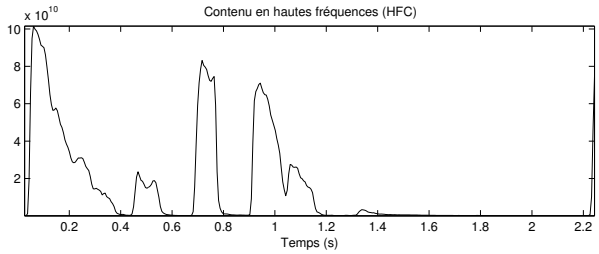


FIG. 4.25 – *Contenu en hautes fréquences $HFC(x,k)$ de la Piste n°18* ♫♪.

Détection de début de note $I_{note}(n)$ (*onset*)

Un indicateur de début de note $I_{note}(n)$ basé sur le contenu en hautes fréquences a été proposé par [Masri and Bateman, 1996] et utilisé par la suite par [Jehan, 1997]. Il s'agit de comparer une fonction d'observation F_{Masri} à un seuil T_D ; lorsqu'elle est supérieure au seuil, alors un début de note est détecté.

$$F_{Masri} = \frac{{HFC_r}^2}{HFC_{r-1}E_r} \tag{4.27}$$

avec r l'indice temporel de la trame courante (et $r-1$ la trame précédente) et $E_r = \sum_{k=1}^{N/2}\left(|X(k)|^2\right)$ l'énergie de la trame courante.

Comme le relève [Rossignol *et al.*, 1998a], la difficulté consiste à bien seuiller cette fonction d'observation de Masri. Nous choisissons la solution de calculer le rapport (4.27) sans le seuiller, afin de laisser l'utilisateur la possibilité d'utiliser la courbe en tant que telle, ou de la seuiller comme bon lui semble.

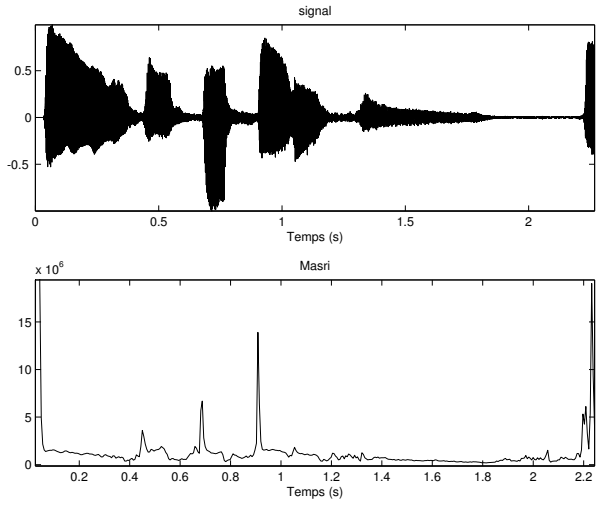


FIG. 4.26 – *Indicateur de début de notes de la Piste n°3* ♫♪.

A titre comparatif, voici représentés ensemble le log en base 10 de F_{Masri}, le contenu en hautes fréquences HFC et l'indice de voisement (*cf. fig.* 4.27).

Flux spectral : Flux$(X, r+1)$

Le flux spectral correspond à la norme 2 de la différence de magnitude des spectres à court-terme de deux trames temporelles successives $X_r X_{r+1}$:

$$\text{Flux}(X, r+1) \quad = \quad \|X_{r+1} - X_r\|_2 \tag{4.28}$$

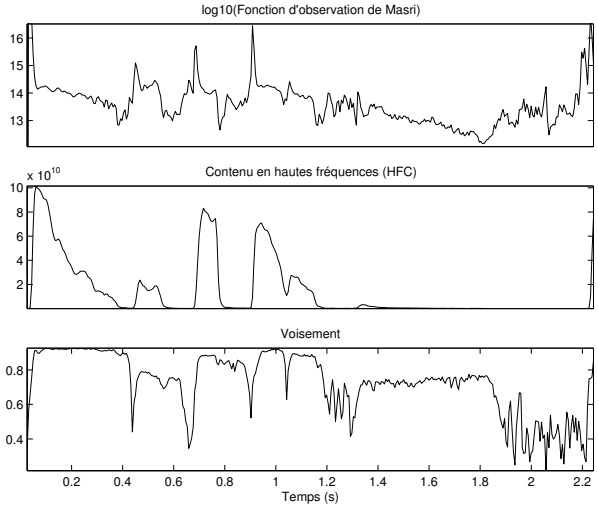


FIG. 4.27 – *Indicateur de début de notes de la Piste n°3 ♫♪.*

$$= \sqrt{\sum_{k=0}^{N/2} \left(\|X_{r+1}(k)\| - \|X_r(k)\|\right)^2} \tag{4.29}$$

Remarque : la transformée de Fourier à court-terme $X^t(k)$ doit être normalisée en énergie.

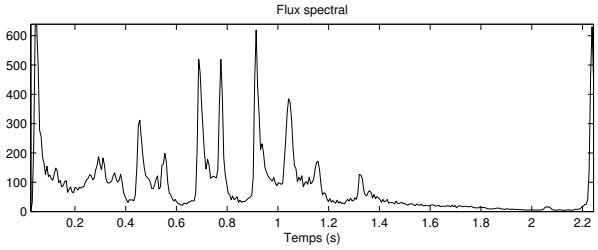


FIG. 4.28 – *Flux spectral de la Piste n°3 ♫♪.*

On peut aussi calculer le flux spectral non plus sur le spectre de magnitude mais sur son enveloppe $\mathcal{E}$ (comme illustré *fig.* 4.29 gauche) [Rossignol, 2000] :

$$Flux_{env} = \sqrt{\sum_{k=0}^{N/2} \left(\mathcal{E}_{r+1}(k) - \mathcal{E}_r(k)\right)^2} \tag{4.30}$$

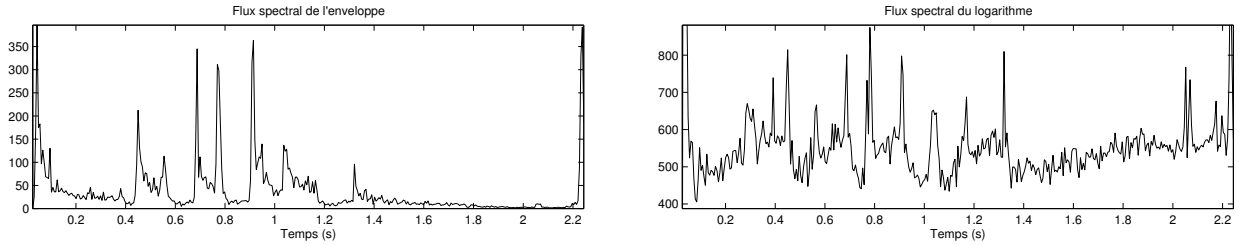


FIG. 4.29 – *Flux spectral de l'enveloppe de la Piste n°3 ♫♪ en échelle linéaire (à gauche) et en échelle* log *(à droite).*

On peut enfin calculer le flux spectral non pas sur le spectre de magnitude mais sur son logarithme [Lu *et al.*, 2001] :

$$Flux_{log} = \frac{1}{N_t N_f} \sum_{k=0}^{N/2} (\log(A_{r+1}(k) + \varepsilon) - \log(A_r(k) + \varepsilon))^2 \tag{4.31}$$

Ne connaissant pas forcément la longueur de signal (et donc le nombre de trames temporelles N_t) à l'avance lors des applications temps-réel, on préférera une formulation ne faisant pas intervenir N_t : il ne s'agit après-tout que d'une constante de normalisation (*fig.* 4.29 droite).

Le flux spectral croisé Flux(X^g, X^d, r)

Il s'agit d'un calcul identique à celui du flux spectral, mais au lieu de le faire sur le spectre à court-terme de deux instants successifs, il se fait au même instant entre les signaux de deux canaux différents. On peut le calculer sur le spectre entier :

$$\text{Flux}(X_g, X_d, r) = \|X_r^g - X_r^d\| \tag{4.32}$$

$$= \sqrt{\sum_{k=0}^{N/2} (X_r^g(k) - X_r^d(k))^2} \tag{4.33}$$

mais aussi sur les enveloppes estimées par la méthode du cepstre (flux des résidus de la resynthèse cepstrale), ou encore en échelle logarithmique, comme précédemment pour le flux spectral.

Différence spectrale DiffSpec$(X, r+1)$

La différence spectrale est définie par :

$$\text{DiffSpec}(X, r+1) = \text{sign}(E_r - E_{r-1}) \left(\sum_{k=2}^{N/2+1} k(X_r(k) - X_{r-1}(k))^2 \right) \tag{4.34}$$

Cette différence tient compte du signe (*cf. fig.* 4.30 gauche). On peut aussi la calculer sans le signe de la différence (*cf. fig.* 4.30 droite), ce qui donne une courbe bien plus proche du flux spectral (*cf. fig.* 4.28 gauche), mais moins pratique pour détecter les débuts (*onset*) et fins (*offset*) de notes.

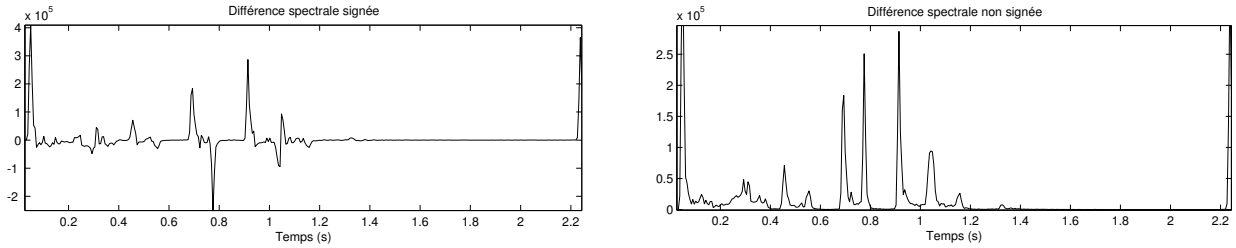


FIG. 4.30 – *Différence spectrale de la Piste n°3 ♫♪, signée (à gauche) et non signée (à droite).*

Pente spectrale $S_{tilt}(x, k)$

La pente spectrale (ou *spectral tilt*) est la pente de la régression linéaire des points utilisés pour représenter la forme spectrale de la composante sinusoïdale[Serra, 1996; Amatriain *et al.*, 2001] :

$$S_{tilt}(x, k) = \frac{1}{\sum_{i=1}^{H} t_i^2(k)} \sum_{i=0}^{H} \frac{t_i a_i(k)}{\sigma_i(k)} \tag{4.35}$$

avec

$$t_i(k) = \frac{1}{\sigma_i(k)} \left(f_i(k) - \frac{\sum_{i=0}^{H} \frac{f_i(k)}{\sigma_i^2(k)}}{\sum_{i=0}^{H} \frac{1}{\sigma_i^2(k)}} \right) \tag{4.36}$$

Cette pente spectrale est corrélée avec l'amplitude du signal [Bennett and Rodet, 1991]. On peut aussi la calculer sur la TFCT (*cf. fig.* 4.31).

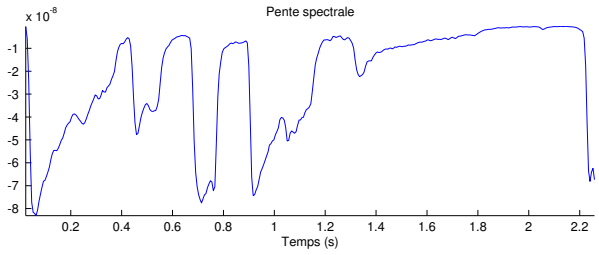


FIG. 4.31 – *Pente spectrale de la Piste n°3* ♫♪.

4.6 Descripteurs de haut niveau

Les descripteurs de haut niveau sont de plusieurs types : des descripteurs psychoacoustiques, des descripteurs de signal de haut niveau, des méta-descripteurs. Les descripteurs psychoacoustiques donnent un accès aux dimensions perceptives du son. On s'est notamment appuyés sur [Meunier and Canévet, 2002]. Les descripteurs de signal se basent sur une analyse des propriétés fines du son, notamment ses propriétés harmoniques. Une fois le signal $x(n)$ séparé en trois composantes (sinus + transitoires + résidu, [Verma *et al.*, 1997]), $S(n)$ la partie sinusoïdale, $T(n)$ la partie transitoire et $B(n)$ la partie bruitée, on peut obtenir de nouveaux indices. Les méta-descripteurs sont par exemple l'enveloppe spectrale, les coefficients Mel-cepstraux, etc. : des descripteurs multidimensionnels.

4.6.1 Descripteurs de signal de haut niveau

Distorsion harmonique

La distorsion harmonique est une mesure du degré de déviation des fréquences des partiels par rapport aux fréquences parfaitement harmoniques [Serra, 1996; Amatriain *et al.*, 2001] :

$$dist = \sum_{i=1}^{H} |f_i - iF_0| \frac{a_i}{\sum_{i=1}^{H} a_i} \tag{4.37}$$

Pour des sons de piano par exemple, la distorsion harmonique varie en fonction de la fréquence fondamentale et des partiels [Fletcher and Rossing, 1998]. Elles est dûe à la légère inharmonicité inhérente à la corde, dûe à des modes de propagation des ondes qui ne sont pas parfaitement harmoniques.

Caractère bruité

Le caractère bruité (ou *noisiness*) est une mesure de la quantité d'information non sinusoïdale présente dans la trame. On la calcule en faisant le rapport de l'amplitude du résidu sur l'amplitude totale du son [Serra, 1996; Amatriain *et al.*, 2001] :

$$Noisiness = \frac{\sum_{n=0}^{M-1} |x_R(n)|}{\sum_{n=0}^{M-1} |x(n)|} \tag{4.38}$$

Irrégularité du spectre

Lorsqu'on dispose d'une représentation additive d'un signal, l'irrégularité du spectre est définie [Krumhansl, 1989; Orio, 1999] à partir de l'équation :

$$IRR(x) = \log |\sum_{k=2}^{H-1} 20 \log \frac{A_k}{\sqrt[3]{A_{k-1}A_kA_{k+1}}}| \tag{4.39}$$

Lorsqu'on dispose d'une représentation transformée de Fourier à Court-Terme d'un signal, l'irrégularité du spectre est définie à partir de l'équation :

$$c_g(x) = \log |\sum_{k=1}^{N/2-1} 20 \log \frac{X(k)}{\sqrt[3]{X(k-1)X(k)X(k+1)}}| \tag{4.40}$$

4.6.2 Méta-descripteurs de haut-niveau

Coefficients Mel-cepstraux MFCC_i

Les coefficients cepstraux en échelle Mel correspondent à une représentation de l'analyse cepstrale en échelle *mel* plutôt qu'en *Hertz* (Hz). L'échelle *mel* approxime les fréquences telles qu'elles sont perçues, c'est-à-dire linéairement en dessous de $1kHz$, puis logarithmique au dessus. Le signal est alors filtré par un banc de filtres triangulaires espacées également en échelle *mel*, et l'on obtient une série de log-énergies E_k, un coefficient étant obtenu en sortie de chaque filtre et donnant la valeur de l'énergie du signal présent dans le filtre. L'ensemble de log-énergies est alors transformé selon la formule :

$$\text{MFCC}_i = \sum_{k=1}^{T} E_k \cos\left[i\left(k-\frac{1}{2}\right)\frac{\pi}{2}\right] \tag{4.41}$$

où T est le nombre de filtres triangulaires. On peut par exemple espacer les filtres de 150 mel centrés sur les fréquences de premiers harmoniques sous $1kHz$: on obtient alors environ 30 coefficients. Un exemple avec 12 filtres est donné *fig.* 4.32.

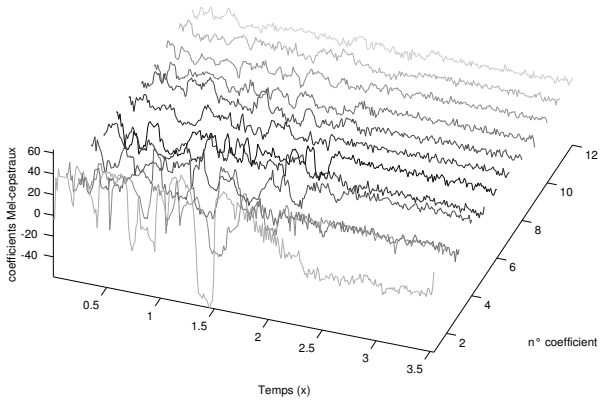


FIG. 4.32 – *Coefficients Mel-cepstraux, de la Piste n°16* ♫♪.

Formants

Dans le cas de la voix humaine ou des instruments de musique, les formants sont définis comme les résonances d'un modèle source-filtre. Chaque formant est défini par une amplitude A_f, une fréquence centrale F_f et une largeur de bande L_f.

Vibrato

L'analyse du vibrato se fait par analyse des modulations de la fréquence fondamentale [Arfib and Delprat, 1998; Rossignol *et al.*, 1998a; Desain and Honing, 1996]. Elle effectue l'extraction de la phase à l'origine, de la fréquence et de l'amplitude du vibrato au cours du temps, ainsi que d'un indicateur de vibrato.

4.6.3 Descripteurs perceptifs

4.6.2.i) Descripteurs de l'intensité sonore perçue : sonie, maximum de sonie, isosonie

Sonie spécifique N' La sonie spécifique, notée N', correspond à la sonie attribuable à un filtre auditif. La sonie spécifique s'étend des basses fréquences (centrées autour de 50 Hz pour les plus basses) aux hautes fréquences (centrées autour de 15 kHz). Cette échelle fréquentielle psychoacoustique explicite la distribution de l'énergie dans la cochlée. L'unité utilisée est le Erb, variant de 2 à 39. La sonie spécifique est calculée tous les 0.25 Erb. Nous l'utilisons comme méta-descripteur, puisqu'il s'agit d'une surface une sorte de sonagramme perceptif).

Sonie N La sonie, notée N (*cf. fig.* 4.33), se calcule à partir de l'intégrale de la sonie spécifique N' sur l'échelle des Erb (z) :

$$N = \int N'(z)dz \tag{4.42}$$

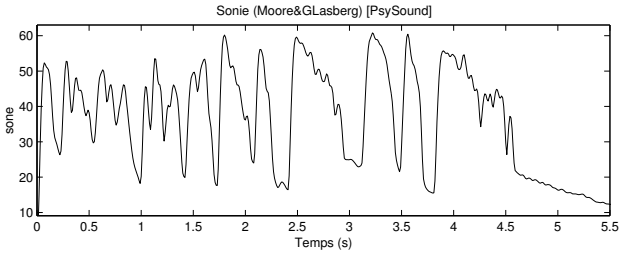


FIG. 4.33 – *Sonie N obtenu de la Piste n°20* ♫♪.

La sonie correspond donc à l'aire sous la courbe de sonie spécifique. Différents modèles de calcul de la sonie spécifique existent, dont celui de Moore et Glasberg [Moore and Glasberg, 1996], mais celui-ci ne convient pas aux sons non stationnaires. Le programme *PsySound* utilise l'algorithme de Zwicker et Fastl [Zwicker and Fastl, 1999], qui lui effectue un masquage temporel pour prendre en compte les évolutions de sons non stationnaires.

Maximum de sonie N_{max} Le maximum de sonie, noté N_{max} (*cf. fig.* 4.34), correspond au maximum de la sonie sur une fenêtre temporelle glissante. Pour un son de sonie variant dans le temps, la sonie est largement déterminée par les grandes valeurs, plutôt que par les faibles valeurs ; N_{max} est donc souvent une meilleure représentation de l'intensité perçue pour un son variable [Cabrera, 2000].

Isosonie L_N Pour une sonie modérée, le niveau de sonie s'approche par la fonction :

$$L_N = 40 + 10\log_2(N) \tag{4.43}$$

Le programme *PsySound* utilise une fonction plus complexe [Cabrera, 2000], non explicitée par son auteur. Un exemple est donné (*fig.* 4.35).

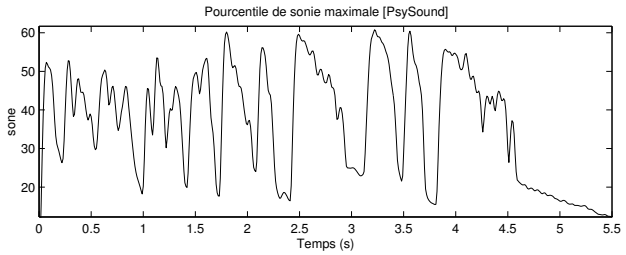


FIG. 4.34 – *Sonie totale N_{max} obtenu de la Piste n°20 ♫♪.*

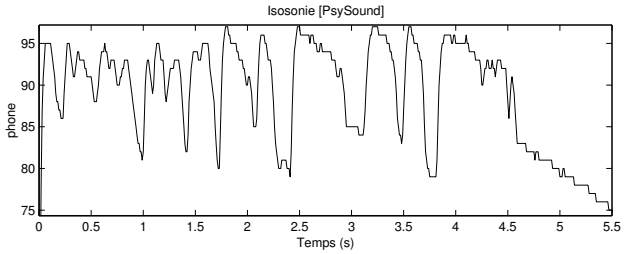


FIG. 4.35 – *Isosonie ou niveau de sonie obtenu de la Piste n°20 ♫♪.*

Comparaisons des courbes pour des niveaux différents La sonie se calcule par rapport à un signal de référence étalonné, dont on connaît la sonie. On peut alors calculer la sonie par rapport à ce signal de référence avec différents étalonnages. Ceci n'aurait pas de sens pour une expérience de psychoacoustique, puisque le signal de référence perdrait sa qualité de signal étalon. Par contre, dans notre cas, cela permet de connaître la sonie du son avec différents signaux de référence. La sonie (et donc l'isosonie) varie en fonction du niveau du signal de référence (exemple de l'isosonie *fig.* 4.36) : ces courbes d'isosonie ont été normalisées afin de pouvoir apprécier l'aspect non linéaire de l'isosonie en fonction du niveau de référence.

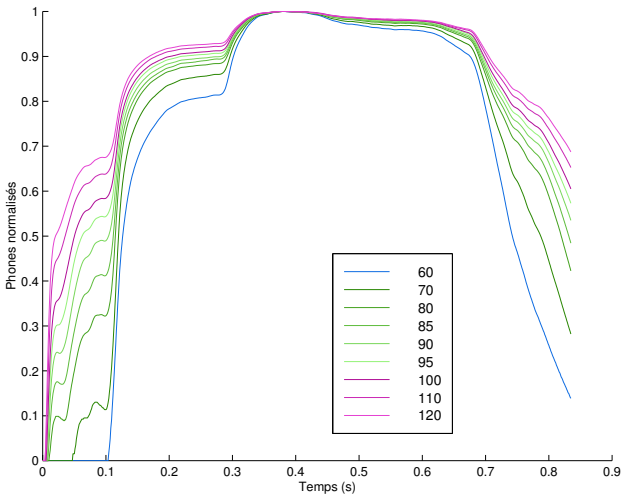


FIG. 4.36 – *Isosonie (*phone*) pour différentes valeurs du signal de référence de la Piste n°12 ♫♪.*

Le fait que les courbes soient différentes est intéressant pour le contrôle d'effets, aussi nous choisissons de calculer la sonie à plusieurs niveaux de référence dans le programme d'effets adaptatifs en *Matlab*. Les calculs n'ont pas été réalisé sous *PsySound*, car ce programme calcule en même temps tellement d'autres descripteurs qu'il est très lent. A la place, nous avons utilisé une implémentation sous *Matlab* réalisé par Alain Marchioni de l'équipe "Acoustique Perceptive et Informatique Musi-

cale" du LMA.

Mesures liées à la sonie

Dans le programme *PsySound*, plusieurs autres paramètres liés à la sonie sont calculés, à partir de la sonie spécifique : l'acuité (mesure de la brillance) ou ***sharpness***, la largeur timbrale, le volume (voir les descripteurs du timbre, *cf.* sec. 4.6.3). Des mesures de sonie statistiques sont ensuite calculées, que l'on peut manipuler comme on manipule les statistiques des mesures de niveau de pressions, afin d'obtenir le contraste de la sonie : on utilise alors plutôt les ratios que les différences. Zwicker et Fastl on montré qu'en utilisant les pourcentiles N_5 ou N_{10}, on obtient une bien meilleure estimation de la sonie globale d'un son, les moments forts ayant une plus grande salience que les moments calmes. Ces valeurs ne sont significatifs que pour la totalité d'un son, aussi il convient de segmenter le son d'abord, puis de réaliser une analyse avec *PsySound* pour chaque segment. La mesure statistique L_A caractérise la distribution des valeurs de la sonie au cours du temps. Il faut au moins 20 valeurs de sonie pour que *PsySound* calcule ce descripteur.

4.6.2.ii) Hauteur perçue

Le pitch ou hauteur perçue correspond à la hauteur que l'être humain entend à l'écoute du son, et se décompose en hauteur tonale et hauteur spectrale. On calcule la hauteur tonale, qui est intimement lié à la composition fréquentielle du son, et évolue en fonction des fréquences et des amplitudes des sinusoïdes, dans le cas des sons quasi-harmoniques ou harmoniques. Le modèle de Terhardt est basé sur l'analyse du spectre et non sur des méthodes d'autocorrélation.

Descripteurs liés à la hauteur

Le modèle de Terhardt mesure la hauteur perçue, sa proéminence, la hauteur virtuel et les décalages de hauteur (***pitch-shifts***). De plus, Parncutt [Parncutt, 1989] (cité par [Cabrera, 2000]) propose des estimations de tonicité (comment le son est tonal) de deux types, et de multiplicité (combien de hauteurs distinctes sont entendues), qui sont intégrées à *PsySound* en ne se limitant pas aux douze demi-tons de la gamme tempérée comme dans le modèle de Parncutt. Ces résultats sont ensuite quantifiés pour correspondre à la gamme tempérée (les hauteurs à mi-chemin entre deux demi-tons sont partagées par les deux demi-tons). Les motifs de salience sont exprimés linéairement selon la hauteur, et circulairement selon l'échelle des chromas.

Caractère tonal simple et complexe (*tonalness*) Le caractère tonal pur indique l'audibilité des hauteurs spectrales (*cf. fig.* 4.37 gauche). La caractère tonal complexe indique l'audibilité de la hauteur virtuelle (fréquence fondamentale), et doit s'interpréter comme une mesure de la similarité du spectre avec celui d'un son harmonique, ou complexe (*cf. fig.* 4.37 droite). Ces descripteurs sont calculés d'après [Parncutt, 1989] (cité par [Cabrera, 2000]).

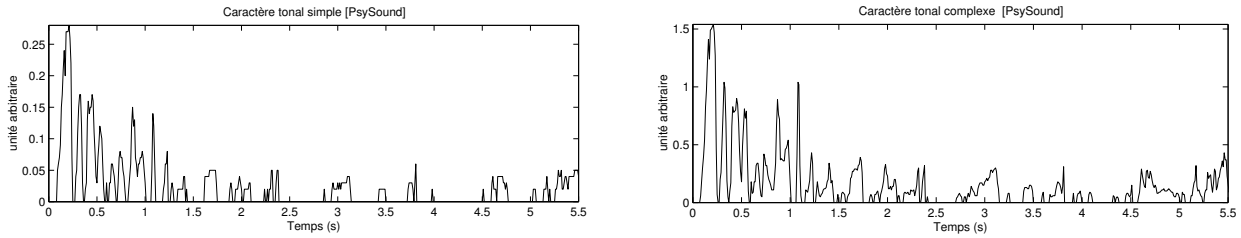


FIG. 4.37 – *Caractère tonal pur (à gauche) de la Piste n°20 ♫♪ et complexe (à droite).*

Multiplicité La multiplicité de hauteur est une estimation du nombre de tons simultanément présents dans le son [Parncutt, 1989] (cité par [Cabrera, 2000]), mise en œuvre dans *PsySound*.

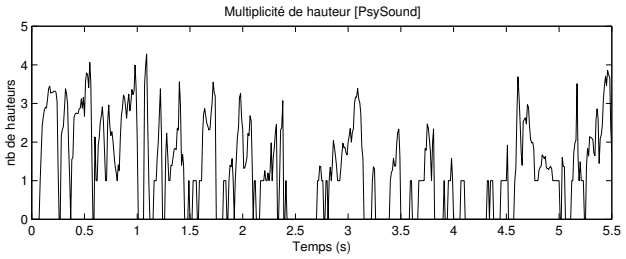


FIG. 4.38 – *Multiplicité de hauteurs obtenu de la Piste n°20* ♫♪.

Proéminence chromatique Les chromas sont les noms des douze demi-tons qui composent la gamme chromatique occidentale. Alors que des variations brusques correspondent probablement à des accidents de calcul de *PsySound*, la plupart du temps, les valeurs à plus faible variations sont plus fiables. La proéminence chromatique, ou *chroma salience* (*cf.* 4.39) est calculée de la manière suivante [Parncutt, 1989] (cité par [Cabrera, 2000]) à l'aide de l'algorithme de calcul de hauteur de [Terhardt, 1979] : on considère qu'une hauteur parfaitement accordée sur une gamme tempérée sera toujours entendue selon le chroma correspondant, et uniquement celui-là. Une hauteur à cheval entre deux chromas contribuera proportionnellement aux deux chromas (hypothèse simpliste).

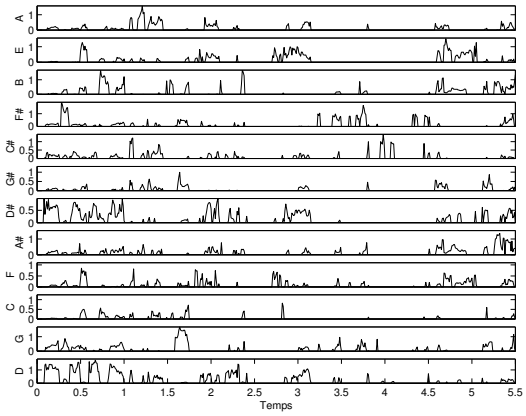


FIG. 4.39 – *Proéminence chromatique obtenue de la Piste n°20* ♫♪.

Proéminence de hauteur tonale *PsySound* extrait aussi la proéminence de chaque hauteur, pour toute la tessiture des instruments occidentaux, par octave (*cf. fig.* 4.40 gauche pour le deuxième octave et *cf. fig.* 4.40 droite pour le troisième octave).

Tonique – tonalité Le programme *PsySound* calcule le coefficient de corrélation du motif d'acuité chromatique pour chacune des tonalités, et celle de plus grande valeur (la plus corrélée) est la tonique supposée. Le coefficient de corrélation R^2 est donné *fig.* 4.41.
La tonalité quant à elle est donnée par un entier dans l'ordre du cycle des quintes (*cf.* 4.2), et doit être lu circulairement (ie. $mod(\cdot, 12)$) :

Majeur / mineur Le caractère majeur ou mineur de la tonalité est donné par un entier dans *PsySound* : 0 pour l'accord Majeur et -3 pour l'accord mineur. Ceci permet d'interpréter la tonalité

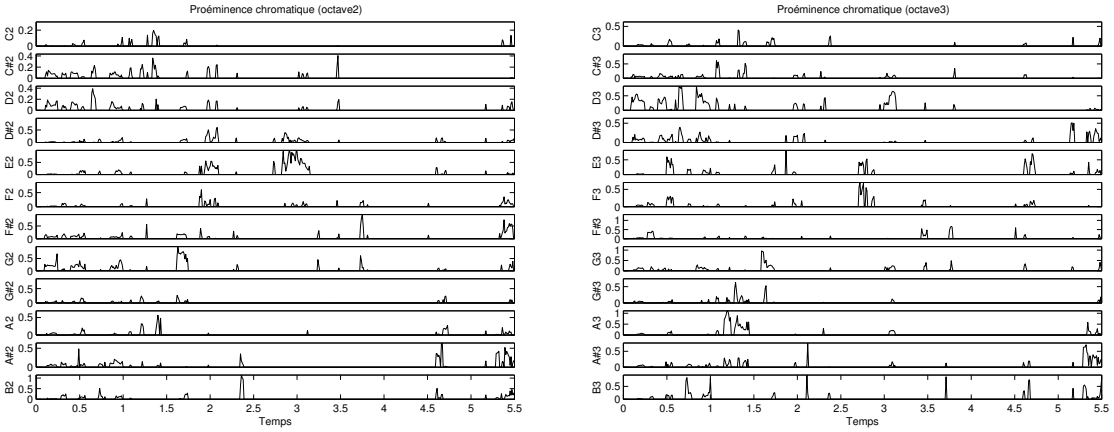


FIG. 4.40 – *Proéminence de hauteur tonale de l'octave 2 (à gauche) et de l'octave 3 (à droite), obtenus de la Piste n°20 ♫♪.*

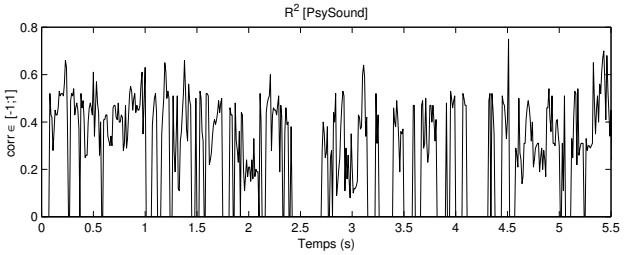


FIG. 4.41 – *Coefficient de corrélation du motif d'acuité chromatique R^2 de la Piste n°20 ♫♪.*

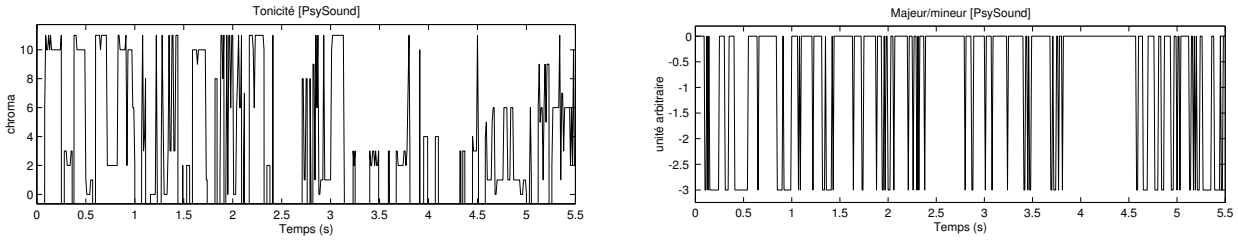


FIG. 4.42 – *Tonicité obtenu de la Piste n°20 ♫♪ (à gauche) et indicateur majeur / mineur : M=0/ $m = -3$ (à droite).*

en terme de gamme majeure ou gamme relative mineure (afin d'interpréter la nomenclature) ; la tonique est alors :

$$Tonique_{rel} = Tonique + Ind_{M/m} \tag{4.44}$$

4.6.2.iii) Descripteurs de la spatialisation

Fonction d'intercorrélation entre les canaux (ICCC) Pour des signaux stéréophoniques, on peut calculer la fonction d'intercorrélation entre les canaux sur une TFCT de 4096 points, avec des décalages allant de -45 à 45 échantillons. La plage de variation du décalage est de l'ordre de $\pm$ 1 ms, ce qui correspond à la plage de variation des délais interauraux qui impliquent la reconnaissance de l'azimut d'une source. La valeur absolue de la fonction d'intercorrélation s'obtient selon la formule :

$$ICCC = \max_{\tau \in [-45;45]} \left| \frac{\sum_{n=45}^{4050} x_g(n) x_d(n+\tau)}{\sqrt{\sum_{n=45}^{4050} x_g^2(n) \sum_{n=45}^{4050} x_d^2(n)}} \right| \tag{4.45}$$

La	**Mi**	**Si**	**Fa ♯**	**Do ♯**	**Sol ♯**	**Ré ♯**	**La ♯**	**Fa**	**Do**	**Sol**	**Ré**
(A)	(E)	(B)	(F♯)	(C♯)	(G♯)	(D♯)	(A♯)	(F)	(C)	(G)	(D)
0	1	2	3	4	5	6	7	8	9	10	11

TAB. 4.2 – *Tonalité ordonnée selon le cycle des quintes, et le numéro associé dans* PsySound

Cette fonction permet de mesurer la similarité des signaux provenant des deux canaux. Dans certaines circonstances, sa valeur absolue correspond à la notion d'espace (*auditory spaciousness*), qui se réfère à la taille apparente ou à l'étalement du son dans l'espace, et peut se relier au volume auditif.

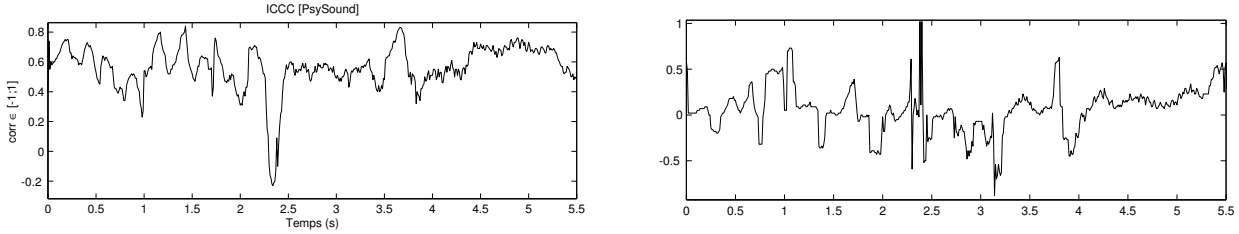


FIG. 4.43 – *A gauche : corrélation croisée inter-canaux ICCC de la Piste n°20 ♫♪ stéréophonique. A droite : temps de meilleure corrélation croisée inter-canaux* τ_{ICCC}.

Temps de meilleure inter-corrélation τ_{ICCC} Le temps de meilleur intercorrélation (*ICCC lag*) correspond au temps de décalage associé à la valeur de la fonction ICCC ; c'est la valeur $\frac{\tau_0}{F_e}$ telle que τ_0 maximise la fonction ICCC. Dans certaines circonstances, ce temps donne une indication de la latéralisation du son. Cependant, il faut lire sa valeur en même temps que celle de l'ICCC et de la différence de pression. Afin de donner des résultats sensés pour des fonctions d'intercorrélation périodiques, une pondération centrale est effectuée pour la recherche du maximum.

Différence de niveau entre canaux ΔL La différence de niveaux intercanaux ΔL est simplement la différence en dB de niveau entre les deux canaux. Le calcul fait dans *PsySound* n'effectue aucune pondération temporelle ou fréquentielle.

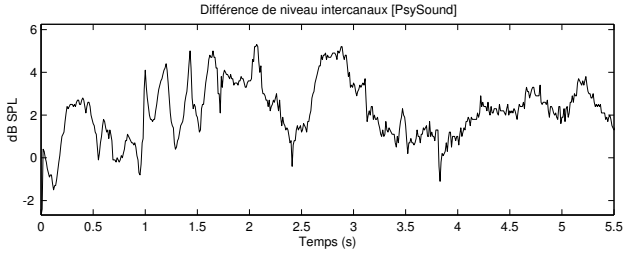


FIG. 4.44 – *Différence de niveau intercanaux de la Piste n°20 ♫♪ stéréophonique.*

Statistique de la différence de niveau Les mesures statistiques sur la différence de niveau sont de même nature que celles sur les mesures SPL, mis à part le fait (important) qu'elles représentent les différences de valeur entre les deux canaux. $\Delta_{10} - \Delta_{90}$ ou $\Delta_{20} - \Delta_{80}$ peuvent donner une mesure du contraste spatial existant entre les deux canaux. Δ_{50} devrait donner une mesure de la latéralisation globale meilleure que la valeur moyenne ΔL.

4.6.2.iv) Descripteurs du timbre

Acuité (brillance) Ac_A **et** $Ac_{Z\&F}$ L'acuité (*sharpness*) est une mesure subjective de la brillance, et correspond à une sorte de mesure de la hauteur du timbre entier (grave/aigu). Son pendant dans le vocabulaire musical est la brillance. Elle est souvent considérée comme corrélée au CGS [McAdams *et al.*, 1995; Krimphoff, 1994]. Le programme *PsySound* calcule l'acuité selon deux modèles :

– modèle d'Aures [von Aures, 1985] (cité par [Cabrera, 2000]) :

$$Ac_{Aures} = 0.585 \frac{\int_{z=0}^{24\ \text{Bark}} N'(z) g(z) dz}{\log\left(\frac{N+20}{20}\right)} \tag{4.46}$$

– modèle de Zwicker et Fastl [Zwicker and Fastl, 1999] :

$$Ac_{Z\&F} = 0.11 \frac{\int_{z=0}^{24\ \text{Bark}} N'(z) z g(z) dz}{N} \tag{4.47}$$

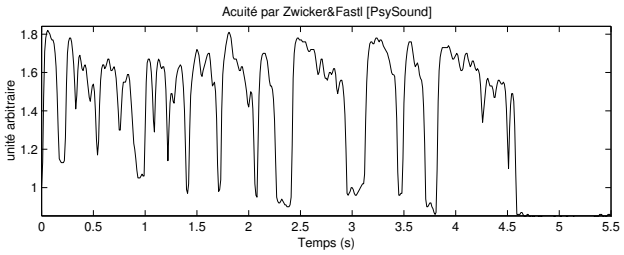


FIG. 4.45 – *Acuité par le modèle de Zwicker et Fastl de la Piste n°20* ♫♪.

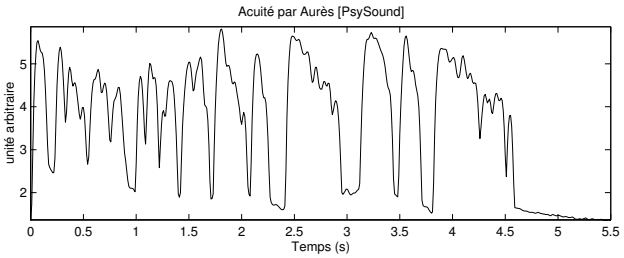


FIG. 4.46 – *Acuité par le modèle de Sétharès de la Piste n°20* ♫♪.

avec z le taux de bande critique et $g(z)$ une fonction de pondération qui accentue les hautes fréquences [Zwicker and Fastl, 1999]. L'unité de l'acuité est l'*acum*. Un *acum* est défini comme l'acuité d'une bande de bruit centrée sur 1 kHz, large d'une bande critique, de niveau SPL 60 dB. Les deux modèles fonctionnent avec l'échelle des $Bark$ au lieu des Erb pour mesurer le taux des bandes critiques. Le modèle de Zwicker et Fastl calcule le centre de gravité pondéré de la sonie spécifique tandis que celui d'Aurès est plus sensible à l'influence de la sonie sur l'acuité. Les fonctions $g(z)$ sont différentes pour les deux modèles.

Spectre compact Le spectre compact est une représentation de la TFCT plus proche de la perception : jusqu'à 441 Hz, les bandes d'analyse sont de largeur constante, puis à partir de 458.5 Hz, la distribution des bandes d'analyse est logarithmique, avec des composants espacés d'un demi-ton (un douzième d'octave). Leurs fréquences sont choisies pour correspondre exactement aux octaves et tiers d'octaves usuels. A partir de ce spectre compact, on peut calculer le spectre par octave, le spectre par tiers d'octaves. C'est ce que fait *PsySound*.

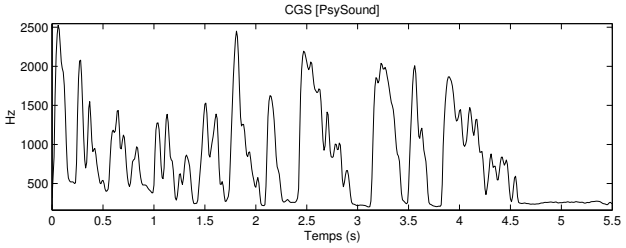


FIG. 4.47 – *Centre de gravité spectrale du spectre compact de la Piste n°20* ♫♪.

Largeur timbrale On appelle largeur timbrale la largeur du pic du spectre de sonie spécifique. C'est une mesure de la platitude (*flatness*) de la sonie spécifique : en effet, plus le spectre est large, plus il est plat (à l'extrême : bruit blanc), et plus le spectre est étroit, moins il est plat (à l'extrême : un son pur). Sa mesure est proposée par Malloch [Malloch, 1997] (cité par [Cabrera, 2000]) à partir de la méthode d'analyse du timbre par tristimulus de Pollard et Jansson [Pollard and Jansson, 1982] (cité par [Cabrera, 2000]), avec quelques modifications dues à la différence des méthodes employées pour calculer les composantes fréquentielles de la sonie spécifique.

Le programme *PsySound* la calcule d'une manière similaire à [Malloch, 1997] (cité par [Cabrera, 2000]) selon :

$$L_{timb} = \left(\frac{\left(N - \int_{z_{max}-0.5}^{z_{max}+0.5} N'(z)dz \right)}{N} \right)^2 \tag{4.48}$$

avec z_{max} le numéro de Erb auquel le spectre de sonie spécifique atteint son maximum. Les valeurs théoriques varient entre 0 et 1 : un ton pur aura une largeur timbrale inférieure à celle d'un bruit à bande étroite. La fonction carrée est utilisée pour diminuer les valeurs, qui sinon seraient toutes très proches de 1.

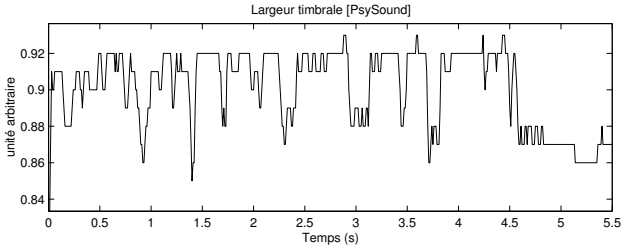


FIG. 4.48 – *Largeur timbrale de la Piste n°20* ♫♪.

Volume Le volume concerne la taille perçue du son, mesure subjective du son allant de "petit" à "grand". Le volume et la notion d'espace prise par le son ont des définitions très similaires : la taille apparente du son. C'est un concept assez ancien : le volume auditif de sons purs était le sujet de recherche de Stevens en 1933. Cabrera [Cabrera, 2000] suggère qu'une composante binaurale intervient dans le calcul du volume. Il a mis en œuvre son propre modèle de volume dans *PsySound*, dont la formule analytique est donné par :

$$V = 3.47\ 10^5 \frac{\sqrt{N}}{\left(\frac{\int N'(z)(z+8.65)dz}{N} \right)^4} \tag{4.49}$$

en vol, avec z le numéro d'Erb. 1 vol est le volume d'un son pur à 1 kHz en champ libre à un niveau SPL de 40 dB. L'idée générale est que le volume augmente avec la sonie, et diminue avec

le centroïde. Ceci explique la fraction au dénominateur. La puissance 4 quant à elle détermine l'influence relative de la sonie et du centroïde. Cette approche est liée à la découverte du fait que le volume de bruits à bande étroite est égale à leur sonie divisée par leur brillance (ou densité). Notons que cette fonction n'est fiable que pour un nombre limité de stimuli, lesquels ne sont pas précisés dans le documentation du logiciel *PsySound*.

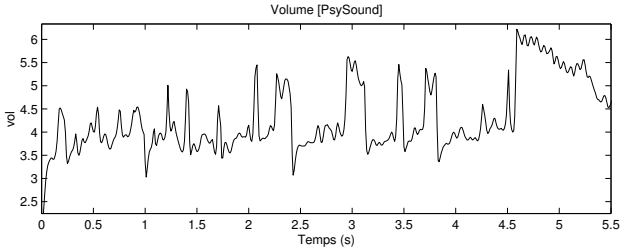


FIG. 4.49 – *Volume de la Piste n°20* ♫♪.

Dissonance La dissonance musicale est déterminée par une combinaison de facteurs acoustiques et contextuels. Les facteurs contextuels concernent le langage musical. On ne cherche pas à les modéliser ici. La composante acoustique quant à elle est encore appelée **rugosité**, et se calcule assez simplement. Deux méthodes de calcul sont utilisées dans *PsySound* : à partir des composants du "spectre compact" (mesure du caractère bruité) et à partir des composants tonaux (extraits lors de l'analyse de la hauteur par l'algorithme de Terhardt ; mesure de la dissonance musicale). Ces deux méthodes de calcul sont mises en œuvre avec le modèle de Hutchkinson et Knopoff [Hutchinson and Knopoff, 1978] (cité par [Cabrera, 2000]) qui normalise les résultats et utilise une amplitude linéaire, puis avec le modèle de Sétharès [Sethares, 1993] (cité par [Cabrera, 2000]) qui ne normalise pas les résultats et utilise une échelle en dB. D'autres méthodes existent [Daniel and Weber, 1997; Pressnitzer and McAdams, 1999; Leman, 2000], mais nous n'avons pas eu le temps de les mettre en œuvre.

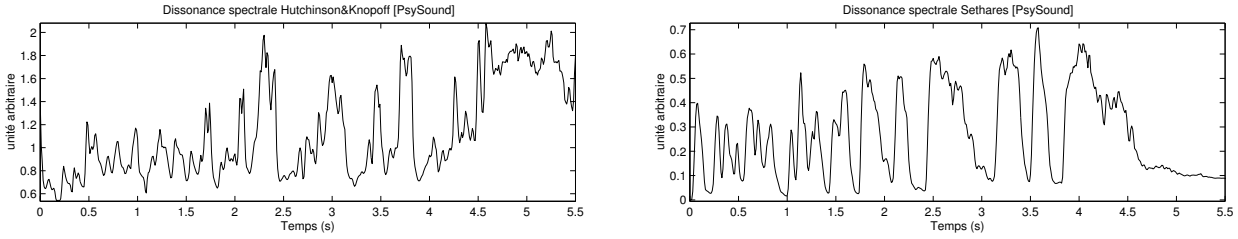


FIG. 4.50 – *Dissonance spectrale par le modèle de Hutchkinson et Knopoff (à gauche) et par le modèle de Sétharès (à droite) de la Piste n°20* ♫♪.

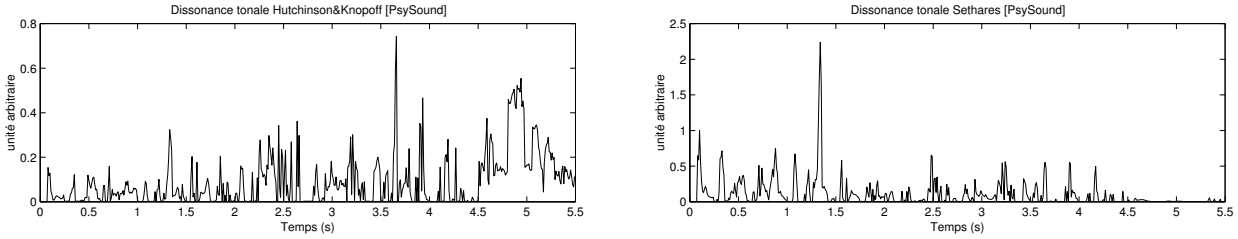


FIG. 4.51 – *Dissonance tonale par le modèle de Hutchkinson et Knopoff (à gauche) et par le modèle de Sétharès (à droite) de la Piste n°20* ♫♪.

4.6.4 Descripteurs musicaux

Il serait intéressant d'utiliser d'autres descripteurs décrivant la musique tels que le rythme, le style de musique, le style de voix, la liste des instruments, les effets appliqués au son. Toutefois, ceci n'a pas été intégré à notre étude, car encore trop complexe à extraire du signal.

4.7 Paramètres dérivés des descripteurs de bas et de haut-niveau

Les paramètres que nous allons présenter maintenant sont calculés à partir des autres descripteurs, à l'instant du taux de basse énergies, par exemple. Il s'agit de paramètres statistiques, de la dérivée et de l'intégrale. On peut les considérer comme une introduction au *mapping*, du fait que les fonctions de dérivation et d'intégration sont aussi utilisées pour transformer les descripteurs, *cf.* sec. 6.3.2. De même, la moyenne sur fenêtre glissante est utilisée pour le zoom *cf.* sec. 6.4.2.

4.7.1 Paramètres statistiques : moyenne, variance, coefficients d'asymétrie et d'aplatissement et segmentation

Il est intéressant d'avoir des informations sur le type de variation des descripteurs. Ces variations servent pour des modèles de timbre, pour la classification, et peuvent être pris comme descripteurs à leur tour, pour contrôler des effets. On considère donc la moyenne, la variance, les moments d'ordre supérieurs (coefficients d'asymétrie et d'aplatissement), ainsi que les dérivées et intégrales.

Moyenne

La moyenne d'un descripteur est le moment statistique d'ordre 1. Elle peut se calculer selon deux approches : si l'on dispose d'une segmentation du signal, on calcule la moyenne sur chaque segment, même si leurs tailles sont différentes, tandis que lorsqu'on ne dispose pas de cette segmentation, on calcule la moyenne sur une fenêtre glissante. Les moyennes mesurées sont alors différentes : dans le cas de la segmentation, le descripteur de moyenne aura une valeur constante sur tout le segment, alors que dans l'autre cas, la valeur sera échantillonnée régulièrement (*cf. fig.* 4.52).

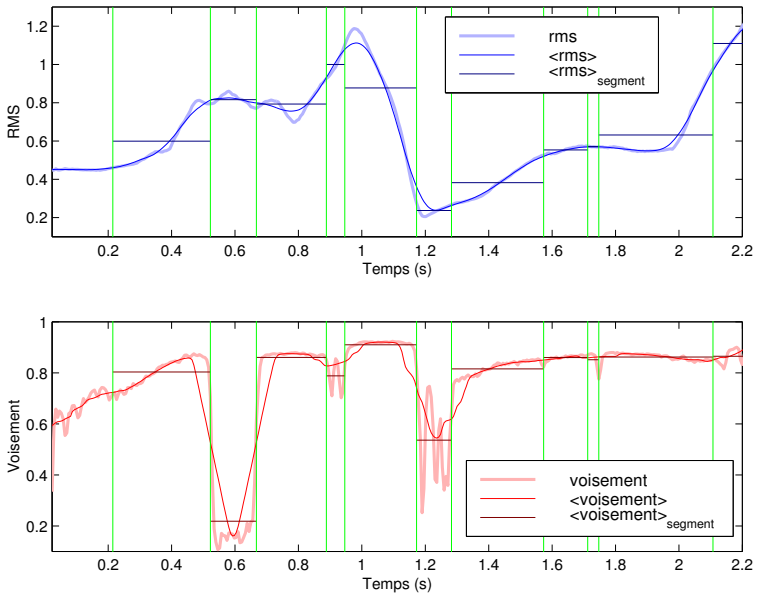


FIG. 4.52 – *Moyenne glissante et moyenne par segments de deux descripteurs : le RMS et l'indice de voisement.*

Variance

La variance correspond à l'écart moyen à la valeur moyenne. C'est le moment statistique d'ordre 2. De grandes valeur signifient que le paramètre analysé varie beaucoup autour de sa moyenne. lorsque la variance est de l'ordre de la moyenne, les résultats statistiques sont considérés comme peu fiables. Une fois encore, on peut la calculer sur des segments ou sur une fenêtre fixe. Notons cependant que la valeur calculée n'est significative que pour un nombre de valeurs supérieur à 30. Dans le cas contraire, elle n'est pas statistiquement significative. Dans notre cas, on ne désire pas faire de test statistique à partir de nos données, donc peu nous importe que notre descripteur ne soit pas tout à fait valide.

Coefficient d'asymétrie

Le coefficient d'asymétrie (ou *skewness*) est le moment statistique d'ordre 3. C'est une mesure du degré d'asymétrie de la distribution de probabilité étudiée autour de sa moyenne. Si la distribution est symétrique, le coefficient d'asymétrie vaudra 0, mais ce n'est pas réciproque. En effet, une valeur nulle ne prouve pas que la distribution est symétrique. Par contre, une valeur non nulle peut indiquer que la distribution n'est pas symétrique. Ainsi, une valeur négative indique que la distribution est plus étalée en dessous de la moyenne qu'au dessus, et une valeur positive indique que la distribution est plus étalée au dessus de sa moyenne qu'en dessous. Une fois encore, on peut calculer la valeur du coefficient d'asymétrie sur un segment du descripteur, ou sur une fenêtre glissante.

Coefficient d'aplatissement

Le coefficient d'aplatissement (ou *kurtosis*) est le moment statistique d'ordre 4. Il est une indication de la gaussianité d'une statistique ; en effet, c'est une mesure de l'aspect en pic ou plat de la distribution de probabilités. Un coefficient d'aplatissement positif indique une distribution en pic, et un coefficient d'aplatissement négatif indique une distribution plutôt plate. Une distribution normale a un coefficient d'aplatissement nul ou quasi-nul (on la dit méso-kurtique). Une valeur positive correspond à une distribution trop élevée (lepto-kurtique), alors qu'une valeur négative correspond à une distribution trop plate (plato-kurtique).

Dubnov utilise cette mesure avec le coefficient d'asymétrie, calculés à partir du bispectre, pour indiquer dans quelle mesure l'hypothèse de gaussianité des variations de fréquence des harmoniques est vérifiée [Dubnov and Tishby, 1996; Dubnov and Rodet, 1997; Dubnov and Tishby, 2002]. Ceci lui permet notamment de classifier les timbres des instruments en fonction de ces deux paramètres. Cependant, les analyse qu'il réalise se fait sur la portion stable de sons relativement longs (de l'ordre de deux secondes, afin que le nombre de fréquences des harmoniques soit statistiquement significatif. On utilisera donc ces mesures dans une optique plutôt expérimentale et musicale que statistique !

4.7.2 Dérivées, intégrale

Dérivées relatives et absolues d'un indice

Pour tous les indices dont nous parlerons par la suite, on peut calculer plusieurs dérivées. En effet, la dérivée première s'obtient par approximation numérique selon différents schémas, et il en est de même pour les dérivées successives. De plus, chaque dérivée peut s'exprimer sous une forme relative ou absolue. Soit $u(t)$ une fonction continue du temps t, que l'on connaît de manière discrétisée $u_n = u(t_n)$ en des instants $t_n = n*\Delta t$. La dérivée relative première $\frac{\partial u}{\partial t}(t)$ peut s'approcher par différences finies de trois manières :

- différences finies centrées : $d_C u(t) = \frac{u_{j+1} - u_{j-1}}{2\Delta t}$
- différences finies décentrées rétrograde (*backward*) : $d_B u(t) = \frac{u_j - u_{j-1}}{\Delta t}$

– différences finies décentrées postgrade (*forward*) : $d_F u(t) = \frac{u_{j+1}-u_j}{\Delta t}$

De la même façon, la dérivée absolue première $\frac{\frac{\partial u}{\partial t}(t_n)}{u(t_n}$ peut s'approcher par différences finies de trois manières :

– différences finies centrées : $\delta_C u(t) = \frac{u_{j+1}-u_{j-1}}{2u_j \Delta t}$

– différences finies décentrées rétrograde (*backward*) : $\delta_B u(t) = \frac{u_j-u_{j-1}}{u_j \Delta t}$

– différences finies décentrées postgrade (*forward*) : $\delta_F u(t) = \frac{u_{j+1}-u_j}{u_j \Delta t}$

La dérivée relative d'ordre deux $\frac{\partial^2 u}{\partial t^2}(t)$ s'approche le plus souvent par la forme centrée : $d_C^2 u(t) = \frac{u_{j+1}-2u_j+u_{j-1}}{(\Delta t)^2}$. La dérivée absolue d'ordre deux $\frac{\frac{\partial^2 u}{\partial t^2}(t)}{u_j}$ s'approche par la forme centrée : $\partial_C^2 u(t) = \frac{u_{j+1}-2u_j+u_{j-1}}{u_j(\Delta t)^2}$.

Intégrale d'un descripteur au cours du temps

Nous pouvons aussi calculer une approximation de l'intégrale d'un descripteur, par la formule des trapèzes :

$$I(u(t)) \approx \sum_{k=1}^{K} \frac{u(t_k) + u(t_{k+1})}{2} = \sum_{k=2}^{K-1} u(t_k) + \frac{u(t_1) + u(t_K)}{2} \tag{4.50}$$

C'est une fonctionnalité offerte dans le contrôle, via l'une des fonctions de conformation (*sec.* 6.3.2). Elle est utilisée par exemple pour la dilatation/contraction temporelle avec synchronisation, pour le contrôle de la conformation fréquentielle ou du trémolo spectral adaptatif.

4.8 Sous-échantillonnage, interpolation, qualité du calcul

4.8.1 Sous-échantillonnage

Pour chaque descripteur extrait du son, l'extraction (ou le calcul) du descripteur se fait à des instants donnés : la courbe descriptive est en fait sous-échantillonnée par rapport au signal sonore, *cf. fig.* 4.53. Si l'on n'applique pas de filtre passe-bas et que l'on calcule le RMS pour un échantillon sur 128, à partir d'un signal échantillonné une fréquence de $F_e = 44,1kHz$, la fréquence d'échantillonnage de ce signal de contrôle est alors $\frac{F_e}{128} = 344,53\ Hz$. Si le signal comporte des fréquences supérieures à la demi fréquence d'échantillonnage $172,27\ Hz$, elles seront repliées. Il conviendrait donc de filtrer ces paramètres au cours de l'extraction.

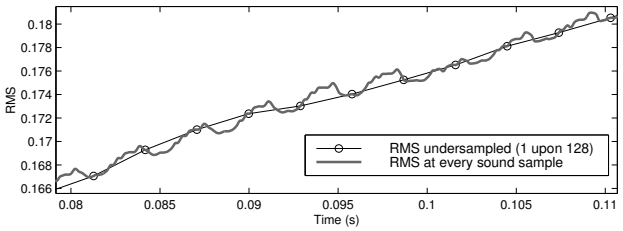


FIG. 4.53 – *Extraction du RMS par pas de 1 échantillon et par pas de 128 échantillons : le sous-échantillonnage implique un lissage et un repliement du spectre, du fait de l'absence de filtrage passe-bas avant décimation.*

4.8.2 Interpolation

Dans une analyse–synthèse par TFCT, tel que le vocodeur de phase, le pas d'analyse R_a et le pas de synthèse R_s ne sont pas forcément identiques, ni constants. Ceci signifie que les courbes

des descripteurs peuvent être interpolées (ré-échantillonnées) afin d'extraire la valeur de contrôle correspondant à l'instant du grain. Là encore, il faut bien choisir le type d'interpolation selon ce que l'on contrôle. Chaque méthode d'interpolation apporte son lot d'erreur : on préféra une interpolation linéaire pour sa simplicité d'implémentation (lorsque la précision du contrôle obtenu à partir du descripteur n'est pas nécessairement grande), et une interpolation par splines ou cubique lorsqu'une grande précision est requise.

4.8.3 Qualité du calcul

La valeur des descripteurs peuvent dépendre de paramètres tels que la taille de la fenêtre d'analyse, le fait qu'elle soit synchrone à la hauteur ou non, comme il a été illustré pour l'amplitude p. 117. Il est donc intéressant de calculer ces descripteurs avec plusieurs valeurs des paramètres de contrôle du calcul, pour proposer des courbe de contrôles différentes.

4.9 Redondances et corrélations des descripteurs sonores

Il a été montré qu'il existe des corrélations entre certains descripteurs. De plus, à la lumière de la présentation des descripteurs et de courbes d'exemples dans la partie précédente, nous avons remarqué qu'il existait des redondances. Nous allons développer ce point, en gardant à l'esprit que l'on utilise ces descripteurs pour contrôler un traitement sonore et non pour effectuer une analyse–synthèse parfaite, par exemple.

4.9.1 Corrélations entre descripteurs

Il existe des corrélations entre certains descripteurs. En effet, le jeu instrumental, la physique des instruments de musique et plus généralement des modes de productions de signal sonore impliquent des évolutions conjointes de descripteurs :

- entre fréquence fondamentale et forme spectrale : pour un même instrument, la forme spectrale n'est pas constante mais évolue en fonction de la fréquence fondamentale ;
- entre fréquence fondamentale et amplitude : selon la fréquence fondamentale, les amplitudes des harmoniques ne sont pas les mêmes, ainsi que l'amplitude de la partie résiduelle ;
- entre amplitude et forme spectrale : plus le son est joué fort (timbre de l'instrument), plus le spectre s'enrichit dans les hautes fréquences, ce qui modifie la forme spectrale (et donc le centroïde spectrale et la pente spectrale) [Beauchamp, 1982]. De plus, le rapport signal sur bruit varie en fonction de l'amplitude totale du son lorsqu'il est joué.

L'existence de ces corrélations signifie que notre ensemble de descripteurs est trop grand, et que l'information qu'ils donnent est redondante.

4.9.2 Redondances entre descripteurs

Il existe d'autres redondances entre descripteurs. Nous avons vu en sec. 4.5.2 que le centre de gravité spectrale possède plusieurs définitions, dont l'évolution des valeurs sont proches. Ces définitions donnent donc des descripteurs dont l'évolution, à défaut d'être identique, est souvent corrélée.
De plus, d'autres descripteurs tel le taux de passage par zéros (*cf.* sec. 4.5.1), le roulement spectral (*cf.* sec. 4.5.2) et le contenu en hautes fréquences (*cf.* sec. 4.5.2) sont corrélés au centroïde. La pente spectrale est corrélée à l'amplitude du signal.

4.9.3 Quelques réflexions à ce sujet

Tout d'abord, les redondances entre les paramètres s'expliquent de deux manières : certains descripteurs correspondent à deux manières de mesure une même quantité, et donnent donc des

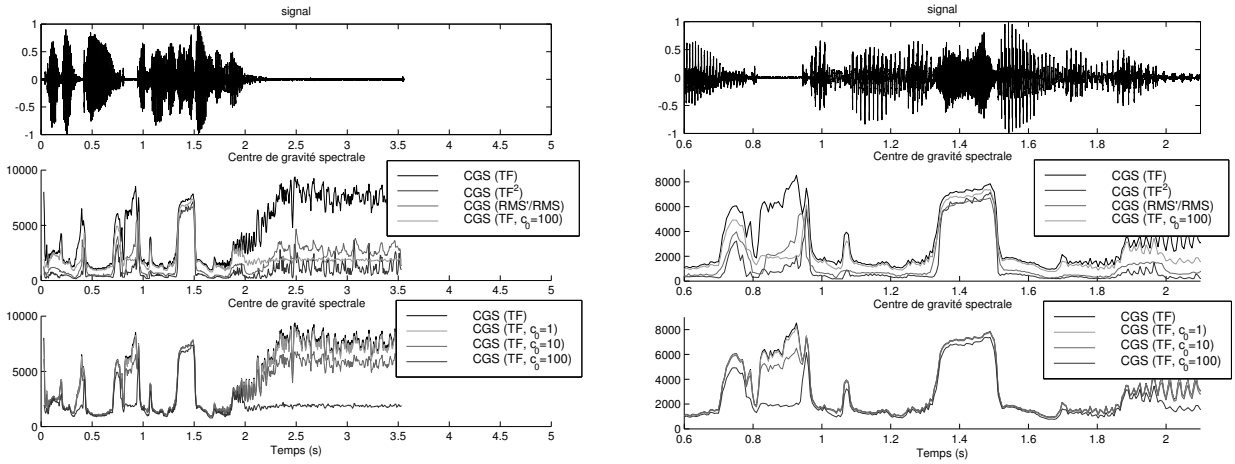


FIG. 4.54 – *A gauche : centre de gravité spectral (Piste n°18 ♫♪) calculé par plusieurs méthodes. A droite : idem, avec zoom. Les différentes méthodes estiment différemment le centroïde. On remarque que l'utilisation de la constante au dénominateur fait bien tendre le CGS vers 0 pour les portions non voisées.*

résultats proches. Ensuite les corrélations existant entre certains descripteurs, dues à la physique, impliquent une redondance dans les mesures que l'on en fait. Si l'on s'était placé dans le cadre d'une recherche de descripteurs satisfaisant une bonne analyse et resynthèse de signaux sonores, il nous aurait fallu absolument passer par une étape de réduction de données [Rochebois, 1997; Drame *et al.*, 1998], par exemple par analyse en composantes principales, ou en utilisant des méthodes non linéaires et implicites tels les réseaux de neurones. Cependant et comme nous l'avons déjà affirmé précédemment, pour contrôler les effets audionumériques adaptatifs, tout descripteur peut être bon. De plus, les légères différences entre deux courbes corrélées et redondantes peuvent donner des contrôles d'effet qui, une fois appliqués, donneront des sons dont la différence est intéressante. Il n'y a donc pas d'a priori pour savoir comment réduire les données ; nous choisissons donc de conserver tous les descripteurs afin d'offrir le plus de créativité possible lors de l'utilisation musicale des effets adaptatifs.

Deuxième partie

Effets audionumériques adaptatifs

Donner aux effets audionumériques un contrôle adaptatif ?

Ce qui fait l'homme, c'est sa grande faculté d'adaptation.
Socrate [Platon, 1997]

Ce qui importe à la simulation du geste instrumental, ce ne sont pas la rapidité, l'uniformité ni le rendement mais la virtuosité, la souplesse, l'amplitude, l'aisance, la justesse, la finesse, l'infinie coordination d'un mouvement d'adaptation finalisé.
Hugues Dufour [Dufour, 1999]

Généralisation des effets audionumériques et de leur contrôle

Le contrôle des effets usuels

Nous avons vu dans les chapitres précédents toute la richesse des effets audionumériques et des transformations sonores qu'ils permettent. Le contrôle proposé dans la plupart des cas est gestuel, temps-réel ou temps-différé selon la méthode, et effectué par le biais d'interfaces graphiques et gestuelles. Une autre manière consiste à donner en temps différé les valeurs de contrôle sous forme de courbes de contrôle que l'on dessine segment par segment : c'est le principe de l'automation dans les séquenceurs audionumériques et MIDI, des partitions *Music V* et *CSound*.

Définition des effets adaptatifs

Une spécificité de quelques effets, dont le compresseur et l'expandeur, la wha-wha sensitive, nous a semblé intéressante à creuser : le contrôle automatique par des paramètres extraits du son. C'est le principe qui est au cœur des effets audionumériques adaptatifs et que nous voulons généraliser à tous les effets possibles. L'intérêt principal de ce contrôle automatique réside dans le fait que l'on peut, en temps-réel, donner des valeurs de contrôle de l'effet dont l'évolution suit celle du son.

Définition Un effet adaptatif est un effet dont les paramètres de contrôle sont pilotés par des paramètres extraits du son (*cf. fig.* 6) [Verfaille and Arfib, 2001; Verfaille, 2002]. Il est aussi appelé "effet ou transformation basée sur le contenu" [Amatriain *et al.*, 2003] (*content based transformation*) ou encore "effet dépendant du contexte et piloté par des descripteurs" [Lindsay *et al.*, 2003] (*context description-driven context-sensitive effect*). Ainsi, le contrôle est automatisé en fonction de propriétés intrinsèques du son. Nous n'avons rien inventé, puisque le principe d'adaptation est à l'origine de

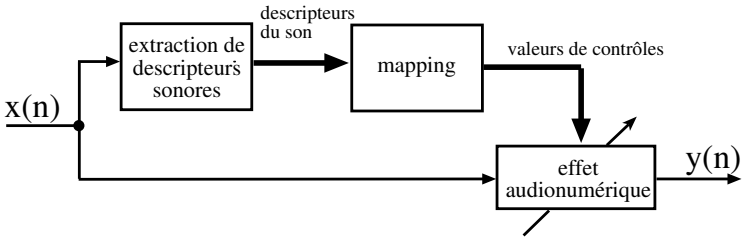


FIG. 6 – *Principe simplifié d'un effet audionumérique adaptatif.*

l'évolution de systèmes simples ou complexes, biologiques ou sociologiques. La nouveauté est la volonté de généraliser cette adaptation aux effets audionumériques.

Les descripteurs du son

Si l'on choisit bien les descripteurs que l'on utilise, ils seront corrélés avec la perception que l'on a du son. Pour les effets sur la dynamique (compresseur, expandeur, *noise gate*, limiteur), c'est le niveau d'entrée qui modifie le niveau de sortie. Pour l'*auto-tune*, c'est la hauteur en entrée qui est discrétisée et corrigée. Pour la wha-wha sensitive, tout change puisque c'est la dynamique du son qui contrôle les modifications du timbre. Une deuxième généralisation consiste à permettre qu'un effet portant sur une dimension perceptive soit contrôlé par un descripteur portant sur une autre dimension perceptive. Et dans ce cas, pourquoi se limiter à un seul descripteur ? Nous pouvons utiliser une combinaison de plusieurs descripteurs d'un son, voire de plusieurs sons : c'est la troisième généralisation du principe d'adaptation.

Réintégrer le contrôle gestuel

Les méthodes d'analyse-synthèse et de *machine learning* [Métois, 1996] utilisent des descripteurs du son afin de le modéliser. Ensuite, ces modèles permettent de synthétiser des sons à l'aide d'un contrôle gestuel. Le contrôle gestuel des effets audionumériques a jusqu'ici été occulté de notre discours ; il est pourtant fondamental pour certains effets dans les contextes d'acquisition et de studio, et pour la plupart des effets dans les contextes de composition en studio et de diffusion. Il nous a semblé très intéressant de s'inspirer du contrôle gestuel de la synthèse sonore, qui fournit des pistes très intéressantes, que ce soit dans ses premières applications hors temps-réel avec les programmes *Music N* et *CSound*, ou dans les environnements temps-réel tel *Max/MSP*. Une question importante que soulève l'utilisation du contrôle gestuel est le *mapping* (mise en correspondance entre paramètres). La quatrième et dernière généralisation que nous proposons est celle du contrôle gestuel, qui peut aussi bien porter sur les contrôles de l'effet (cas des effets usuels) que sur le *mapping* entre descripteurs et contrôle de l'effet.

Des outils à développer ou créer

Quels outils mettre en place ?

Pour permettre le contrôle adaptatif et gestuel des effets audionumériques, nous devons nous construire toute une palette d'outils. Les premiers outils sont des programmes d'effets et de traitements supportant des contrôles variables : des effets classiques (écho, chorus, transposition, etc.) et des effets de haut niveau, permettant la modification ou conservation de paramètres perceptifs (par exemple la dilatation/contraction temporelle conservant l'expressivité). Nous devons ensuite écrire ou détourner des programmes d'extraction de descripteurs sonores pour offrir une grande variété de contrôles au musicien. Nous devons aussi définir une structure de *mapping* en détaillant les

fonctionnalités et leur ordonnancement entre descripteurs et contrôles. Nous devons enfin donner des points d'accès au contrôle gestuel sur le *mapping*.

Nécessité d'une mise en œuvre spécifique

La mise en œuvre des effets audionumériques va nécessiter la plupart du temps des méthodes différentes de celles utilisées lors de la mise en œuvre de la version non adaptative de l'effet. Nous allons le montrer dès à présent avec la modification de la longueur d'une ligne à retard, afin de sensibiliser le lecteur sur ce point très important, qui justifie la plupart des développements qui viennent par la suite de ce mémoire.

Une ligne à retard permet entre autres de simuler les réflexions d'une onde sur une paroi parfaitement réfléchissante, et sert notamment pour créer un écho artificiel. Si l'on rend cet effet d'écho adaptatif, cela signifie que l'on va, à un instant donné, modifier la longueur de la ligne à retard. Cela peut se faire en raccourcissant la ligne à retard (troncature) ou en la rallongeant (bourrage de zéros). De plus, le temps de retard τ peut être inversement proportionnel à la fréquence fondamentale du signal, s'il est harmonique, ou pas.

Troncature d'une ligne à retard On utilise pour l'exemple un signal parfaitement harmonique, composé d'une seule sinusoïde à $f = 235.2\ Hz$ échantillonnée à $F_e = 44.1\ kHz$. La ligne à retard comporte 1500 échantillons, puis au temps $T = \frac{3000}{F_e} s$, elle est raccourcie à 1125 échantillons[1]. Cela signifie que les 375 échantillons de la fin de la ligne à retard sont perdus, ils ne circuleront plus dans la ligne. Le taux de réinjection est de 0.9. On voit (*fig.* 7 gauche) que ceci introduit une rupture du signal qui se répercute sur toutes ses répliques. L'enveloppe effectue un saut, ce qui introduit des hautes fréquences dans le signal (voire des *clicks*). Nous avons choisi dans cet exemple de retirer un nombre entier de périodes du signal ($\frac{f}{F_e/375} = 8$), si bien que le signal d'entrée et le signal en sortie de la ligne auquel il s'ajoute sont en phase. Si l'on effectue maintenant une troncature d'un nombre non entier de périodes (*fig.* 7 droite), le signal d'entrée et le signal en sortie de la ligne à retard ne sont plus en phase : on remarque alors que leur somme donne un signal dont l'amplitude est moindre, du fait du déphasage. Notons toutefois que ceci est classique, et existe de toute façon pour n'importe quelle ligne à retard non modifiée dont la longueur n'est pas un multiple de la période du signal.

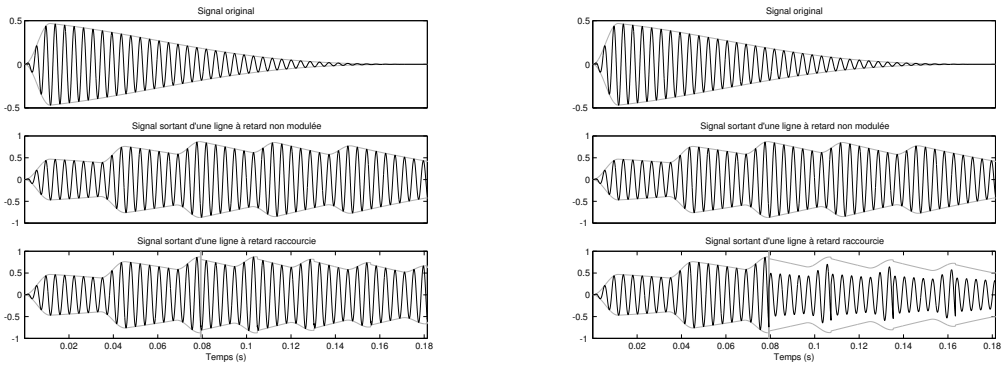


FIG. 7 – *Modification de la longueur d'un ligne à retard par troncature de N périodes (N entier à gauche, N non entier à droite).*

Allongement d'une ligne à retard Allonger une ligne à retard consiste à ajouter des zéros à la fin. il s'agit d'un bourrage de zéros classique. On utilise le même signal que précédemment, mais au lieu de retirer 375 échantillons, on ajoute 375 zéros en fin de ligne, avant de poursuivre les itérations

[1]Par soucis de lisibilité des figures, nous avons choisi des délais de l'ordre de 30 ms. Ils correspondent du point de vue perceptif à un filtrage en peigne RII. Cependant, les illustrations restent valides pour des délais plus grands, perçus comme des échos

du filtre. Dans les deux cas, on introduit deux clicks, l'un à la fin de la longueur d'origine, l'autre à la fin de la nouvelle longueur de la ligne. On remarque à nouveau la diminution de l'amplitude due au déphasage des signaux (*fig.* 8 droite).

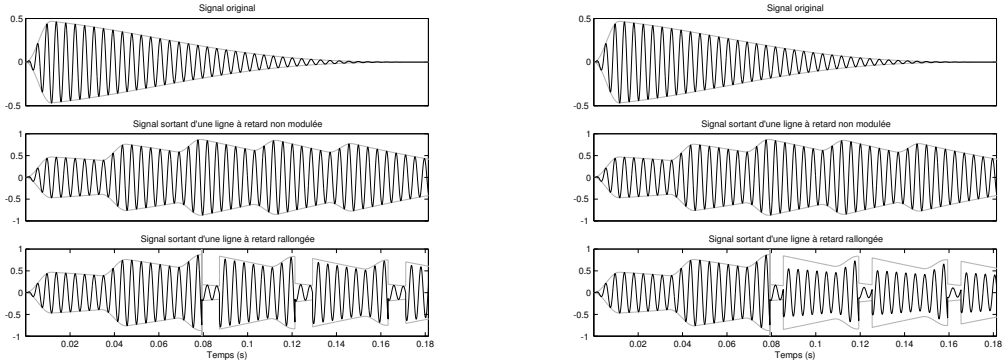


FIG. 8 – *Modification de la longueur d'un ligne à retard par ajout de N périodes (N entier à gauche, N non entier à droite).*

Quid des sons à traiter ?

Nous avons besoin d'une base de sons dont les descripteurs varient : on s'intéressera alors à tout type de son, et pas seulement à des notes musicales isolées. Ainsi, des phrases de voix parlée ou chantée, des sons instrumentaux monophoniques et polyphoniques, des sons électroacoustiques, des sons écologiques correspondent tout à fait au cadre de notre étude. En effet, ils ont souvent des effets perceptifs intéressants (par exemple le vibrato [Desain and Honing, 1996], les transitions [Strawn, 1987]). De plus, de nombreux descripteurs évoluent puisque ces sons ne sont pas stationnaires. Enfin, ils sont souvent utilisés en musique électroacoustique, musique pour laquelle se destinent en premier lieu les outils que nous réalisons. Ceci dit, pour les exemples sonores, nous avons choisi principalement des sons instrumentaux et de voix parlée ou chantée, parce que c'est sur ces sons que la perception de la plupart des auditeurs est la plus apte à juger des transformations sonores. Si nous avions exclusivement utilisé des sons électroacoustiques, seul le public habitué à cette musique aurait perçu les potentialités des effets adaptatifs. Nous utiliserons donc certains sons pour les exemples, et d'autres pour la mise en œuvre musicale de ces effets.

Quel mapping ?

Le *mapping* est le terme usuel pour parler de la mise en correspondance entre paramètres. Dans notre cas, il s'agit de la manière de mettre en correspondance les traits caractéristiques du son et du geste avec les contrôles de l'effet audionumérique. Cette mise en correspondance entre traits et contrôles peut être explicite (donnée par des formulations mathématiques explicites) ou implicite (donnée par des modèles statistiques ou combinatoires, tels les réseaux de neurones). Nous développerons principalement le *mapping* explicite, que nous formalisons en deux étages : le premier se constitue d'un contrôle automatique de l'effet par des descripteurs extraits du son à l'aide de *mappings* spécifiques à définir ; le second se constitue d'un contrôle gestuel sur le *mapping* du premier étage.

Nous nous sommes inspiré du modèle de *mapping* en trois couches (*cf. fig.* 9) développé par notre équipe pour le contrôle gestuel de la synthèse sonore [Arfib *et al.*, 2002b], amélioration d'autres *mappings* [Wanderley *et al.*, 1998; Hunt *et al.*, 2000; Wanderley and Depalle, 1999]. Dans ce *mapping*, la première couche de *mapping* va du transducteur gestuel à une représentation perceptive des gestes, la troisième couche relie les paramètres perceptifs du son aux paramètres de synthèse, et la couche intermédiaire relie les paramètres perceptifs du geste à ceux du son. Nous avons conservé

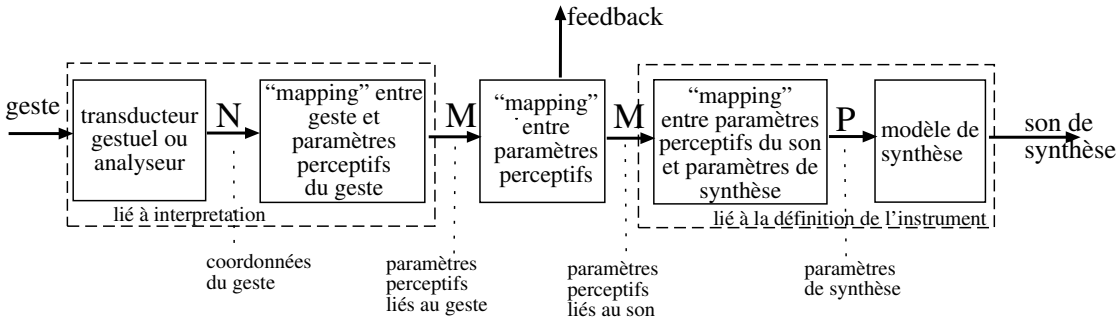


FIG. 9 – Mapping *en trois étapes, avec utilisation d'une couche perceptive comme intermédiaire entre paramètres du geste et paramètres de synthèse ou de l'effet.*

cette couche perceptive, par l'utilisation de descripteurs perceptifs. Notre *mapping* comprend aussi des étapes d'ajustement (*scaling* ou *fitting*), de façon à caler les plages de variation des paramètres de contrôle.

Plan de la suite de l'exposé

Nous allons dans un premier temps définir les effets audionumériques adaptatifs et leurs différentes formes. Nous pourrons alors détailler un à un les effets que nous avons réalisés, classifiés selon la taxonomie perceptive (*cf.* chap. 5). Il sera alors temps de parler de leur contrôle et des *mappings* utilisés. La structure de *mapping* en deux niveaux (un pour le contrôle par le son, un pour le contrôle par le geste) que nous avons mise au point sera détaillée. (*cf.* chap. 6). Nous donnerons des indications sur les adaptations de la mise en œuvre nécessaires à leur réalisation. Nous terminerons en insistant sur l'intérêt que présente le contrôle adaptatif et décrirons quelques applications musicales réalisées (*cf.* chap. 7).

Chapitre 5

Effets adaptatifs

Il y a quelques années, de jeunes compositeurs, à Vancouver, à Marseille, et au MIT, se sont mis à faire de la musique avec des quanta sonores : je ne sais pas où ils arriveront, cela n'a pas vraiment d'importance, car c'est une exploration qu'il faut faire. Il s'agit donc là d'une théorie différente de celle de la synthèse additive qui provient de Fourier. Et il y aurait d'autres possibilités d'exploration à partir de théories situées en amont de l'utilisation des machines à calculer.
Iannis Xenakis [Albèra, 1997a]

Sommaire

5.1 Principe d'adaptation d'un effet

Nous allons présenter en détail les effets audionumériques du point de vue de leur structure. En effet, plusieurs configurations sont possibles, selon si l'on utilise un son ou plusieurs sons, selon si le descripteur est extrait du son avant ou après traitement, etc. Nous allons développer chaque configuration, afin de proposer un ensemble exhaustif de structures d'effets adaptatifs. Une fois le cadre général posé, nous verrons la chaîne de traitement employée, puis les effets adaptatifs, un par un, et par paramètre perceptif modifié d'après la classification des effets classiques, sachant qu'un effet adaptatif peut modifier d'autres paramètres perceptifs que ceux modifiés par l'effet classique dont il découle.

5.1.1 Définition des effets adaptatifs (A-DAFx)

Les effets audionumériques adaptatifs ont été présentés dans différents articles [Verfaille and Arfib, 2001; Verfaille and Arfib, 2002; Amatriain *et al.*, 2003; Verfaille, 2003; Verfaille, 2002; Arfib *et al.*, 2002b]. Ils ont pour acronyme A-DAFx, de l'anglais *Adaptive digital audio effects*. Il s'agit d'une généralisation des effets audionumériques existants (qu'ils soient mono ou multi-canaux) incluant un contrôle automatique. Ce contrôle utilise des valeurs extraites du signal [Arfib, 1998b], puis modifiées par des fonctions de correspondance (ou *mapping*) pour aboutir à des valeurs de contrôle de l'effet. Les fonctions de correspondances quant à elle peuvent être modifiées par les descripteurs sonores, mais aussi par un contrôle gestuel. Nous allons dans un premier temps regarder les effets existant sous ce nouveau jour, puis détailler plusieurs configurations et raffinements de cette description générale.

5.1.2 Effet adaptatif déjà existants

Nous avons vu que certains effets effectuaient une analyse du signal dont dépendait leur contrôle (*cf.* chap. 3). Il s'agit du compresseur, de l'expandeur, du limiteur, du *noise gate* , de la discrétisation de hauteur sur une échelle prédéfinie ou *autotune* [Antares, 2003], le changement de genre, etc. Ces effets sont des exemples d'effets adaptatifs. Ils ont été définis de manière à répondre à une tâche particulière : accorder la hauteur du signal (contrôle de la hauteur par la hauteur), imposer des critères à la dynamique du signal (amplitude contrôlée par l'amplitude). La synthèse croisée applique l'enveloppe spectrale d'un son à un second son.

D'autres effets adaptatifs ont été développé dans cette optique. Ainsi, l'effet *Contrast* des GRM Tools permet d'appliquer des compresseurs différents à différentes bandes de fréquence [INA-GRM, 2003]. Dans le même ordre d'idée, le *Spektral Delay* de Native Instruments permet d'appliquer des délais différents à des bandes de fréquences définies par l'utilisateur, ainsi que des modulations [Native Instrument, 2002]. La wha-wha sensitive permet de retrouver le comportement des cuivres notamment, en corrélant la brillance du son avec son intensité. Très récemment ont été proposés des effets, dont le chorus, contrôlés par le RMS ou le CGS [The Sound Guys, 2003]. Le *rack* Voice One de TC Helicon permet d'appliquer des changements de genre [TC-Helicon, 2002], selon le principe décrit par [Amatriain *et al.*, 2003], ainsi que des harmonisations adaptatives, dépendant de la hauteur et du contexte harmonique.

Toutes ces informations sont regroupées *tab.* 5.1, en considérant les dimensions perceptives (p-dim°) que sont la durée et le rythme (R), la dynamique (D), la hauteur (H) et le timbre (T).

Nom	**Effet (technique)**	**p-dim° modifiée**	**Contrôle**	**p-dim° de contrôle**	**Loi de conformation**
compresseur, expandeur	dynamique	D	RMS	D	segment de droites et hystérésis
AutoTune	transposition	H	F_0	H	quantification
wha-wha sensitive	filtrage	T	attaque	T	hystérésis
synthèse croisée	filtrage	T	$\mathcal{E}(f)$	T	directe
Sfx chorus	chorus	T	RMS, CGS	D, T	direct
changement de genre	transp. et filtrage	T & H	F_0	hauteur	
contrast (GRM Tools)	dynamique	D	bandes fréq.	D et T	
harmoniseur intelligent	transposition	T & H	F_0	H	règles d'harmonie
harmoniseur intelligent	filtrage	T & H	F_0, $\mathcal{E}(f)$	T & H	règles d'harmonie
changement de *swing*	dilat./contr. temp.	D & R	tempo	rythme	binaire =: ternaire

TAB. 5.1 – *Tableau des effets adaptatifs pré-existant à cette étude.*

Cette première étape dans l'adaptation en laisse présager d'autres, auxquelles cette étude va s'attacher. Tout d'abord, il faut une phase exploratoire de l'adaptation d'un bien plus grand nombre

d'effets. Ensuite, le descripteur d'une dimension du son peut contrôler un effet portant sur une autre dimension du son, comme c'est le cas de la wha-wha sensitive, où la dynamique du son module le timbre. On peut aussi considérer le cas où plusieurs sons entrent en jeu (*cf.* la synthèse croisée). De plus, les fonctions de transformation des descripteurs peuvent différer de celles du compresseur. Enfin, un contrôle gestuel peut porter sur la transformation du descripteur en paramètre de contrôle de l'effet. Ces différentes pistes sont l'objet de la suite de cette étude.

5.1.3 Effet auto-adaptatif

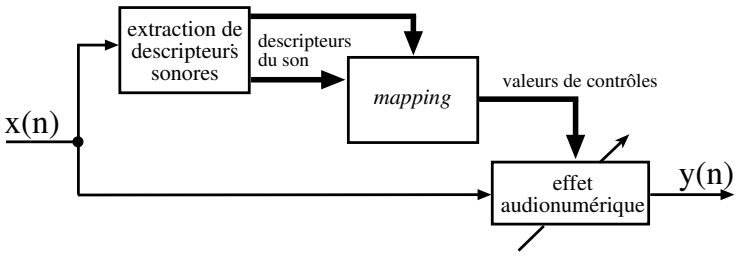


FIG. 5.1 – *Diagramme de l'effet auto-adaptatif avec un signal d'entrée.*

Le diagramme en bloc *fig.* 5.1 illustre les principales étapes de l'adaptation d'un effet audionumérique simple. Il est dit **auto-adaptatif** du fait qu'un seul signal d'entrée est utilisé. Le signal d'entrée sert à l'extraction des descripteurs du son, à la fois de bas et de haut niveau, dont des paramètres perceptifs (*cf.* chap. 4). La mise en correspondance des descripteurs du son avec les paramètres de contrôle de l'effet peut se faire via l'utilisation de non linéarités, afin de conformer les descripteurs ; elle peut aussi se faire via la combinaison de descripteurs (*cf.* chap. 6).

5.1.4 Effet adaptatif avec contrôle gestuel

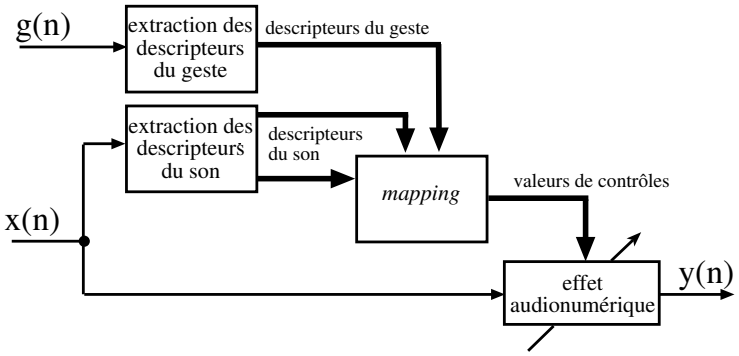


FIG. 5.2 – *Diagramme de l'effet auto-adaptatif avec contrôle gestuel.*

Pour un effet usuel, le contrôle gestuel porte sur les paramètres de contrôle de l'effet, et en permet la modification à l'aide de transducteurs tels que des interrupteurs, des potentiomètres linéaires ou circulaires, qu'ils soient réels (et pilotant via des données MIDI ou via le port série) ou bien virtuels (c'est-à-dire représentés à l'écran et manipulés à l'aide de la souris ou du contact du doigt sur une surface tactile), aussi bien que tout autre transducteur (instrument MIDI, gant de réalité virtuelle, pédale, etc.).

Pour un effet adaptatif, le **contrôle gestuel** porte à la fois sur le contrôle de l'effet et surtout sur le *mapping* entre descripteurs du son et contrôles de l'effet. Ce *mapping* peut être modifié par le geste (à l'aide d'un transducteur gestuel identique à ceux précédemment cités), comme illustré *fig.* 5.2. De

cette manière, le contrôle gestuel est ré-introduit de manière plus général : à la fois comme contrôle direct (ce qui jusqu'ici a toujours été le cas) et comme contrôle de plus haut niveau, sur le *mapping*. Le *mapping* peut être modifié en passant d'une fonction non-linéaire à une autre, en changeant les bornes de variation d'un paramètre, en changeant la combinaison des descripteurs sonores (et leur pondération), etc. Nous verrons (*cf.* chap. 6) les différentes fonctions de correspondance utilisées ainsi que tous les paramètres permettant de les modifier. Pour chaque effet adaptatif (*cf.* sec. 5.2, *cf.* sec. 5.7) pour lequel une mise en œuvre en temps-réel a été réalisée, nous expliciterons quelques contrôles gestuels utilisés et leur pertinence.

5.1.5 Effet adaptatif croisé

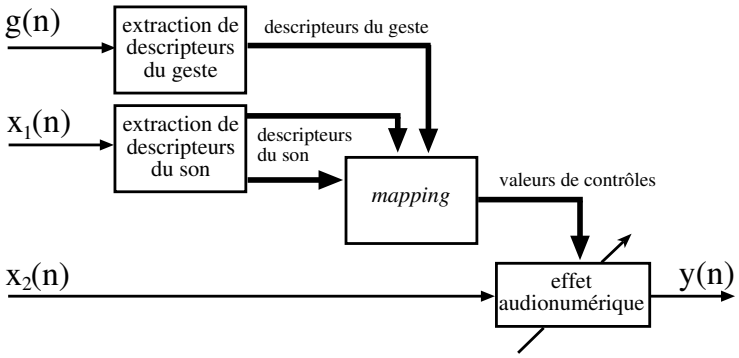


FIG. 5.3 – *Diagramme de l'effet adaptatif croisé, avec deux signaux d'entrée.*

Jusqu'à présent, l'effet adaptatif n'utilise qu'un seul signal d'entrée. Si maintenant on utilise deux signaux d'entrée, l'un utilisé pour l'extraction des descripteurs sonores et l'autre sur lequel on va appliquer l'effet proprement dit, on obtient un effet **adaptatif croisé** (*cross-adaptive*), illustré par le diagramme *fig.* 5.3. La dénomination "effet adaptatif croisé" est une analogie avec "synthèse croisée" (*cf.* sec. 3.6.4), où des propriétés de deux signaux (la source et le filtre, ou la partie harmonique et le résidu, par exemple) sont utilisées pour synthétiser un troisième signal. Bien évidemment, lorsque les deux signaux d'entrée sont identiques, on retrouve l'effet auto-adaptatif présenté précédemment *fig.* 5.2. C'est une généralisation du limiteur croisé (*cf.* sec. 3.2.4), effet adaptatif qui modifie le niveau d'un signal (musique) en fonction d'un second signal (parole). Cette généralisation permet différents types d'effets :

- un son peut être considéré comme contrôle de l'autre son, l'effet ne s'appliquant que sur le second son (cas du limiteur croisé) ;
- les deux sons contrôlent l'effet via des descripteurs croisés, et l'effet s'applique sur un seul des deux sons ;
- l'effet s'applique sur les deux sons (principe général de la synthèse croisée), et est contrôlé par l'un des deux sons, voire par les deux.

5.1.6 Effet adaptatif à rétrocontrôle

En physiologie, les systèmes à **rétrocontrôle** (ou réinjection, ou *feedback*) sont des systèmes qui s'auto-régulent. Pour ce faire, l'état du système est analysé à intervalles réguliers, et les modifications nécessaires pour atteindre ou conserver un état particulier sont appliquées. La notion de rétroaction a depuis largement dépassé son domaine originel, et s'utilise notamment dans des systèmes dynamiques (discipline de l'Automatique), et dans les systèmes d'Interface Homme Machine (IHM). Les effets adaptatifs n'y échappent pas : en effet, les descripteurs servant à contrôler l'effet peuvent provenir du signal en entrée de l'effet (effet auto-adaptatif ou *feedforward*, *cf.* sec. 5.1.3) mais aussi du signal en sortie de l'effet (*cf. fig.* 5.4). L'effet adaptatif à rétrocontrôle

consiste donc à prendre pour $x_1(n)$, le signal utilisé pour l'extraction des descripteurs, le signal de sortie $y(n)$. A l'inverse, lorsque les deux signaux d'entrée sont identiques (effet auto-adaptatif), il s'agit d'un **postcontrôle**, du fait que les descripteurs sonores sont d'abord extraits avant que ne soit appliqué l'effet sur la portion de son décrite par les descripteurs.

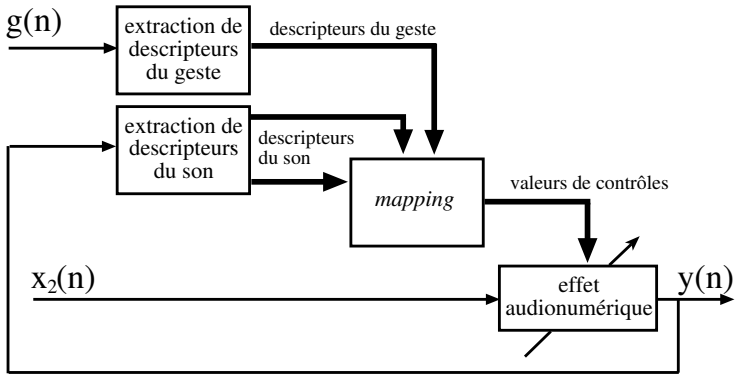


FIG. 5.4 – *Diagramme de l'effet adaptatif à rétrocontrôle : les descripteurs sont extraits du signal après traitement.*

L'effet adaptatif à rétrocontrôle permet par exemple, de manière itérative, de faire coïncider certains critères du signal de sortie avec des exigences données à l'avance. Cela permet aussi de rendre moins évidente la relation entre le son traité et son traitement, du fait que la cohérence descripteur–contrôle en terme d'évolution ne se fait plus sur le son d'entrée uniquement.

5.1.7 Effet adaptatif multi-canal

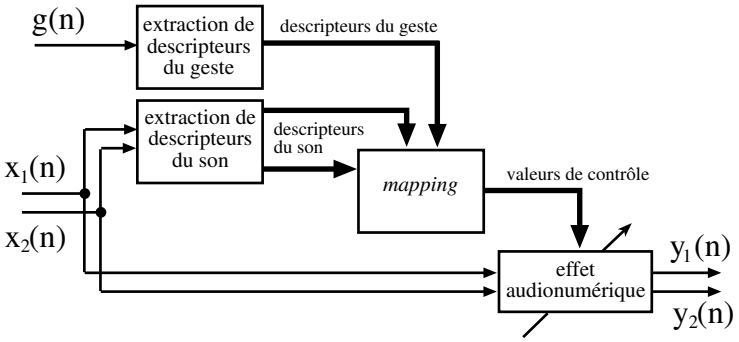


FIG. 5.5 – *Diagramme de l'effet adaptatif croisé stéréophonique : un signal stéréophonique est utilisé à la fois pour l'extraction de descripteurs et comme entré d'un effet stéréophonique. La sortie de l'effet est elle aussi stéréophonique.*

Enfin, lorsque le signal d'entrée est **multi-canal** (par exemple un signal stéréophonique, *cf. fig.* 5.5), le traitement de l'effet peut être beaucoup plus complexe, du fait que les descripteurs sonores proviennent de chaque canal séparé ainsi que de l'intercorrélation entre les deux canaux (descripteurs de corrélation). L'effet adaptatif peut avoir :

- en entrée un signal mono-canal et en sortie un signal multi-canal (cas de la spatialisation, adaptative ou non, de l'écho stéréophonique) ;
- en entrée et en sortie un signal stéréophonique (soit par des traitements parallèles et indépendants des canaux, soit par des traitements prenant en compte les deux canaux d'entrée ; un exemple est donné par l'équaliseur adaptatif, *cf.* sec. 5.6.1).

5.1.8 Relation entre filtre adaptatif en télécommunications et effets adaptatifs

Un cas particulier de l'effet adaptatif à rétrocontrôle est le filtrage adaptatif utilisé dans les télécommunications, différent du filtrage à but musical que nous présenterons (*cf.* sec. 5.6.1). Connaissant un signal $x_s(n)$ composé du mixage d'un signal utile $x_u(n)$ (la voix du locuteur B) et d'un signal parasite $x_p(n)$ (la voix du locuteur A arrivant par l'écouteur, filtrée, et réinjectée dans le microphone ou s'exprime le locuteur B) ainsi que le signal parasite $x_p(n)$, on cherche à retrouver le signal $x_u(n)$. On procède de manière itérative le meilleur filtrage $\mathcal{F}$ du signal mixé $x_s(n)$, c'est-à-dire le filtre qui fasse disparaître de la mixture le signal parasite : $x_s * \mathcal{F}(n) = x_u(n)$. Le but de ce filtrage est de faire disparaître l'écho de sa propre voix qui apparaît dans les téléphones [Haykin, 1996]. On comprend l'intérêt de telles techniques, tout en remarquant que pour l'application de contrôle variable et automatique que nous proposons dans un but musical avec les effets adaptatifs, cette formulation ne soit pas adaptée.

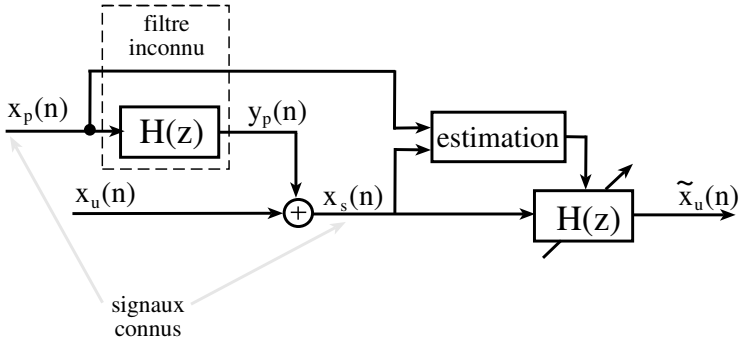


FIG. 5.6 – *Diagramme du filtrage adaptatif utilisé en télécommunications : au signal utile $x_u(n)$ inconnu est ajouté un signal parasite $x_p(n)$. La somme de ces signaux est connue. Le système estime le filtre $H(z)$ appliqué à $x_p(n)$ afin de pouvoir soustraire l'estimation de ce signal $\tilde{x}_p(n)$.*

5.1.9 Le tour des effets adaptatifs

Munis des descriptions des différents types d'effet adaptatif, nous pouvons prendre en compte toutes les éventualités que nous nous proposons d'investiguer par la suite : des signaux mono ou multi-canaux, un seul signal en entrée ou deux signaux différents pour l'effet et l'extraction de descripteurs, l'utilisation du signal d'entrée ou de sortie pour l'extraction des descripteurs. Les possibilités de traitements offerts par les effets adaptatifs sont encore plus nombreux qu'avec les effets traditionnels, aussi nous présentons ici (*cf. tab.* 5.2) un ensemble "assez complet" mais qui ne se veut pas plus exhaustif que la présentation des effets traditionnels (*cf.* chap. 3).

5.2 Effets adaptatifs sur la dynamique

Les effets adaptatifs portant sur la dynamique utilisent un paramètre de contrôle pris comme signal $c(n)$ pour modifier le signal, échantillon par échantillon, à l'aide d'une simple multiplication $y(n) = c(n) \cdot x(n)$. Dans le cas d'une sinusoïde pure $c(n) = \sin(2\pi f_{mod} n)$, la multiplication du signal $y(n)$ par la $c(n)$ s'appelle modulation en anneau, *cf.* sec. 2.3.1. La multiplication de $y(n)$ par $1 + c(n)$ s'appelle modulation en amplitude *cf.* sec. 2.3.1. Dans les deux cas, une modulation à basse fréquence (inférieure à 20 *Hz*) est perçue comme une modification de l'amplitude. C'est le cas qui nous intéresse ici. Cette modification s'appelle trémolo et est habituellement de fréquence et de profondeur constante, du fait que la modulante $c(n)$ est une sinusoïde pure. Cependant, on imagine plusieurs scénarios différents : une sinusoïde dont la fréquence varie (trémolo adaptatif), une courbe $c(n)$ provenant de paramètres du son (changement de niveau adaptatif, incluant le compresseur,

nom	dynamique	durée	hauteur	espace	timbre	temps-réel
a-changement de niveau	×					oui
a-changement de nuance	×					non
violoning	×					oui
a-trémolo	×					oui
a-pseudo-trémolo	×					non
a-changement de trémolo	×					non
a-trémolo spectral	×					non
a-dilatation		×				non
a-dilatation avec conservation de la durée		×				non
a-dilatation avec conservation de l'attaque		×				non
changement de *groove*		×				non
a-transposition (sans formants)			×		(×)	oui
a-transposition (avec formants)			×			oui
a-changement de hauteur			×			non
a-discrétisation de hauteur			×			oui
a-harmoniseur			×			oui
a-écho granulaire				×		oui
a-panoramisation				×		oui
a-panoramisation spectrale				×		oui
a-révérbération				×		oui
a-changement de distance				×		oui
a-spatialisation				×		oui
a-directivité				×		oui
effets sur l'enveloppe						
a-modif. enveloppe spectrale					×	non
a-changement du centroïde					×	non
a-équaliseur					×	oui
a-filtre en peigne					×	oui
a-filtre résonant					×	oui
a-filtre avec résolution arbitraire					×	oui
a-compresseur spectral					×	oui/non
a-wha-wha					×	oui
changement de voyelle					×	non
voice impersonator					×	oui
effets sur la phase						
a-*flanger*					×	oui
a-chorus					×	oui
a-*phaser*					×	oui
a-chuchotement					×	non
effets sur le spectre et sa structure						
a-vibrato			(×)		×	oui
changement de vibrato			(×)		×	non
a-changement d'échelle			(×)		×	non
a-dual detune			(×)		×	non
effets sur le spectre et l'enveloppe						
a-distorsion					×	non
a-conformation spectrale					×	non
a-synthèse croisée					×	oui
a-modulation en anneau spectrale					×	non
a-robotisation			×		×	oui
a-rééchantillonnage			×		×	non
a-brassage		×			×	non
martianisation			×		×	non
changement de prosodie/d'intonation			×		×	non
a-modulation en anneau			×		×	oui
a-modulation en anneau spectrale			×		×	oui
a-transposition sans conservation formants			×		×	non
a-panoramisation-octaviation			×	×	×	non

TAB. 5.2 – *Tableau des principaux effets audionumériques adaptatifs, indiquant les paramètres perceptifs modifiés par l'effet, et le type de mise en œuvre réalisée (temps-réel ou non).*

l'expandeur, etc.), un changement de niveau d'après des indices de segmentation (changement de nuance), un effet de *violoning* et un changement de sonie.

5.2.1 Changement de niveau adaptatif (TR, NTR)

Les effets de changement de niveau (modification du gain) tels que le compresseur et le limiteur (*cf.* sec. 3.2.4), l'expandeur et le *noise gate* (*cf.* sec. 3.2.3) sont adaptatifs et utilisent le gain et une fonction de conformation [Zoelzer, 1997; Dutilleux and Zoelzer, 2002]. Ils sont conçus pour répondre à des besoins spécifiques. En utilisant une fonction de conformation différente, ainsi que des paramètres différents du RMS, on obtient des effets différents. Ainsi, nous avons réalisé un atténuateur de voyelles (utilisation d'un indice de voisement), de consonnes (idem), d'attaques (indice de rupture de Masri, flux spectral et CGS) en utilisant la fonction de conformation sin. C'est le principe, en plus simple, du *de-esser*, qui filtre les sifflantes. Le signal est donné par $y(n) = x(n).c(n)$.

Pour éviter le problème de l'apparition de *clicks* lorsque les variations du contrôle sont brusques, il faut effectuer un filtrage passe-bas. La courbe de contrôle est sous-échantillonné par rapport au signal, donc il faut dans un premier temps la ré-échantillonner, puis passer les échantillons dans un filtre passe-bas. Ceci concerne la mise en œuvre du changement de niveau échantillon par échantillon. Une autre solution consiste à mettre en œuvre cet effet par blocs : chaque petit grain de signal se voit appliqué un gain $g \in [0; 1]$, et l'effet de fenêtrage diffuse l'information temporellement, ce qui revient à un filtrage, il n'y aura alors pas de variations trop brusques même si le gain passe de 0 à 1 d'un grain au suivant.

Exemples sonores

- Piste n°21-CD1 ♫♪ : changement de niveau adaptatif appliqué à la Piste n°3 ♫♪. L'amplitude $c(n)$ est donnée par une combinaison linéaire de l'amplitude RMS_{512} et du centroïde $cgs_3(n)$ normalisée. Le son obtenu voit ses attaques et ses décroissances ainsi que les passages de faibles brillance diminués, et le rendu perceptif est que les notes sont plus sèches.
- Piste n°22-CD1 ♫♪ : changement de niveau adaptatif appliqué à la Piste n°5 ♫♪. Pour ce son de type "trame temporelle" (ou "nappe"), l'utilisation de n'importe quel paramètre permet de rajouter de grandes variations de niveau : ici, on utilise le RMS, qui varie très peu dans le son original, mais qui, une fois normalisé avec $\mathcal{N}_1$ puis modifié par la fonction $\mathcal{H}_{sin}$, varie dans $[0; 1]$ en renforçant la proximité du 0 ou du 1, ce qui creuse davantage la courbe de contrôle.
- Piste n°23-CD1 ♫♪ : appliqué à la Piste n°5 ♫♪. On utilise maintenant la courbe du RMS inversée en amplitude, avec le même *mapping* : les observations sont identiques, et cette fois ci, c'est un peu le "son complémentaire" qui a été réalisé, puisque d'amplitude maximal lorsque l'autre est d'amplitude minimal, et réciproquement.
- Piste n°24-CD1 ♫♪ : changement de niveau adaptatif appliqué à la Piste n°10 ♫♪. En utilisant l'indice de voisement auquel on applique la fonction de conformation $1 - \mathcal{H}_{sin}$, on obtient une sorte de sélecteur de consonnes.

5.2.2 Changement de nuance adaptatif (TR)

Il a été montré que la perception des niveaux d'intensité, connus sous le nom de "nuance" par les musiciens, correspond à une échelle de variation par pas de 3 dB. Ainsi, une note *forte*, notée **f** sur une partition, est 3 dB plus forte qu'un *mezzo-forte*, noté **mf**. A partir du moment où l'on dispose d'une segmentation du signal, on peut appliquer une modification automatique de nuance en fonction d'un paramètre du son. Par exemple, on peut appliquer des *crescendi* (augmentation régulière en dB de l'intensité) lorsque l'on monte la gamme, et des *decrescendi* lorsqu'on descend la gamme. On peut aussi estimer la nuance de chaque note, et ne modifier que certaines notes sur des critères de jeu par exemple. On peut encore utiliser le contrôle gestuel pour renforcer les effets de nuance, ou les diminuer (dans ce cas, il faut d'abord réaliser une analyse hors-temps réel du son, puis extraire une courbe décrivant la nuance N_{in}, afin de modifier le ratio $\frac{N_{out}}{N_{in}}$).

5.2.3 Violoning (NTR)

Le violoning consiste à gommer les attaques des instruments à cordes pincées (notamment la guitare). Cela s'obtient en limitant la pente de la courbe du RMS. Soit $p = \frac{RMS(t)-RMS(t-1)}{R_A}$ la pente du RMS, avec R_A le pas d'analyse et d'extraction du RMS. Soit p_{max} la pente maximale autorisée. Si $p < p_{max}$, $RMS(t)$ inchangé, si $p > p_{max}$, alors

$$grain_{out} = grain_{in}\frac{RMS(t-1) + pR_A}{RMS(t)}$$

Notons qu'un effet similaire peut s'obtenir avec un réglage particulier du compresseur (avec un grand temps de montée par exemple).

5.2.4 Trémolo adaptatif (NTR)

Le trémolo adaptatif est une modulation sinusoïdale d'amplitude (*cf.* sec. 3.2.5), dont la fréquence de modulation est fonction de descripteurs du son. On utilise ici l'échelle logarithmique, d'où la modulation d'amplitude en fonction du temps n :

$$y(n) = x(n).\left(1 - 10^{-\left(d(n)\frac{1-\sin(2\pi f_m n/F_e)}{40}\right)}\right) \tag{5.1}$$

avec $d(n)$ la profondeur du trémolo en dB et f_m la fréquence de modulation en Hz. L'idée du trémolo adaptatif consiste à faire varier la profondeur et la fréquence de modulation d'après des descripteurs du son (par exemple le RMS et le CGS). Pour ce faire, on utilise la modulation de phase et d'amplitude combinés. Soient la profondeur $d(n)$ et la fréquence de modulation $f_m(n)$ variable dans le temps. L'équation itérative donne la valeur de la fonction d'amplitude $c(n)$:

$$\varphi(n) = \varphi(n-1) + 2\pi f_m(n)\Delta t \tag{5.2}$$

$$c_{dB}(n) = d(n)\frac{1-\sin\varphi(n)}{2} \tag{5.3}$$

$$c_{log}(n) = 10^{-c_{dB}(n)/20} \tag{5.4}$$

Ceci nous conduit à l'équation (5.5) donnant le signal après traitement :

$$y(n) = (1 - c_{log}(n))\,.x(n) \tag{5.5}$$

La figure *fig.* 5.7 donne un exemple de courbe de contrôle et de trémolo adaptatif appliqué à la Piste n°2 ♫♪. Selon l'échelle que l'on utilise, le trémolo est plutôt "mou" (Piste n°25-CD1 ♫♪ en échelle linéaire) ou plutôt sec (Piste n°26-CD1 ♫♪ en échelle logarithmique).

5.2.5 Pseudo-trémolo adaptatif (TR, NTR)

Pour un pseudo-trémolo, la fonction de modulation (porteuse) n'est plus sinusoïdale, mais prédéfinie sous forme d'une table d'amplitude (comme une table d'onde), lue à vitesse variable. Ce contrôle est identique à celui du trémolo adaptatif : seule la courbe d'amplitude est différente (sinusoïdale ou donnée par l'utilisateur). Un exemple est donné Piste n°27-CD1 ♫♪ avec pour table d'amplitude *fig.* 5.8 et pour contrôle *fig.* 5.9, appliqué à la Piste n°2 ♫♪. Si l'on compare ce son avec Piste n°26 ♫♪ pour lequel la table utilisée est une sinusoïde (trémolo adaptatif), on saisit l'intérêt que l'on peut avoir de donner une courbe autre que la sinusoïde. Ici, la courbe possède deux pics, l'un en-dessous du second de 15 dB. L'effet de double rebond ne s'entend que pour les modulations lentes, à la fin de l'exemple. Puisque l'utilisateur dessine lui-même la période de la courbe de contrôle d'amplitude, il peut utiliser une forme basique telle qu'une forme triangulaire, (Piste n°28-CD1 ♫♪) un "pallier" (Piste n°29-CD1 ♫♪), etc. Selon la forme de la courbe utilisée, la perception du pseudo-trémolo adaptatif est légèrement altérée.

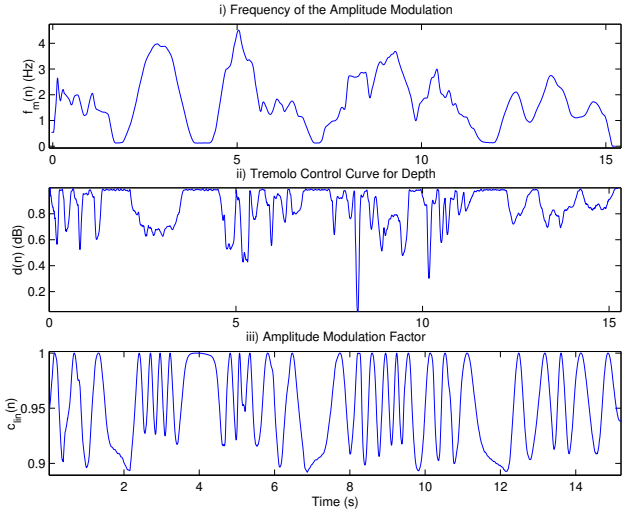


FIG. 5.7 – *Courbe de contrôle du trémolo adaptatif : i) fréquence de modulation $f_m(n)$, ii) profondeur $d(n)$ et iii) le rapport de correction d'amplitude résultant $c_{lin}(n)$.*

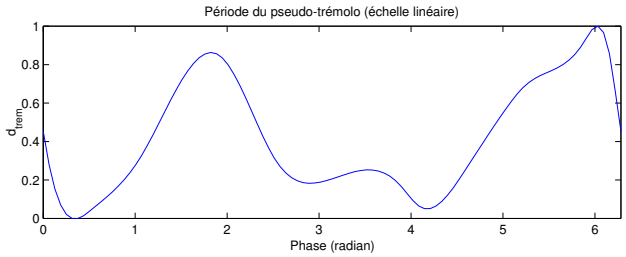


FIG. 5.8 – *Période de la table d'amplitude utilisée pour le pseudo-trémolo.*

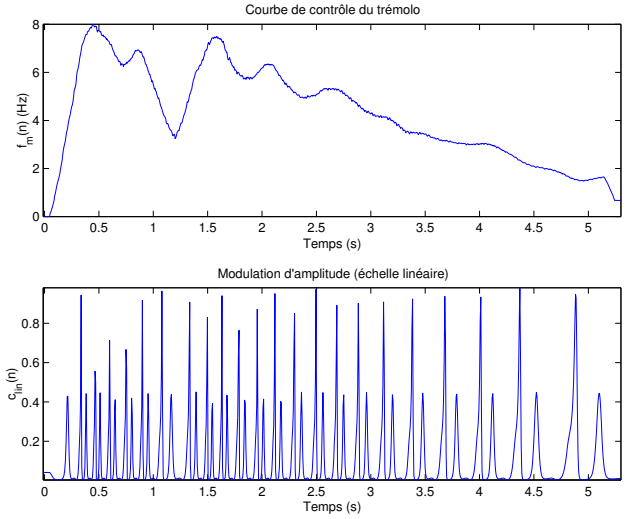


FIG. 5.9 – *Courbe de contrôle du pseudo-trémolo adaptatif : fréquence de modulation $f_m(n)$, et ii) le rapport de correction d'amplitude résultant $c_{lin}(n)$.*

5.2.6 Trémolo spectral adaptatif (NTR)

Le trémolo spectral adaptatif consiste à appliquer à chaque composante d'une TFCT un trémolo adaptatif avec une modulation de fréquence et d'amplitude différente pour chaque composante (ceci

peut aussi se faire avec une analyse additive, mais nous ne l'avons pas expérimenté). Du fait que l'on travaille dans le domaine spectral, il faut d'abord s'assurer que le traitement grain par grain permet lui aussi d'appliquer un trémolo, malgré le lissage de l'enveloppe dû à l'ajout-superposition des grains lors de la synthèse. Ceci ne pose aucun problème pour le trémolo en échelle linéaire *eq.* (3.6) puisqu'on utilise une fenêtre de Hanning, qui est une fonction sinusoïdale. Par contre, dès que l'on utilise l'échelle logarithme *eq.* (3.7), c'est plus délicat, comme nous allons le montrer.

Restrictions dues à la méthode granulaire

Problème de la taille des grains et du pas de synthèse Pour des fréquences relativement basses, par exemple $f_{tr} = 5\ Hz$ (*cf. fig.* 5.10 gauche et droite), l'utilisation d'une méthode granulaire pour appliquer un trémolo donne sensiblement la même enveloppe temporelle qu'une méthode temporelle, appliquée échantillon par échantillon. Cependant, si on y regarde de plus près, l'utilisation de petits grains ($R_s = 256$) au lieu de longs grains ($R_s = 2048$) fait passer l'erreur maximale de 6.45 dB à 0.47 dB. On a donc tout intérêt à utiliser de petites fenêtres et de petits grains.

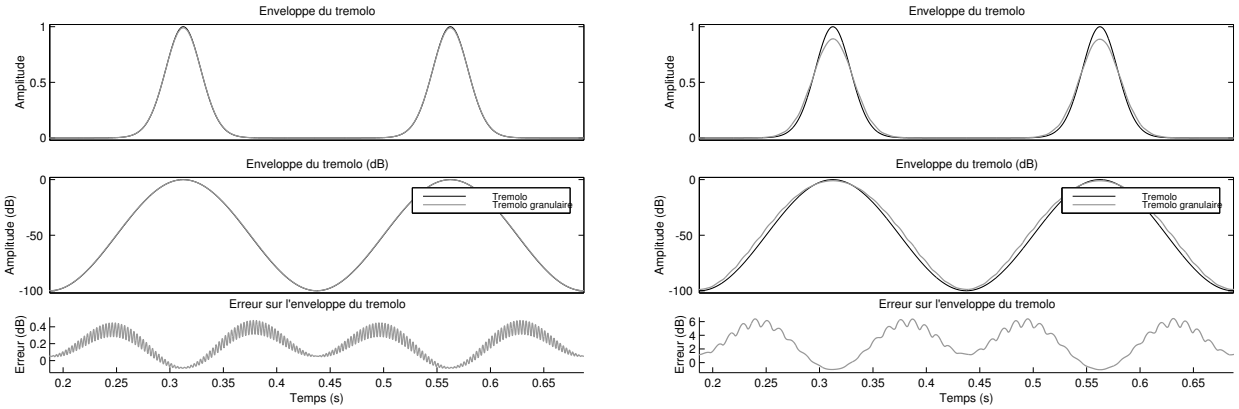


FIG. 5.10 – *Amplitude du trémolo granulaire ($f_{tr} = 5\ Hz$, $d_{tr} = 100\ dB$, $N = 512$, $R_s = 128$ à gauche et $N = 2048$, $R_s = 512$ à droite) : pour cette fréquence fixe et pour $N = 512$, $R_s = 128$, on retrouve avec la méthode granulaire l'enveloppe voulue, avec une erreur maximale de* $0.47\ dB$.

Pour des fréquences supérieures à 10 Hz, par contre, la méthode granulaire pose plus de problèmes. Le premier problème est que l'utilisation de grandes fenêtres implique un lissage de l'enveloppe créée, et ne permet par d'avoir la finesse de définition que l'on voudrait. Ceci se voyait déjà un peu à 5 Hz, mais cela devient flagrant pour des fréquences plus élevées. Ainsi, pour une modulation à 10 Hz, une fenêtre de 2048 échantillons est trop grande, quel que soit le pas (*cf. fig.* 5.11 gauche pour un pas de 128 échantillons). Dans tous ces exemples et sauf indication contraire, la fenêtre appliquée est une fenêtre de Hanning.

Le second problème qui se pose est le choix du pas entre deux grains successifs. Si le ratio R_s/N_s (avec N_s la taille de grain) est trop proche de 1 (tout en s'assurant que l'on utilise un pas tel que la somme des puissances des fenêtres donne bien un signal normalisé, *cf.* sec. 7.1), on risque de ne pas avoir assez de précision temporelle pour respecter la forme de la courbe d'amplitude. Il s'agit du problème de la modulation d'amplitude par ajout-superposition. Comme on peut le voir pour des grains de $n = 512$ (*cf. fig.* 5.11 droite), un pas de $R_s = 128$ échantillons donne une meilleure définition. En effet, pour un pas $R_s = 128$, la courbe est beaucoup plus lisse, et l'erreur moindre ($2,69\ dB$) : la différence n'est pas audible. Bien évidemment, pour des fenêtres encore plus petites (256 ou 128 échantillons), l'erreur passe en dessous de $1,5\ dB$.

Nous nous sommes demandé si la fenêtre de Hanning était la plus appropriée pour ce trémolo. Nous donnons dans la table 5.3 quelques ordres de grandeur de l'erreur en dB, pour différentes fenêtres en fonction du pas R_s et de la taille de la fenêtre (et donc du grain). Ils montrent que c'est

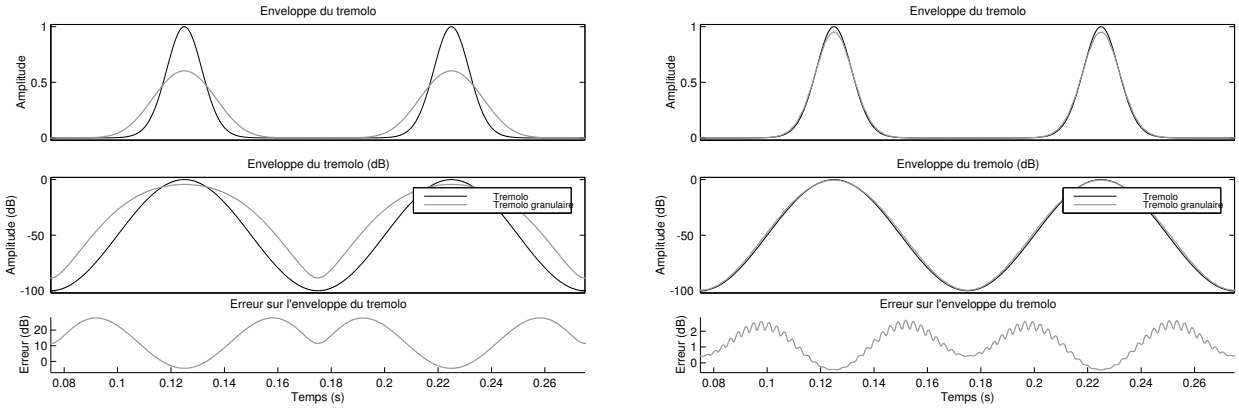


FIG. 5.11 – *Amplitude du trémolo granulaire ($f_{tr} = 10\ Hz$, $d_{tr} = 100\ dB$, $R_s = 128$).*
A gauche : même en diminuant le pas R_s (avec $N = 512$) par rapport à la fig. *5.10, l'enveloppe n'est toujours pas respectée. La forme est plus ressemblante, mais l'erreur encore trop importante ($27,8\ dB$).*
A droite : pour de petites fenêtres ($N = 128$), l'enveloppe est bien respectée, avec une erreur évoluant comme la courbe de trémolo et de valeur maximale de l'ordre de $2,5\ dB$.

Taille de grain	256	256	256	512	512	512	1024	1024	1024	2048	2048
Pas	32	64	128	64	128	256	128	256	512	128	256
Hanning	0.63	0.72	1.51	2.48	2.69	4.22	9.04	9.68	12.74	27.80	28.12
Blackman	1.63	1.63	3.64*	1.86	1.86	4.22*	5.83	6.10	8.98*	22.63	22.85
Blackman-Harris	3.20	3.20	9.31*	3.37	3.38	8.92*	4.01	4.07	11.37*	15.76	15.78
Triangulaire	0.85	0.93	1.22	3.21	3.50	4.55	11.43	12.42	15.46	32.95	33.71

TAB. 5.3 – *Erreurs d'amplitude en décibels pour un trémolo granulaire à $f_{tr} = 10\ Hz$ de profondeur $100\ dB$, réalisé avec différentes fenêtres, pour plusieurs tailles de grains et pas de synthèse. Les $*$ correspondent à des situations où le ratio pas sur taille de grain est supérieur à $1/4$, et où les fenêtres de Blackman et Blackman-Harris ne se recouvrent pas unitairement.*

bien la fenêtre de Hanning qui génère le moins d'erreur d'amplitude. Ceci a aussi été vérifié pour un trémolo à fréquence variable.

Mise en œuvre

Nous avons montré que pour des grains assez grands (à partir de $N_a = 512$ échantillons), il n'était pas possible de bien modéliser la courbe d'amplitude par une méthode granulaire. Si l'on se limite à $N_a = 256$ échantillons, la précision en fréquence, pour une mise en œuvre avec le vocodeur de phase, sera limitée à 256 paniers de fréquences. Si l'on désire plus de précision en fréquence, on peut utiliser la technique du bourrage de zéros. Une autre solution consiste à utiliser une fenêtre et un grain d'analyse de grande taille, et un grain de synthèse de taille N_s inférieure à N_a (une fenêtre de Hanning) avec bourrage de zéros des deux côtés, $[0...0\ w_{hanning}(N_s)\ 0...0]$ pour obtenir une fenêtre de synthèse de même taille que la fenêtre d'analyse, mais réduisant la zone temporelle où le grain est synthétisé. Le grain est ainsi réduit en taille, mais conserve sa cohérence de phase (la position temporelle de chaque échantillon).

On utilise donc à l'analyse une fenêtre de Hanning, avec des grains d'analyse de taille variant entre $N_s = 256$ et 4096, selon la définition en fréquence désirée par l'utilisateur. Pour la synthèse, chaque grain se voit appliquer une fenêtre qui réduit le grain en le gardant centré. Ainsi, on est sûr d'avoir une bonne définition temporelle de la modulation d'amplitude (faible erreur, en dessous de $1.5\ dB$ pour un pas inférieur ou égale à $R_s = 128$ échantillons) et une précision en fréquence. La contrainte à respecter scrupuleusement est d'avoir un pas de synthèse R_s (et donc un pas d'analyse)

du quart de la taille du grain de synthèse N_s [Allen, 1977; Allen and Rabiner, 1977; Harris, 1978; Arfib and Zoelzer, 2002b] :

$$R_s = \frac{N_s}{4}$$

Une fois le schéma d'analyse-synthèse mis en place, il nous faut définir quels paramètres vont contrôler cet effet. Pour contrôler N_a trémolos (un par panier de fréquence), il nous faut N_a données de contrôle. On peut utiliser les échantillons directement, le spectre de magnitude, le spectre de phase, l'enveloppe spectrale. Avec la pratique, on remarque qu'il est préférable d'avoir des valeurs de fréquences de modulation relativement proches (par exemple dans un intervalle restreint, du type $[5; 6\ Hz]$), ou variant lentement d'un panier de fréquence à l'autre. Pour cette raison, nous avons utilisé l'enveloppe spectrale extraite par la méthode du cepstre dans les exemples qui vont suivre.

Résultats

Tout d'abord, visualisons le sonagramme d'un son auquel on a appliqué un trémolo spectral adaptatif, contrôlé par l'enveloppe spectrale dans une plage de fréquence $[2; 3]$ *Hz*, *cf. fig.* 5.12. On remarque qu'effectivement, le trémolo appliqué dans chaque panier de fréquence est légèrement différent, ce qui va impliquer des déphasages entre les trémolos de chaque panier, et donc un effet de *phasing*, puisqu'à certains moments, certaines fréquences seront quasi-inexistantes tandis qu'à d'autres moments, elles seront présentes.

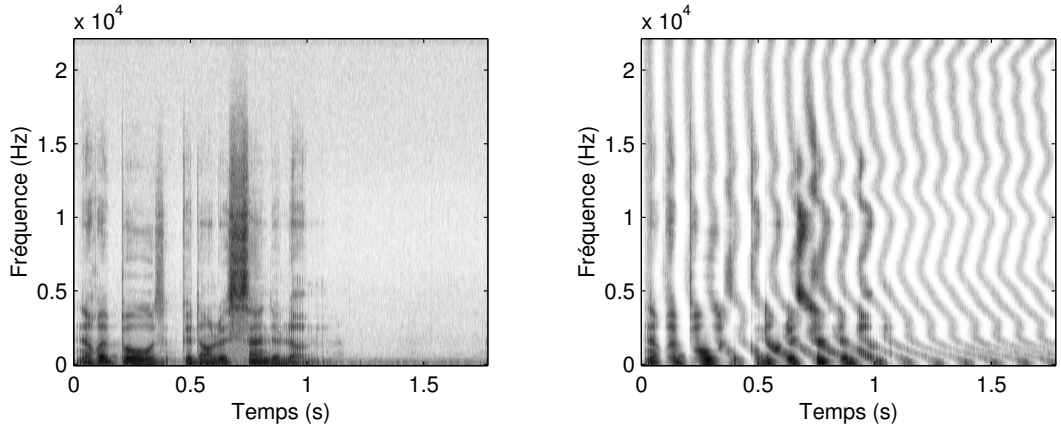


FIG. 5.12 – *Sonagramme de la Piste n°16 ♫♪ et sonagramme après trémolo spectral adaptatif : on voit apparaître des trémolos de fréquences variables, différentes et déphasées pour chaque composante spectrale.*

Effet de la taille de fenêtre et du pas Nous avons vu qu'il est préférable d'avoir de petites fenêtres pour respecter l'enveloppe temporelle. Malheureusement, l'utilisation de ces petites fenêtres laisse mieux passer les hautes fréquences que les grandes fenêtres. Finalement, nous avons préféré utiliser des fenêtres de synthèse de 512 ou 1024 éch. et des pas allant de 128 à 512 éch. (*cf.* Piste n°30-CD1 ♫♪ avec pour référence la Piste n°2 ♫♪).

Effet de la profondeur du trémolo Selon la profondeur du trémolo en dB, le trémolo est plus ou moins sensibles, de même que l'effet de *phasing*. Pour la Piste n°31-CD1 ♫♪, la profondeur est de 100 dB (avec pour référence la Piste n°2 ♫♪), tandis qu'elle est de 50 dB pour Piste n°32-CD1 ♫♪ et de 10 dB pour Piste n°33-CD1 ♫♪.

Effet du choix du méta-descripteur Nous avons appliqué un trémolo spectral adaptatif avec les mêmes fenêtres, la même fonction de contrôle de la fréquence de modulation $2[1 - \mathcal{N}_1(20\log_{20}(\mathcal{D}))] + 4 \in [4; 6]$, et différents méta-descripteurs. L'enveloppe spectrale $\mathcal{E}(nf)$ (de la Piste n°31-CD1 ♫♪) utilisée comme méta-descripteur donne un trémolo relativement homogène, où toutes les fréquences suivent à peut près la même modulation, puisque la courbe de contrôle est molle (avec pour référence la Piste n°2 ♫♪). Ceci explique le fait que l'aspect de trémolo soit prépondérant, et que l'aspect de *phasing* soit secondaire. L'utilisation du spectre de magnitude $\rho(n)$ permet de renforcer l'effet de *phasing*. Les variations de fréquences intra-paniers sont très grandes, si bien que les modulations se désynchronisent très vite entre un panier et ses voisins. L'utilisation de la forme d'onde $x(n)$ donne un trémolo quasiment normal, du fait que la forme d'onde est très mobile, contrairement au spectre d'amplitude ou à l'enveloppe spectrale. Les fréquences de modulation changement beaucoup, avec une valeur moyenne identique pour chaque panier de fréquence, si bien que les trémolos sont globalement synchrones.

5.2.7 Quelques autres effets

Nous avons imaginé quelques autres effets que nous n'avons pas eu le temps de réaliser. Le premier est un changement de sonie. Nous pensions modifier la sonie uniquement par le biais d'un changement de niveau du signal, grain par grain, sans modifier la composition spectrale du signal. Un algorithme itératif basé sur le calcul de sonie variant dans le temps, tel celui de [Zwicker and Fastl, 1999], devrait convenir. En ne modifiant pas la composition spectrale du son, on pense modifier le moins possible le timbre du son, et seulement son intensité perçue.

Le second effet est un changement de trémolo. L'idée nous est venue trop tard pour le réaliser, mais nous pouvons en donner les points clefs. Nous avons vu lors de la présentation des descripteurs sonores (*cf.* sec. 4.5.1), et notamment de l'amplitude calculée par RMS, que nous pouvons estimer l'amplitude du signal sans tenir compte de la modulation d'amplitude due au trémolo. On peut donc estimer la fréquence et la profondeur du trémolo en calculant le RMS avec une grande fenêtre (typiquement 8192 points), et appliquer un changement d'amplitude qui enlève le trémolo original, puis appliquer un autre trémolo.

$$y(n) = \frac{x(n)}{1 + \tilde{c}_{lin}(n)} \times (1 + c_{lin}(n)) \tag{5.6}$$

avec $\tilde{c}_{lin}$ la modulation d'amplitude estimée, et c_{lin} la nouvelle modulation d'amplitude.

5.3 Effets adaptatifs sur l'échelle temporelle

Modifier l'échelle temporelle d'un son consiste à modifier sa durée sans pour autant modifier la hauteur du son (*cf.* sec. 3.3.2). Une dilatation/contraction temporelle adaptative (ou non-linéaire, ou encore sélective) consiste à modifier de manière non-linéaire γ le facteur de dilatation ou de contraction du son. Ceci se réalise en utilisant des pas d'analyse R_A et de synthèse R_S différents, dont l'un est variable. Selon la méthode que l'on utilise (TD-PSOLA *cf.* sec. 2.4.1, additive *cf.* sec. 2.4.3, vocodeur de phase *cf.* sec. 2.4.2), on peut préférer que ce soit la pas d'analyse ou le pas de synthèse qui soit constant. Des explications sont données en sec. 7.1. Nous avons choisi d'utiliser le vocodeur de phase avec des pas de synthèse constant, afin que le résultat soit normalisé en amplitude.

Une fois que l'on sait réaliser une dilatation/contraction temporelle adaptative, on peut calculer la longueur du son de synthèse, et modifier la courbe de contrôle de manière à ce que le son de synthèse soit de même longueur que le son original : il s'agit de la dilatation/contraction temporelle adaptative préservant la durée globale. On peut aussi appliquer ce principe pour plusieurs portions de son, et ainsi imposer des temps de synchronisation entre le son original et le son de synthèse : il s'agit alors de la dilatation/contraction temporelle adaptative préservant la durée globale et la synchronisation localement (changement de *groove*). On peut aussi utiliser la courbe de contrôle

comme temps de lecture de la courbe, et en proposer une version réorganisée temporellement : c'est le brassage adaptatif, qui sera présenté avec les effets modifiant plusieurs paramètres perceptifs. Un dernier effet que nous allons présenter consiste à appliquer une dilatation temporelle en respectant des attributs perceptifs (tel que l'attaque des notes, le vibrato) : ceci permet alors un traitement plus réaliste, y compris pour une dilatation/contraction "linéaire".

5.3.1 Dilatation/contraction temporelle adaptative avec conservation de la hauteur

Définition

La dilatation/contraction temporelle adaptative (ou non linéaire) est une dilatation/contraction temporelle dont le facteur de dilatation γ varie au cours du temps et dépend d'un descripteur extrait du son à traiter. Ainsi, le facteur de dilatation est une fonction du ou des descripteurs $\mathcal{D}_i(t)$ utilisés, et donc du temps t :

$$\gamma(t) = f\left(\mathcal{D}_1(t), \ldots, \mathcal{D}_I(t)\right) \tag{5.7}$$

Fonctionnement

Le paramètre de contrôle est transformé en un facteur de dilatation/contraction, de façon à ce que ses bornes varient dans un intervalle adapté. Son minimum peut être inférieur à 1 (contraction) mais doit rester strictement supérieur à 0. Son maximum doit suivre la même contrainte, et peut être supérieur à 1. Typiquement, l'intervalle utilisé sera $[1/M; N]$ avec $M, N \in \mathbb{N}$.

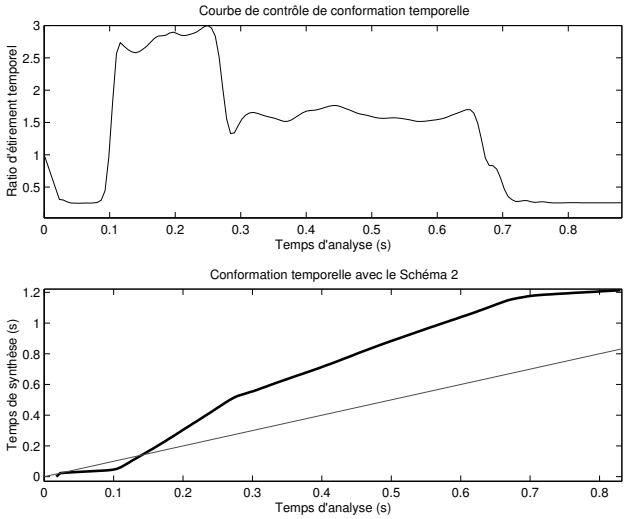


FIG. 5.13 – *Courbe de contrôle $\gamma(t)$ (*fig. *sup.) et temps du grain de synthèse (*fig. *inf.) en fonction du temps du grain d'analyse : une contraction temporelle correspond à une pente supérieure à 1 et une dilatation à une pente inférieure à 1.*

Intérêt de cet effet

D'une manière générale, cela permet de ré-interpréter une phrase musicale en changeant son expressivité. Par exemple, en utilisant l'indice de voisement comme contrôle, on peut créer le ralenti sélectif, qui ne ralentit que les voyelles, et laisse les consonnes d'une voix parlée ou chantée intelligible. Un autre exemple consiste à prendre un facteur de dilatation/contraction allant d'une valeur inférieure à 1 à une autre supérieure à 1. Cela permet alors de ralentir ou d'accélérer la lecture du son original. Prenons un premier exemple avec un son instrumental : une phrase musicale de flûte Piste n°8 ♫♪ . Sa forme d'onde est donnée *fig.* 5.14 gauche et sa fréquence fondamentale

fig. 5.14 droite. On effectue tout d'abord une contraction de la note la plus aiguë et une dilatation des autres notes (Piste n°34-CD1 ♫♪ , forme d'onde *fig.* 5.15 gauche et courbe de contrôle *fig.* 5.15 droite). On effectue ensuite la transformation inverse, à savoir une dilatation de la note la plus aiguë et une contraction des autres notes (Piste n°35-CD1 ♫♪ , forme d'onde *fig.* 5.16 gauche et courbe de contrôle *fig.* 5.16 droite). La dilatation se fait jusqu'à un facteur 4 et la contraction un facteur 1/4. Un autre exemple sonore est obtenu à partir de la Piste n°3 ♫♪ : Piste n°36-CD1 ♫♪ . Cette improvisation, originellement joué "dans le temps", devient hors-temps avec ce traitement. Il pourra être intéressant de synchroniser le son de synthèse avec le son d'analyse en certains points, pour conserver partiellement le rythme. Pour ce faire, il faut d'abord estimer la durée du son de synthèse, puis pouvoir modifier la courbe de contrôle de façon à imposer de conserve la durée globale sur chaque segment entre deux temps de synchronisation. C'est le propos des sections 5.3.2 et suivantes.

FIG. 5.14 – *Forme d'onde (à gauche) et fréquence fondamentale (à droite) du son de flûte original Piste n°8 ♫♪.*

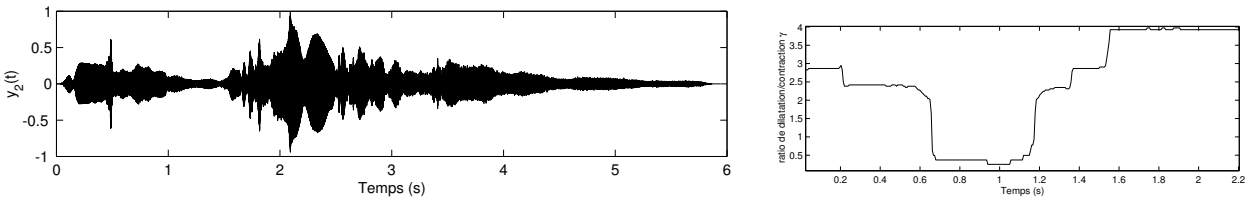


FIG. 5.15 – *Dilatation/contraction adaptative du son Piste n°8 ♫♪ : forme d'onde (à gauche) après dilatation des notes basses et contraction de la note la plus aiguë ; courbe de contrôle (à droite).*

FIG. 5.16 – *Dilatation/contraction adaptative du son Piste n°8 ♫♪ : forme d'onde (à gauche) après dilatation de la note la plus aigüe et contraction des notes basses ; courbe de contrôle (à droite).*

Dès lors que l'on applique un changement de durée (adaptatif ou non) à une voix, l'intelligibilité du contenu peut être altérée. Prenons pour exemple l'extrait sonore Piste n°11 ♫♪, constitué d'une phrase dans un langage imaginaire. Sa transformation est contrôlée par l'indice de voisement, ce qui donne un texte prononcé différemment mais toujours compréhensible (Piste n°37-CD1 ♫♪ , et avec une loi non-linéaire de conformation de la courbe de contrôle Piste n°38-CD1 ♫♪). Un autre exemple est donné avec l'extrait de voix de Pierre Schaeffer parlant de la musique électroacoustique (Piste n°15 ♫♪). Nous avons appliqué différentes courbes de contrôle, et les sons obtenus donnent tous une expressivité différente :

– Piste n°39-CD1 ♫♪ le locuteur est pressé, sans pause de respiration entre les mots ;

– Piste n°40-CD1 ♫♪ le locuteur ralenti par moment, sans respiration ; semble réfléchir à ce qu'il dit ;
– Piste n°41-CD1 ♫♪ le locuteur ralenti aux pauses respiratoires et au milieu de certains mots ; il semble réfléchir entre chaque mot.

Bien sûr, ces appréciations sont personnelles et n'ont pas été analysées en détail, mais des discussions avec des chercheurs du Laboratoire Parole et Langage (LPL) à Aix-en-Provence nous ont confirmé que c'est bien une porte d'accès à l'expressivité que nous nous sommes ouverte. Hélas, il n'y a pour l'instant aucune mesure quantitative de ces modifications, dont l'appréciation est laissée à l'utilisateur.

5.3.2 Estimation de la durée du son transformé

On utilise deux schémas de "conformation temporelle" ("*time warping*") pour effectuer le *mapping* entre les positions temporelles d'analyse et celles de synthèse : l'un avec pas d'analyse constant, l'autre avec pas de synthèse constants.

Pas de synthèse R_S constant Lorsque l'on utilise le schéma 1 avec pas de synthèse R_S constant, le temps d'analyse est donné en fonction de $\tilde{\gamma}$ par :

$$T_A^1(k) = T_A^1(k-1) + \tilde{\gamma}\left(T_S^1(k-1)\right) R_S \tag{5.8}$$

avec $\tilde{\gamma}$ une valeur interpolée de γ. Le temps de synthèse est donné par :

$$T_S^1(k) = T_S^1(k-1) + R_S \tag{5.9}$$

En notant $t_0 = T_A^1(0) = T_S^1(0)$, nous obtenons la formule récursive :

$$T_A^1(k) = t_0 + R_S \sum_{l=1}^{k} \tilde{\gamma}\left(t_0 + lR_S\right) \tag{5.10}$$

$$T_S^1(k) = t_0 + k\, R_S \tag{5.11}$$

Ce schéma est utilisé pour ne pas avoir à renormaliser le signal de sortie : si le pas de synthèse est constant, le signal de synthèse est normalisé, alors que si le pas n'est pas constant, il faut renormaliser à l'aide de l'enveloppe constituée des somme des puissances des fenêtres, échantillon par échantillon. Le problème est que lorsque le facteur de contraction est grand, des portions du son original ne sont pas traitées avec ce schéma *cf.* sec. 7.1.2.

Pas d'analyse R_A constant Lorsque l'on utilise le schéma 2 avec pas d'analyse R_A constant, le temps d'analyse est donné par :

$$T_A^2(k) = T_A^2(k-1) + R_A \tag{5.12}$$

et le temps de synthèse en fonction de $\tilde{\gamma}$:

$$T_S^2(k) = T_S^2(k-1) + \tilde{\gamma}\left(T_A^2(k-1)\right) R_A \tag{5.13}$$

Remarquons que $T_S^2(0) = t_0 = T_A^2(0)$, d'où l'on obtient la formule récursive :

$$T_A^2(k) = t_0 + k\, R_A \tag{5.14}$$

$$T_S^2(k) = t_0 + R_A \sum_{l=1}^{k} \tilde{\gamma}\left(t_0 + lR_A\right) \tag{5.15}$$

Ce schéma est utilisé lorsque l'on veut s'assurer de traiter le son original dans son intégralité. Il nécessite par contre de renormaliser le signal de synthèse par une post-correction (*cf.* sec. 7.1.2).

5.3.3 Dilatation/contraction temporelle non-linéaire avec conservation de la hauteur et de la durée globale

Définition

Lors d'une dilatation/contraction temporelle non-linéaire hors temps-réel, il peut arriver que le signal obtenu soit de même durée que le signal original. Cependant, la plupart du temps, la durée du signal obtenu est différente, car le facteur de dilatation/contraction γ, $l = 1, \ldots, NT$ ne vérifie pas les bonnes conditions. Nous désirons pouvoir imposer au son traité d'avoir une durée prédéfinie (imposée par l'utilisateur), par exemple la durée du signal original. Une dilatation/contraction temporelle non linéaire avec conservation de la hauteur et de la **durée globale** consiste alors à appliquer un changement local de durée tout en imposant que la durée du son traité soit égale à la durée du son initial. C'est un raffinement de la dilatation/contraction temporelle non linéaire avec conservation de la hauteur, présentée précédemment (*cf.* sec. 5.3.1).

Fonctionnement

Hypothèses et expression de la synchronisation Le facteur de dilatation/contraction $\gamma(l)$ est calculé d'après les contraintes de l'utilisateur (choix des descripteurs et de leur conformations et combinaison). On lui applique ensuite une transformation de façon à ce que la durée du son obtenue soit la même que le son de départ : il s'agit d'une conformation spécifique à l'effet (*cf.* sec. 6.4). De manière à pouvoir utiliser les deux schémas, il nous faut connaître les conditions pour qu'ils soient équivalents, c'est-à-dire les conditions pour qu'à partir de la même courbe, donnent la même dilatation/contraction temporelle. Notons t_s le temps de synchronisation. La condition de synchronisation est donnée par :

$$T_A^i(\mu_1) = T_S^i(\mu_2) = t_s \tag{5.16}$$

Le premier schéma donne la condition suivante :

$$\frac{R_A}{R_S} \sum_{l=1}^{\mu_2} \tilde{\gamma}(t_0 + l\ R_A) = \mu_1 \tag{5.17}$$

et le second schéma donne :

$$\frac{R_S}{R_A} \sum_{l=1}^{\mu_1} \tilde{\gamma}(t_0 + l\ R_S) = \mu_2 \tag{5.18}$$

De plus, la condition de synchronisation (5.16) implique que :

$$\frac{\mu_1}{\mu_2} = \frac{R_A}{R_S} \tag{5.19}$$

La question que l'on se pose alors est la suivante : existe-t-il μ_0 vérifiant

$$\begin{cases} \mu_0 &= \frac{R_A}{R_S} \sum_{l=1}^{\sigma} \tilde{\gamma}(t_0 + l\ R_A) \\ \sigma &= \frac{R_S}{R_A} \sum_{l=1}^{\mu_0} \tilde{\gamma}(t_0 + l\ R_S) \end{cases} \tag{5.20}$$

Une hypothèse qui semble naturelle consiste à prendre un pas constant et commun aux deux schémas :

$$R_A = R_S = R_U \tag{5.21}$$

Cette hypothèse est équivalente à l'hypothèse selon laquelle il faut autant d'itérations de chaque schéma pour arriver au temps de synchronisation, ce qui peut encore s'écrire :

$$\mu_1 = \mu_2 = \mu \tag{5.22}$$

Sous cette hypothèse, la condition donnée par le système (5.20) se réduit à :

$$\mu = \sum_{l=1}^{\mu} \tilde{\gamma}(t_0 + l\ R_U) \tag{5.23}$$

Il est difficile de s'assurer que cette condition est vérifiée pour tout temps de synchronisation. En effet, μ doit être entier tout en vérifiant :

$$t_s = t_0 + \mu.R_U \tag{5.24}$$

ce qui peut se réécrire :

$$\mu = \frac{t_s - t_0}{R_U} \tag{5.25}$$

Si l'on peut se permettre de prendre n'importe quelle valeur pour le pas R_U (y compris réelle) sans problème de normalisation lors du traitement, alors la valeur de μ est donnée par l'équation (5.25). Sinon, si par exemple R_U doit être entier, diviseur de la taille du grain de traitement, alors on choisit le temps de synchronisation sur la grille des $t_0 + k.R_U$, de manière à toujours vérifier (5.24). Dans tous les cas, nous obtenons des conditions pour que (5.20) soit vérifiée. Nous pouvons donc chercher de quelle manière modifier la courbe γ et la transformer (par conformation) en $\tilde{\gamma}$.

Six solutions pour synchroniser Nous proposons trois manières de modifier la courbe γ : par addition, par multiplication et par puissance. Pour chacune, on peut décider de respecter les bornes de variation originales $\mathcal{I}_\gamma$ ou non, ce qui implique des schémas de calcul différents. Au total, nous avons donc six solutions pour imposer une durée au son traité. Nous présentons chaque méthode d'abord sans respect des bornes, puis avec respect des bornes. Par la suite, on notera $\gamma(l)$ la valeur $\gamma(t_0 + l\ R_U)$.

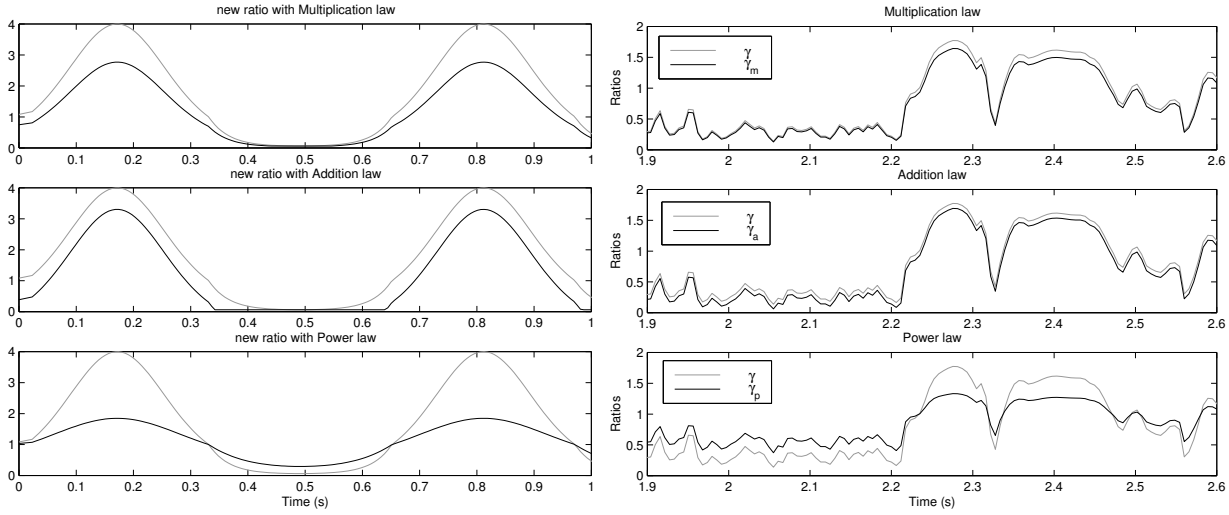


FIG. 5.17 – *Modification du facteur de dilatation/contraction temporelle par addition, multiplication, puissance pour un facteur de courbe sinusoïdale (à gauche) ou extraite d'un descripteur réel (à droite).*

Synchronisation par addition d'une constante γ_a aux valeurs de γ On désire trouver γ_a un facteur additif tel que la courbe $\tilde{\gamma} = \gamma + \gamma_a$ vérifie la condition (5.23). Ceci revient à translater la courbe vers le haut ou vers le bas. Si l'on ne tient pas compte de $I_\gamma = [\gamma_{min}; \gamma_{max}] = [\frac{1}{M}; N]$, l'intervalle de

variation de γ, le calcul est direct et donne :

$$\gamma_a = 1 - \frac{\sum_{l=1}^{\mu} \gamma(l)}{\mu} \tag{5.26}$$

En ne tenant pas compte de I_γ, on prend le risque d'avoir soit un facteur de dilatation/contraction trop grand (ce qui peut être accepté selon les critères esthétiques de l'utilisateur), soit un facteur de dilatation/contraction négatif. Ceci par contre ne peut pas être accepté, sinon cela signifierait de créer dans le son un attracteur (*cf.* Piste n°42-CD1 ♫♪ avec pour référence la Piste n°16 ♫♪), une zone pour laquelle la courbe de γ impose un déplacement tel qu'on resterait bloqué dans cette portion du temps. En effet, pour des valeurs négatives du facteur de dilatation/contraction, le son est lu à rebours. La courbe du paramètre de contrôle est alors elle aussi lue à rebours, jusqu'à une valeur de contrôle positive qui fera qu'on lit le son à l'endroit, jusqu'à une valeur de contrôle négative où l'on va à nouveau partir à rebours, et ainsi de suite. Pour éviter ce désagrément, on utilise une autre méthode de recherche de γ_a, en imposant $\tilde{\gamma}(l) \in I_\gamma,\ l = 1, \ldots, \mu$. Nous utilisons la fonction de troncature (*cf.* sec. 6.3.2) :

$$\mathcal{H}(f(t), I_f) = \begin{cases} f(t) & \text{si } f(t) \in I_f \\ f_{min} & \text{si } f(t) < f_{min} \\ f_{max} & \text{si } f(t) > f_{max} \end{cases} \tag{5.27}$$

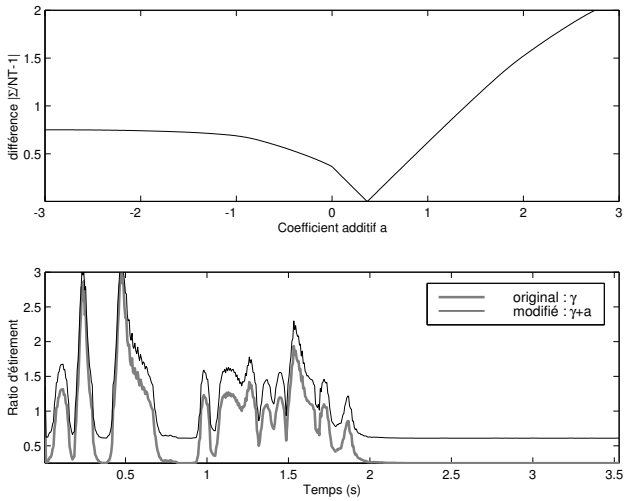


FIG. 5.18 – *Recherche du facteur additif γ_a permettant une synchronisation en respectant $\mathcal{I}_\gamma$.*

Nous calculons D la valeur absolue de la différence de durée normalisée (qui représente l'écart à la condition de synchronisation)

$$D(a) = \left| \frac{\sum_{l=1}^{\mu} \mathcal{H}\left(\gamma(l) + \gamma_a, I_\gamma\right)}{\mu} - 1 \right| \tag{5.28}$$

et ceci pour plusieurs valeurs de $a \in [-\gamma_{max}; \gamma_{max}]$. Nous cherchons ensuite $\overline{\gamma}_a$ la valeur de a qui minimise cette fonction : c'est le facteur additif optimum pour conserver la durée tout en respectant les bornes.

Synchronisation par multiplication des valeurs de γ par γ_m On désire trouver γ_m un facteur multiplicatif tel que la courbe $\tilde{\gamma} = \gamma \times \gamma_m$ vérifie la condition (5.23). Si l'on ne tient pas compte de I_γ (intervalle de variation de γ), le calcul est direct et donne :

$$\gamma_m = \frac{\mu}{\sum_{l=1}^{\mu} \gamma(l)} \tag{5.29}$$

Ceci revient à calculer la durée du son étiré par la courbe $\tilde{\gamma}$, puis à multiplier γ par le rapport γ_m de la durée initiale sur la durée calculée. On remarquera qu'une fois encore, si l'on ne respecte pas l'intervalle de variations I_γ, les risques sont identiques à ceux existant dans le cas de la translation. Pour éviter ces désagréments, on utilise une méthode identique au cas précédent. Nous calculons D la valeur absolue de la différence de durée normalisée

$$D(m) = \left| \frac{\sum_{l=1}^{\mu} \mathcal{H}\left(m\gamma(l), I_\gamma\right)}{\mu} - 1 \right| \tag{5.30}$$

pour plusieurs valeurs de $m \in [-\gamma_{max}; \gamma_{max}]$. Nous cherchons ensuite $\overline{\gamma}_m$ la valeur de m qui minimise cette fonction (*cf. fig.* 5.19) : c'est le facteur multiplicatif optimum pour conserver la durée tout en respectant les bornes.

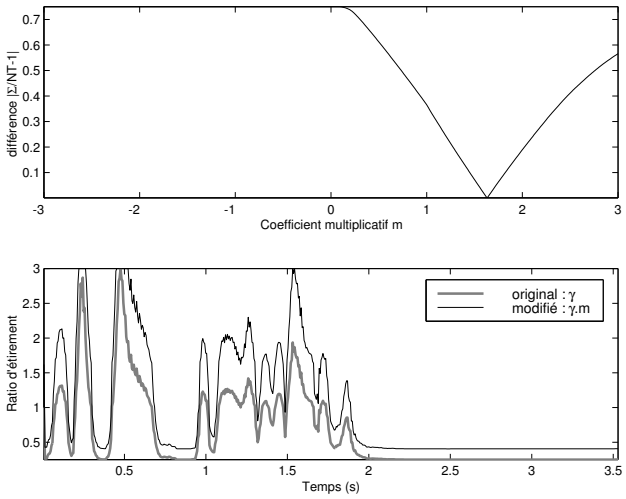


FIG. 5.19 – *Recherche du facteur multiplicatif permettant une synchronisation en respectant $\mathcal{I}_\gamma$.*

Synchronisation par mise à la puissance p des valeurs de γ On désire trouver une puissance $p \in [0; 2]$ tel que la courbe $\tilde{\gamma} = \gamma^p$ vérifie la condition (5.23). Cette condition de synchronisation s'écrit alors :

$$\mu = \sum_{l=1}^{\mu} [\gamma(l)]^p \tag{5.31}$$

On ne peut trouver cette valeur de la puissance p de manière analytique. Aussi, on calcule pour plusieurs valeurs de $p \in [0; 2]$ la valeur absolue de la différence

$$D(p) = \frac{\sum_{l=1}^{\mu} [\gamma(l)]^p}{\mu} - 1 \tag{5.32}$$

qui représente l'écart à la condition de synchronisation. Cette courbe présente un minimum en $p = 0$, mais cette valeur ne nous intéresse pas : elle signifie que l'on remplace la courbe γ par une valeur constante égale à 1, le son n'est alors plus modifié (et le contrôle n'est plus adaptatif) ! On va donc cherche le minimum local dont l'abscisse (la valeur p) est la plus grande. Ainsi, on s'assure de prendre l'exposant le plus grand. En règle générale, on trouve facilement une puissance non nulle qui permet de vérifier la condition de synchronisation, *cf. fig.* 5.20 gauche.

Si l'on désire respecter les bornes de départ, il faut alors utiliser la différence D définie par :

$$D_{tr}(p) = \left| \frac{\sum_{l=1}^{\mu} \mathcal{H}_{trunc}([\gamma(l)]^p, \gamma_{min}, \gamma_{max})}{\mu} - 1 \right| \tag{5.33}$$

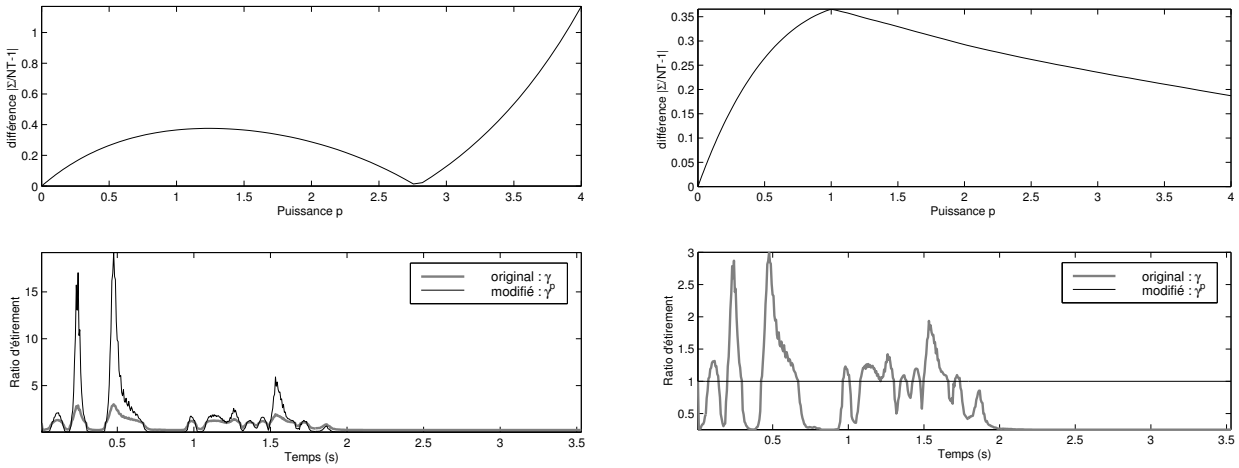


FIG. 5.20 – *Recherche de la plus grande puissance permettant une synchronisation sans respecter $\mathcal{I}_\gamma$, à gauche (fonction de différence en haut ; courbes γ et sa version modifiée $\tilde{\gamma}$ avec $p = 2.49$ en bas), et en respectant $\mathcal{I}_\gamma$, à droite (pour une courbe mal équilibrée autour de 1, la seule puissance valide est très souvent $p = 0$).*

Comme on peut le voir *fig.* 5.20 droite, dès que la courbe n'est pas bien équilibrée autour de la valeur 1, la seule puissance permettant de respecter la condition de synchronisme est $p = 0$. Pour cette valeur, le son n'est ni dilaté ni contracté ; aussi, pour s'assurer de pouvoir conserver la durée globale, il est intéressant de prendre une courbe ayant à peu près autant de valeurs supérieures à 1 qu'inférieures à 1. La solution que nous avons développé consiste dans ce cas à utiliser une autre méthode, utilisant deux puissances, et une séparation de la courbe γ en trois composantes :

i) $\gamma_{inf} = \gamma \mathbb{1}_{\gamma<1}$ ne comportant que les valeurs de γ inférieures strictement à 1 ;
ii) $\gamma_{sup} = \gamma \mathbb{1}_{\gamma>1}$ ne comportant que les valeurs de γ supérieures strictement à 1 ;
iii) $\gamma_{egal} = \gamma \mathbb{1}_{\gamma=1}$ ne comportant que les valeurs de γ strictement égales à 1.

Il faut alors rechercher p_{inf} et p_{sup} tels que :

$$\sum_{l=1}^{\mu} [\gamma_{inf}(l)]^{p_{inf}} + [\gamma_{sup}(l)]^{p_{sup}} + \gamma_{egal}(l) = \mu \tag{5.34}$$

La manière de procéder est une généralisation de la méthode à une puissance : on calcule une fonction $D(p_{inf}, p_{sup})$ à deux dimensions qui correspond à la valeur absolue de la différence normalisée de synchronisation :

$$\left| D(p_{inf}, p_{sup}) = \frac{\sum_{l=1}^{\mu} [\gamma_{inf}(l)]^{p_{inf}} + [\gamma_{sup}(l)]^{p_{sup}} + \gamma_{egal}(l)}{\mu} - 1 \right| \tag{5.35}$$

Cette fonction des deux puissances (*cf. fig.* 5.22 gauche) est calculée à partir de la courbe de contrôle *fig.* 5.21. Remarquons que nous avons utilisé une courbe γ légèrement différente du cas de la fonction à une seule puissance, afin de mieux illustrer la spécificité de cette fonction à deux puissances (et des courbes qui y sont associées).

Nous allons extraire ce que nous appelons la "vallée" de cette fonction de deux variables, à savoir la ligne des minima de $D(,)$, qui apparaît sur la figure *fig.* 5.22 droite (notons que cette courbe est toujours monotone). Pour ce faire, nous allons choisir l'un des axes, puis pour chaque valeur selon cet axe, chercher le minimum selon l'autre axe. Nous effectuons cette démarche pour chaque axe, et obtenons donc deux vallées de minima possibles, qui ne sont identiques que lorsque la pente de la courbe est 1. Dans tous les autres cas (autrement dit quasiment à chaque fois), l'un des deux courbes de vallée est plus lisse que l'autre, du fait que la pente est supérieure ou inférieure à 1. Pour

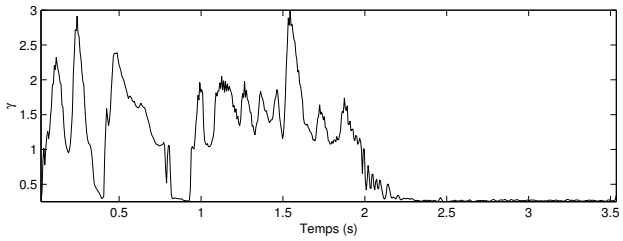


FIG. 5.21 – *Courbe de contrôle de la dilatation/contraction temporelle, avant conformation.*

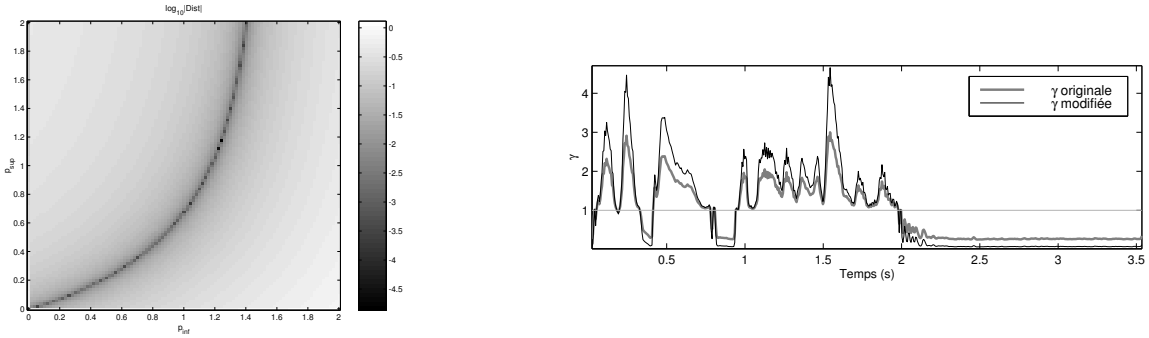


FIG. 5.22 – *A gauche : logarithme de la fonction de distance* $\log_{10}(Dist)$ *en fonction de* p_{inf} *et de* p_{sup}. *A droite : courbe* $\tilde{\gamma}$ *modifié par fonction puissance à deux coefficients* (p_{inf}, p_{sup}), *sans respecter* $\mathcal{I}_\gamma$, *avec* $p_{inf} = 1.84$ *et* $p_{sup} = 1.38$.

chaque courbe de vallée, on calcule la mesure N pour tous les couples (p_{inf}, p_{sup}) (hors extrémités) de chacune des deux courbes de vallée :

$$N = \sqrt{\sum_{l=1}^{\mu} \left([\gamma_{inf}(l)]^{p_{inf}} - 1\right)^2 + \sum_{l=1}^{\mu} \left([\gamma_{sup}(l)]^{p_{sup}} - 1\right)^2} \tag{5.36}$$

Cette mesure correspond à la distance de la courbe $\gamma(p_{inf}, p_{sup})$ à la droite $y = 1$. Puisque nous cherchons les valeurs d'exposants p_{inf} et p_{sup} les plus éloignées de 0, on conserve le couple (p_{inf}, p_{sup}) qui maximise la norme donnée en (5.36). On obtient alors une courbe $\tilde{\gamma}$ modifiée (*cf. fig.* 5.22 droite). Toute cette démarche peut aussi bien s'effectuer en choisissant au départ de respecter les bornes de variation de γ. L'expression de la fonction de différence utilise une fois de plus la fonction de troncature :

$$Dist(p_{inf}, p_{sup}) = \left| \frac{\sum_{l=1}^{\mu} \mathcal{H}_{trunc}\left([\gamma_{inf}(l)]^{p_{inf}}, I_\gamma\right) + \mathcal{H}_{trunc}\left([\gamma_{sup}(l)]^{p_{sup}}, I_\gamma\right) + \gamma_{egal}(l)}{\mu} - 1 \right| \tag{5.37}$$

On notera que la fonction de différence est d'aspect différent lorsque l'on respecte les bornes (*cf. fig.* 5.23 gauche) et lorsqu'on ne les respecte pas (*cf. fig.* 5.22 gauche). La courbe de vallée est moins régulière tout en restant monotone. Les irrégularités apparaissent aux valeurs de p_{inf} et de p_{sup} pour lesquelles la troncature entre en jeu. La courbe de $\tilde{\gamma}$ modifiée est bien évidemment différente, comme l'atteste la *fig.* 5.23.

Choix des valeurs de p**,** p_{inf}**,** p_{sup} L'intervalle imposé aux exposants est $[0; 2]$. Il est totalement arbitraire : on peut vouloir utiliser des puissances plus grandes que 2. Des valeurs négatives n'auraient pas de sens dans ce contexte, car elles modifieraient la courbe γ d'une manière contraire aux exigences de l'utilisateur, en faisant passer les valeurs puisque l'on veut pouvoir passer à des exposants

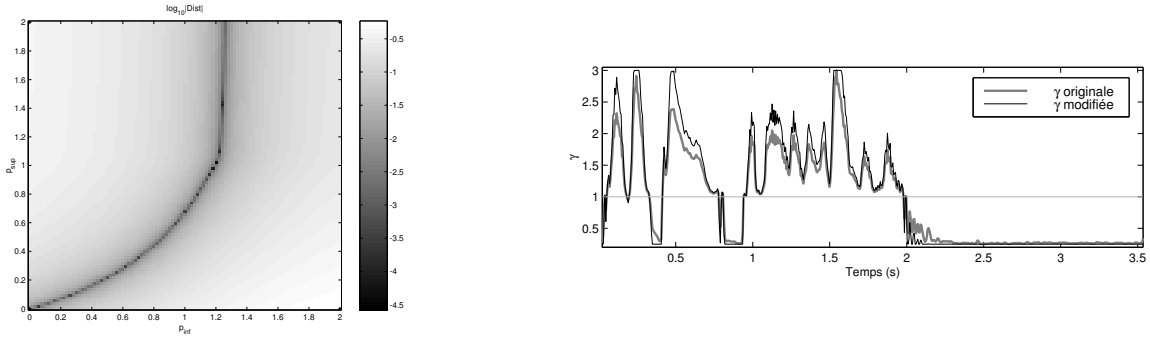


FIG. 5.23 – *A gauche : logarithme de la fonction de distance* $\log_{10}(Dist)$ *en fonction de* p_{inf} *et de* p_{sup}. *A droite : courbe* $\tilde{\gamma}$ *modifié par fonction puissance à deux coefficients* (p_{inf}, p_{sup}), *en respectant* $\mathcal{I}_\gamma$, *avec* $p_{inf} = 1.42$ *et* $p_{sup} = 1.24$.

positifs et croissants, afin que la modification de $\tilde{\gamma}(l)$ soit monotone lorsque la puissance p croît. Parmi les valeurs positives, nous avons par l'expérience trouvé que l'intervalle $[0;2]$ était pratique, d'abord parce qu'il ne contient que des valeurs positives, ensuite parce qu'il contient le 0, qui permet de remplacer la courbe $\gamma(l)$ par une constante $\tilde{\gamma}(l) = 1$, enfin parce qu'il permet des modifications de type racine $\sqrt[\alpha]{x}, \alpha \in [0;+\infty[$ aussi bien que des modifications de type puissance carré, ainsi que les valeurs intermédiaires dans $[1;2]$.

Intérêt de cet effet En utilisant cette notion de synchronisation ou de conservation de durée, on peut obtenir des sons de même durée, mais dont l'expressivité diffère. En effet, pour une voix, les voyelles et les consonnes peuvent être traitées différemment en contrôlant l'effet par l'indice de voisement par exemple (*cf.* Piste n°43-CD1 ♫♪ d'après le son de référence Piste n°16 ♫♪, Piste n°44-CD1 ♫♪ d'après le son de référence Piste n°18 ♫♪). En utilisant l'un ou l'autre des lois de modifications, les expressivités changent. Ceci reste vrai pour les sons instrumentaux. Ainsi, l'introduction au saxophone de Sylvain Beuf Piste n°3 ♫♪ peut être rendu hors-temps de différentes manières : Piste n°45-CD1 ♫♪ pour le première modification seule, et Piste n°46-CD1 ♫♪ en stéréophonique avec le son original ; Piste n°47-CD1 ♫♪ pour la seconde modification, en stéréphonique avec le son original Piste n°48-CD1 ♫♪. On peut donc modifier l'expressivité d'une voix, mais aussi changer le rythme et provoquer des arythmies sur un son auparavant rythmé. Remarquons que si on combine cet effet avec une transposition adaptative (*cf.* sec. 5.4.2), on peut réaliser un changement de prosodie ; ce sujet sera traité en détails en sec. 5.7.5.

5.3.4 Dilatation/contraction temporelle non-linéaire préservant l'attaque, le vibrato

Le problème de la conservation d'attributs perceptifs tels que l'attaque des sons lors de la dilatation/contraction est classique, et notamment utilisé pour conserver l'intelligibilité de la parole [Arons, 1992]. Plusieurs solutions ont été proposées récemment, certaines utilisant des méthodes temporelles [Pallone, 2003], d'autres des méthodes spectrales [Amatriain *et al.*, 2003; Verma *et al.*, 1997], d'autres encore utilisant les modèles temps-fréquence [Arfib and Zoelzer, 2002b; Verfaille and Arfib, 2001]. Dans tous les cas, la méthode est adaptative et consiste à traiter différemment les attaques des portions stables des notes. On utilise les descripteurs de l'aspect transitoire du son (grandes variations du contenu en hautes fréquences, du flux spectral, de l'harmonicité, de l'énergie).

D'autre part, le vibrato pose lui aussi des problèmes lors de la dilatation/contraction temporelle. Une solution consiste à le supprimer par une transposition inverse, avant de procéder à la dilatation/contraction temporelle. Le vibrato peut alors être à nouveau appliqué sur les portions de son où il a été effacé [Arfib and Delprat, 1998].

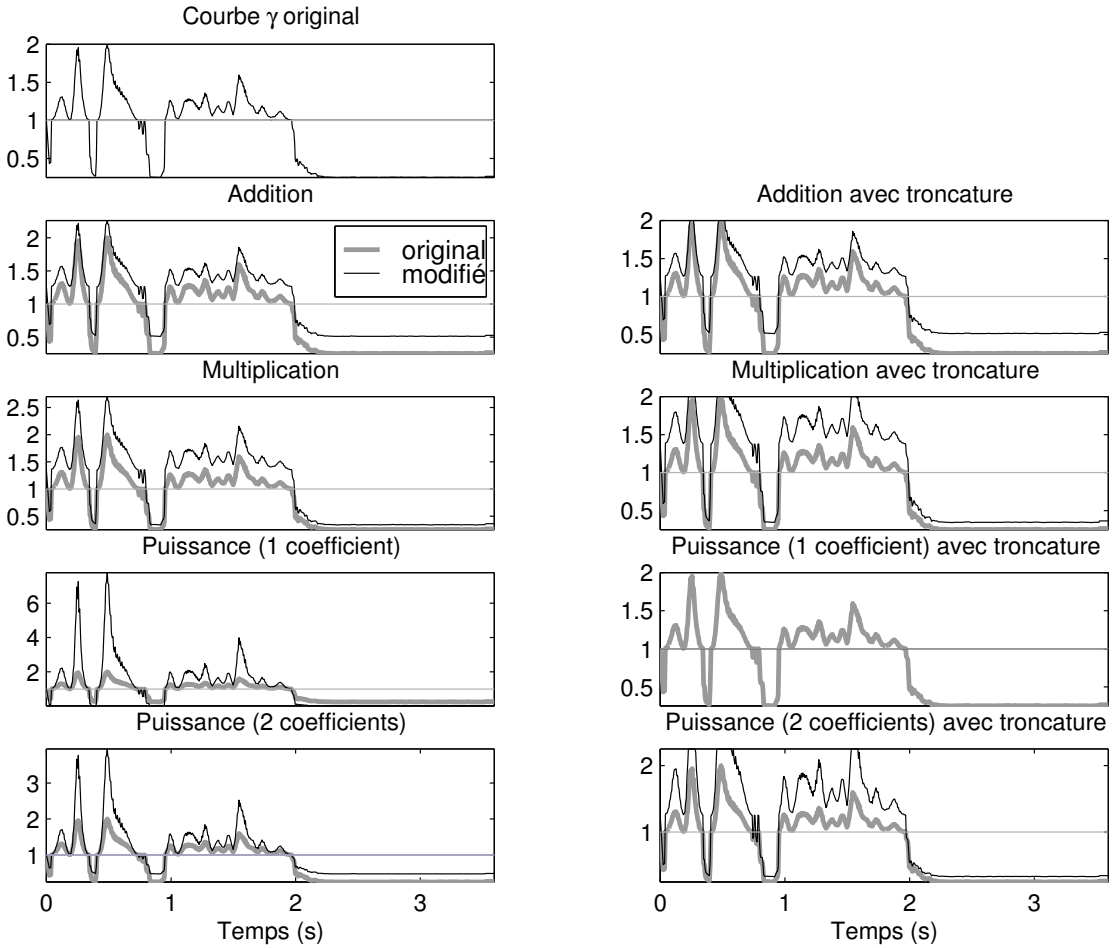


FIG. 5.24 – *Comparatif des modifications du facteur de dilatation/contraction $\tilde{\gamma}$ par addition, multiplication, puissance à 1 ou 2 coefficients (de haut en bas), sans respecter (colonne de gauche) ou en respectant $\mathcal{I}_\gamma$ (colonne de droite). Remarquons que la courbe de contrôle originale γ est encore différente ; elle a été choisie de manière à bien illustrer les différences entre les méthodes de synchronisations.*

5.3.5 Contrôle gestuel de la dilatation/contraction temporelle non-linéaire

Etant donnée une courbe de facteur de dilatation/contraction $\gamma(t)$, on propose de la modifier par le geste, donné par la valeur $g(t)$. On envisage plusieurs manières de le faire :

1. le contrôle $g(t)$ sert à modifier $\gamma(t)$ par une loi de puissance : $\tilde{\gamma}(t) = \gamma(t)^{g(t)}$ avec $g(t) > 0$. Dans ce cas, on risque de perdre la synchronisation si jamais il y en avait une. En effet, $g(t)$ varie dans le temps ; on ne se place donc plus dans le cadre des résultats de la synchronisation, pour lesquels les facteurs additif, multiplicatif ou puissance sont constants.
2. le contrôle $g(t)$ sert à modifier $\gamma(t)$ par addition : $\tilde{\gamma}(t) = \gamma(t) + g(t)$ avec $g(t) > 0$. Dans ce cas, on risque de perdre la synchronisation si jamais il y en avait une.
3. le contrôle $g(t)$ sert à modifier $\gamma(t)$ par addition d'une courbe qui préserve la synchronisation : $\tilde{\gamma}(t) = \gamma(t) + s(t)$ avec $s(t) = g(periode) \sin(2\pi T_{sync}(t))$ périodique et de somme nulle sur une période, vérifiant $g(t) + \gamma(t) > 0$. A chaque nouvelle période, la période de la sinusoïde est donnée par la période de synchronisation T_{sync}, et l'amplitude de la sinusoïde est donnée par $g(t)$;
4. idem avec une courbe autre que sin, donnée par une table d'amplitude (*cf.* sec. 5.2.5) ;

5. le contrôle $g(t)$ donne le facteur de dilatation/contraction globale, et correspond donc à une modification de $\gamma(t)$ par multiplication : $\tilde{\gamma}(t) = \gamma(t)g(t)$ avec $g(t) > 0$. Le changement d'échelle du son est alors non-linéaire par le son, et à nouveau modifié non linéairement par le geste.

Nous allons montrer analytiquement que pour une fonction périodique de somme nulle sur une période (par exemple respectant une symétrie centrale par rapport au milieu du segment d'une période), la modification de $\gamma(t)$ par ajout respecte toujours le temps de synchronisation, et donc la durée initiale de la portion de son. Partons de l'équation de synchronisation *eq.* (5.23). Si elle est vérifiée pour $\tilde{\gamma}(t)$, on aura :

$$\sum_{l=1}^{s} \tilde{\gamma}(t_0 + l\ R_U) = s \tag{5.38}$$

ce qui équivaut encore à :

$$\sum_{l=1}^{s} (\gamma(t_0 + l\ R_U) + g(t_0 + l\ R_U)) = s \tag{5.39}$$

d'où l'on extrait :

$$\sum_{l=1}^{s} g(t_0 + l\ R_U) = 0 \tag{5.40}$$

Cette somme se fait sur une seule période, ce qui signifie que la fonction $g(t)$ peut changer à chaque période, tant qu'elle vérifie la condition (5.40), qui est nécessaire et suffisante. Pour une fonction $g(t)$ sinusoïdale, sa fréquence est alors donnée par $\frac{1}{s\ R_U}$.

5.3.6 Effet de changement de *groove*

Etant donné un signal musical possédant un rythme et des marqueurs du rythme, on peut associer à certains marqueurs un point de synchronisation. Un exemple est donné *fig.* 5.25, avec des points de synchronisation donnés par l'utilisateur.

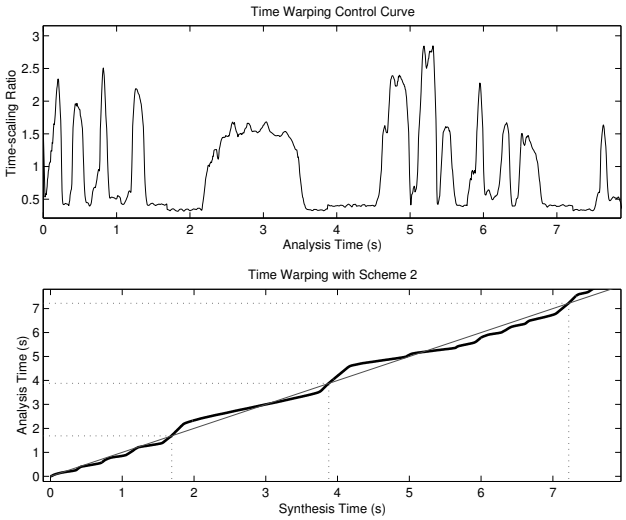


FIG. 5.25 – *Courbe de contrôle du changement d'échelle adaptatif (facteur de dilatation/contraction,* fig. *sup.) et temps de synthèse $T_S^2(k)$ en fonction du temps d'analyse $T_A^2(k)$ (*fig. *inf.). La courbe $T_S = T_A$ est donnée en gris, et les points de synchronisation par les lignes pointillées.*

Intérêt de cet effet Cet effet est un raffinement de la dilatation/contraction temporelle non linéaire avec préservation de la durée globale, qui permet par exemple d'éviter les effets d'arythmie ou d'obtenir un effet chorus.

Pour éviter les effets d'arythmie, on place les temps de synchronisation sur une sous-pulsation du rythme initial. Par exemple, pour un rythme binaire basé sur une mesure à 4 temps, on peut placer des points de synchronisation au premier temps de chaque mesure, ou aux premier et troisième temps, voire aux deuxième et quatrième temps. Ainsi, l'échelle temporelle est localement modifiée, mais des aspects du rythme sont conservés. Dans le cas où on place un point de synchronisation sur chaque attaque de note, ce n'est plus le rythme mais chaque note qui est localement modifiée, ce qui peut modifier finement le timbre.

Avec des points de synchronisation placés de manière moins stricte, on obtient un effet de chorus, où plusieurs locuteurs prononcent la même phrase différemment. Cet effet est plus naturel que par les méthodes de modulation de ligne à retard usuelles, du fait que les décalages temporels sont audibles, mais surtout n'ont pas l'aspect systématique habituel. Ainsi, en imposant un point de synchronisation toutes les 2 secondes et en utilisant des contrôles différents (choix des descripteurs et de leurs conformations différents, mais aussi type de synchronisation entre chaque couple de points de synchronisation différents), on obtient des sons distincts. Une première série d'exemples est donnée avec la voix de Pierre Schaeffer Piste n°49-CD1 ♫♪ et Piste n°51-CD1 ♫♪ (avec pour référence la appliquée à la Piste n°18 ♫♪), où la différence d'expressivité est bien révélée, surtout lorsqu'on les entend en même temps que le son original (Piste n°50-CD1 ♫♪ et Piste n°52-CD1 ♫♪). L'effet a été exagéré volontairement dans les exemples. Une deuxième série d'exemples est obtenu avec le saxophone de Sylvain Beuf Piste n°53-CD1 ♫♪ et Piste n°55-CD1 ♫♪ (avec les versions stéréophoniques pour comparer avec le son original Piste n°54-CD1 ♫♪ et Piste n°56-CD1 ♫♪), d'après la référence Piste n°3 ♫♪. Le fait d'imposer des points de synchronisation permet d'avoir un flux temporel de son globalement préservé, mais au prix de quelques sacrifices : on remarque dans les exemples sonores que l'expressivité sur chaque segment temporel peut différer d'un segment à l'autre.

5.4 Effets adaptatifs sur la hauteur

Nous avons présenté quelques effets adaptatifs sur la hauteur dans la bibliographie : l'harmoniseur (sec. 3.4.5), la discrétisation sur une échelle prédéfinie (*cf.* sec. 3.4.4). Nous présentons maintenant d'autres effets adaptatifs modifiant la hauteur, que nous avons développé. Ils sont tous basés sur la transposition adaptative. Nous différencions la transposition adaptative du changement de hauteur, où c'est la hauteur et non le facteur de transposition qui est donnée par une courbe de contrôle. Nous proposons ensuite une modification de la discrétisation sur une échelle de hauteur prédéfinie. Enfin, nous proposons un harmoniseur/inharmoniseur adaptatif.

Pour chacune de ces modifications de la hauteur, on peut utiliser le vocodeur de phase *cf.* sec. 2.4.2, TD-PSOLA *cf.* sec. 2.4.1 ou un modèle additif *cf.* sec.2.4.3. Nous avons utilisé une mise en œuvre sous Matlab [Arfib and Zoelzer, 2002b] de la transposition par vocodeur de phase. Nous l'avons choisi parce qu'elle utilise la FFT et est donc rapide ; de plus, elle ne dépend pas d'une détection de hauteur, contrairement à TD-PSOLA et au modèle additif (même si, pour des sons harmoniques, une détection de hauteur permet de mieux régler la quéfrence de coupure de façon à bien extraire l'enveloppe spectrale). Nous avons aussi utilisé le modèle additif pour certains effets, tels que la discrétisation sur une échelle prédéfinie.

Dans un premier temps, nous allons montrer les modifications de mise en œuvre auxquelles il faut procéder pour que la transposition à facteur variable avec le vocodeur de phase soit possible sans ajouter de nouveaux artefacts, puis nous présenterons les effets adaptatifs portant sur la hauteur.

5.4.1 Transposition adaptative avec le vocodeur de phase

La transposition par vocodeur de phase utilise à la fois le modèle source-filtre pour préserver les formants, et le modèle temporel afin de ré-échantillonner la source et donc de la transposer (*cf.* sec. 3.4.3). Sachant que ce ré-échantillonnage fait varier la taille du grain de resynthèse selon le facteur de transposition, on s'attend à ce que différents artefacts apparaissent dans le son, dû au traitement. Nous allons corriger un à un ces artefacts. On synthétise le son transformé par ajout de grains à pas constant et de tailles variables, l'enveloppe du traitement ne peut pas être constante. Le son traité dont alors être normalisé par une post-correction (*cf.* sec. 7.1.2). Dans ce cas, étant donné un facteur minimal γ_m et un facteur maximal γ_M de multiplication de la fréquence fondamentale f_0, le pas R_U utilisé à l'analyse et à la synthèse doit être calculé de manière à ce que le recouvrement soit bon (*cf. fig.* 5.26), c'est-à-dire que l'amplitude de l'enveloppe soit toujours supérieure ou égale à 1, *cf.* sec. 2.4.2.

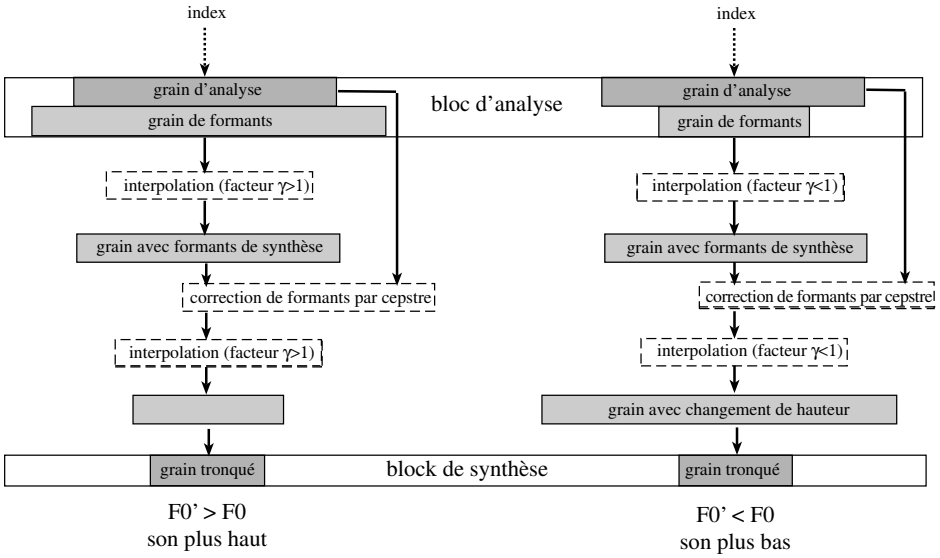


FIG. 5.26 – *Exemples de transposition adaptative par le vocodeur de phase : les rectangles gris représentent les grains, depuis le grain d'analyse jusqu'au grain de synthèse (avec un taille proportionnelle à la taille du grain). Les boîtes en pointillé sont les opérations appliquées aux grains.*

Ensuite, lors de la transposition avec préservation de l'enveloppe spectrale, le principe est de rééchantillonner la source (pour réaliser la transposition), avant de lui appliquer le filtre (pour préserver l'enveloppe spectrale). Le ré-échantillonnage par interpolation doit être de qualité suffisante, que le facteur de transposition soit une fraction entière ou non. Enfin, il faut s'assurer que le temps de référence de chaque grain avant transposition et ajout-superposition soit le centre de ce grain. En effet, lorsque le facteur de transposition ne change pas, on peut utiliser comme référence le début ou le centre du grain : cela ne pose pas de problème de normalisation du son traité, même si la valeur de contrôle utilisée est décalée. Par contre, lorsque le facteur change, il faut que le temps de référence soit le centre du grain, sinon le grain de synthèse sera décalé, puisque plus court ou plus long que le grain initial, mais de même temps de référence (*cf. fig.* 5.27). Ceci sera développé en sec. 7.1.2.

5.4.2 Transposition adaptative (TR, NTR)

La **transposition adaptative** consiste à transposer d'un facteur donné par une courbe de contrôle $\mathcal{C}(t)$, en respectant les formants. Si le son est harmonique de fréquence fondamentale $\mathcal{H}_0$, la nouvelle

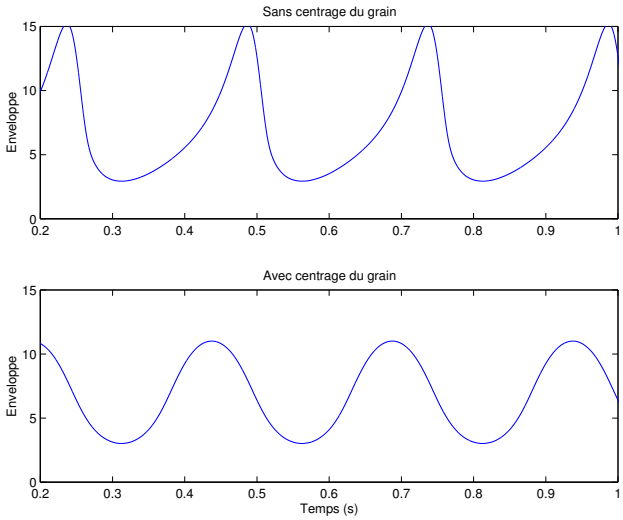


FIG. 5.27 – *Enveloppe d'amplitude pour un changement de hauteur contrôlé par un facteur* $2^{\sin(2\ \pi\ 4\ t)}$, *sans centrage du grain (*fig. *haut) et avec centrage du grain (*fig. *bas).*

hauteur $\mathcal{P}(t)$ du son est alors :

$$\mathcal{P}(t) = \mathcal{C}(t).\mathcal{H}_0(t) \tag{5.41}$$

Tous les types de sons ne se prêtent pas à cet effet. La voix parlée (*cf.* Piste n°57-CD1 ♫♪ et Piste n°58-CD1 ♫♪, d'après la référence Piste n°10 ♫♪) se prête bien à cet effet, alors que le résultat est assez indigeste pour les sons instrumentaux (*cf.* Piste n°59-CD1 ♫♪, d'après la Piste n°3 ♫♪) dès lors qu'on ne prend pas en compte des données de segmentation afin de transposer d'un facteur constant sur chaque note. Pour les sons électroacoustiques, cet effet apporte des possibilités de variations intéressantes, mais il est souvent préférable de ne pas conserver les formants, du fait que l'enveloppe spectrale de ces sons est très différente de celle des sons musicaux.

Une contrainte non négligeable existe sur la courbe de contrôle. Des variations de fréquence entre 4 et 8 *Hz* et d'une profondeur d'un demi-ton seront perçues comme un vibrato, ce qui n'est pas un problème en soi si c'est l'effet voulu. Par contre, des variations au dessus de 20 *Hz* seront trop rapides, et donneront des modifications du timbre n'ayant plus grand chose à voir avec un changement de hauteur, concernant la perception. Cependant, rappelons que le modèle additif et le vocodeur de phase utilisent des fenêtres d'analyse et de synthèse, de tailles et de pas donnés et fonctionnent par ajout-superposition. Généralement, la taille de fenêtre varie entre 512 et 2048 échantillons, et le pas entre 1/8 et 1/2 taille de fenêtre, soit des pas allant de 64 à 1024 échantillons. Ceci signifie que l'échantillonnage de la valeur de contrôle varie entre une valeur allant de 43 *Hz* (le plus grand pas : 1024) à 689 *Hz* (le plus petit pas : 64). Ces fréquences sont dans tous les cas au-dessus de 20 *Hz*, un filtrage passe-bas est nécessaire à la fois pour éviter le repliement, mais aussi pour éviter d'atteindre les 20 *Hz*, tout simplement.

5.4.3 Changement de hauteur adaptative (NTR)

Lorsqu'on veut remplacer la hauteur par une courbe prédéfinie, nous préférons utiliser l'expression "changement de hauteur" plutôt que de transposition. Le contrôle est en effet différent : dans le cas de la transposition, on indique le facteur de transposition alors que dans le cas du changement de hauteur, on indique la hauteur cible (il faut alors calculer le facteur de transposition correspondant). Le **changement de hauteur** adaptatif consiste donc à remplacer la hauteur originale par la courbe de contrôle $\mathcal{C}(t)$, et s'obtient par une transposition de facteur :

$$\rho(t) = \frac{\mathcal{C}(t)}{F_0(t)} \tag{5.42}$$

où $\overline{F_0}(t)$ est la fréquence fondamentale lorsqu'elle existe. Le plus simple lorsque la fréquence fondamentale n'existe pas est de ne pas transposer. Cependant, lorsqu'il y a une co-articulation entre voyelles et consonnes par exemple, il convient de transposer tout de même sur les passages non voisés, de manière à ce que les transitions soient plus souples. Dans ce cas, on utilise une valeur correspondant à la moyenne de F_0 sur les segments de son où la hauteur existe.

5.4.4 Discrétisation sur une échelle prédéfinie

La discrétisation sur une échelle prédéfinie (*cf.* sec. 5.4.4) sert habituellement à accorder un signal par rapport à une échelle prédéfinie. Lorsqu'il fonctionne en temps-réel, cet effet (l'*autotune*) est assez difficile à régler, et ne fonctionne pas toujours bien, du fait qu'il y ait une latence entre la détection de la fréquence fondamentale et sa correction, ce qui provoque des glissés de fréquences notamment aux attaques. Hors temps-réel, cet artefact est tout à fait évitable. Cette échelle est habituellement supposée fixe [Amatriain *et al.*, 2001]. Nous proposons que cette échelle soit variable, et que ce soit les descripteurs du son qui servent à passer d'une échelle à l'autre, de manière discrète ou continue. Notons que cela pourrait tout aussi bien être le geste qui contrôle le glissement d'une échelle à l'autre. Une seconde amélioration consiste à utiliser le modèle additif et à imposer à chaque harmonique à respecter des critères de variabilités. Ainsi, en imposant des valeurs au *jitter* et au *shimmer*, on s'attaque directement à la texture du son, et donc au timbre. Nous n'avons hélas pas suffisamment avancé dans cette direction pour proposer des solutions techniques précises et des exemples sonores.

5.4.5 Harmoniseur/inharmoniseur adaptative (NRT)

L'harmoniseur présenté en 3.4.5 est conçu pour de la musique tonale, et ajoute des versions transposées du signal de façon à respecter une tonalité (donnée par la fréquence fondamentale et le contexte d'harmonie dans lequel la note est jouée). Les facteurs de transposition sont donc discrétisés. Nous proposons d'utiliser des facteurs de transposition non discrétisés, afin de réaliser des glissements continus de la tierce, par exemple, et garder une quinte juste.

Ainsi, dans la première série d'exemple, la Piste n°12 ♫♪ est transposée à la quinte juste ainsi que selon un facteur qui passe de manière continue d'une tierce mineure à une tierce majeure (en fonction de la courbe de contrôle), ce qui donne un effet bizarre, déstabilisant où le son passe de l'accord majeur à l'accord mineur de manière continue (*cf. fig.* 5.28 à gauche et Piste n°60-CD1 ♫♪). Tout aussi dérangeant, l'effet présenté *fig.* 5.28 à droite et Piste n°61-CD1 ♫♪) a pour accords extrêmes l'accord majeur pour la valeur nulle de la courbe de contrôle et la double octaviation (ajout de l'octave supérieure et de l'octave inférieure) pour la valeur unité de la courbe de contrôle. Dans cette configuration, de plus, les trajets des versions transposées car la tierce majeure est associée à l'octave inférieure et la quinte juste à l'octave supérieure.

Une deuxième série d'exemples d'harmonisation adaptative a été réalisée avec le son Piste n°7 ♫♪. La Piste n°62-CD1 ♫♪ correspond à une harmonisation non adaptative à l'accord de septième majeur. On peut aussi appliquer une transposition adaptative au son, puis une harmonisation non adaptative. Ainsi, la Piste n°64-CD1 ♫♪ correspond à une harmonisation en accord de majeur septième, à laquelle s'ajoute une transposition adaptative allant jusqu'à moins un ton, conduite par le RMS. La Piste n°63-CD1 ♫♪ est une harmonisation adaptative passant de l'accord mineur sixième à l'accord de majeur septième en fonction du RMS, en ajoutant une transposition adaptative allant jusqu'à moins un ton, conduite par le taux de basses énergies.

Comme on le voit par ces deux séries d'exemples simples, il est possible de développer une grande quantité de stratégies différentes, en donnant par exemple les accords correspondant aux valeurs extrêmes de la courbe de contrôle ainsi que la manière de combiner les trajectoires des composantes de l'accord. On peut aussi donner une courbe de contrôle différente à chacune des transpositions de l'harmonisation. On peut aussi effectuer une harmonisation "approximative", où l'harmonisation

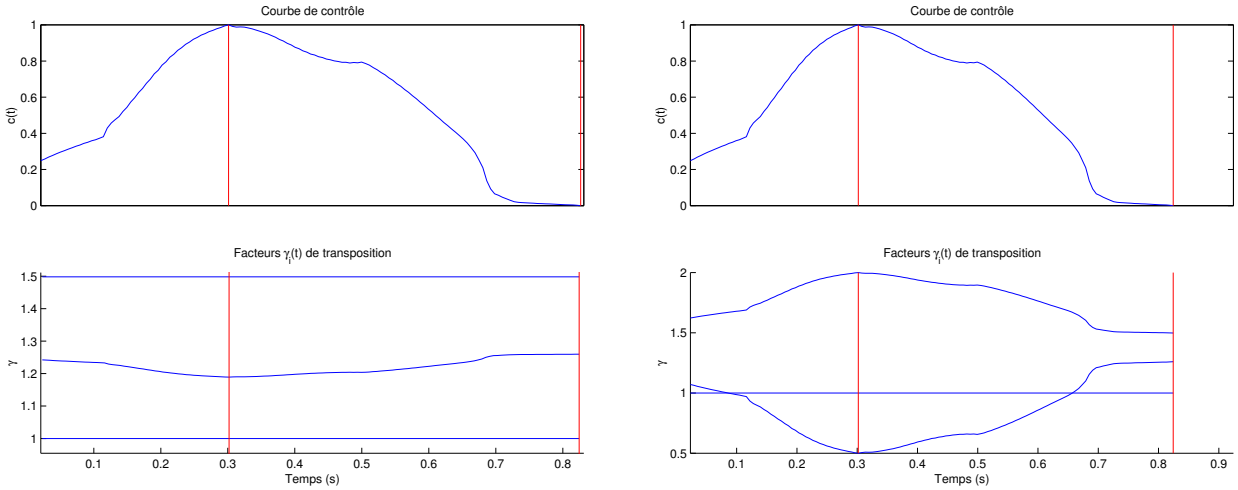


FIG. 5.28 – *Exemples d'harmonisation adaptative.*
A gauche : les deux accords de référence sont l'accord mineur et l'accord majeur Piste n°60-CD1 ♫♪.
A droite : les deux accords de référence sont l'accord majeur et l'octaviation supérieure et inférieure Piste n°61-CD1 ♫♪

se fait globalement sur tel accord, avec de micro-variations (par exemple entre la tierce mineure et la tierce majeur, ou sur la quinte, etc).

5.5 Effets adaptatifs sur la spatialisation

5.5.1 Echo adaptatif granulaire (TR, NTR)

Définition L'écho adaptatif granulaire est un écho dont le gain et le temps de délai peuvent varier. Cet effet est en fait basé sur des méthodes de synthèse granulaire. Nous avons montré précédemment que pour rendre adaptatif un effet, des améliorations doivent être portées aux algorithmes. Dans l'exemple donné en sec. II, la modification de la longueur d'une ligne à retard classique ne donne pas l'effet perceptif de retard variable auquel on s'attend, à cause des ruptures de phase et d'amplitude lors des raccourcissement de la ligne, et à cause des ajouts de zéros lors de l'allongement, mais aussi à cause du fait que tout le son de la ligne à retard subit uniformément la modification. La solution que nous proposons afin d'appliquer à un son un retard variable consiste à appliquer un écho granulaire. Le son est décomposé en grains, et chaque grain se voit parcourir une ligne à retard (réelle ou simulée) de propriétés différentes (*cf. fig.* 5.29).

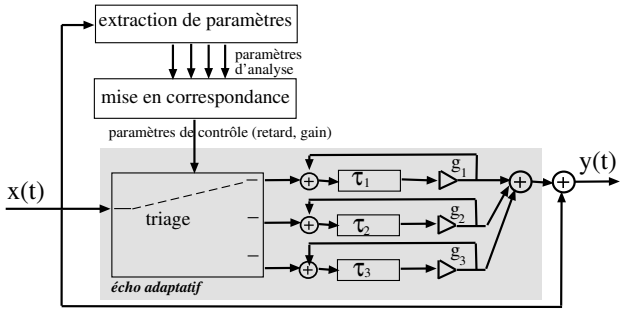


FIG. 5.29 – *Diagramme de l'écho adaptatif granulaire.*

Mise en œuvre Pour chaque grain du signal d'entrée $x_{in}(t) = x(t)\ w_A(t)$, on applique des répétitions (numérotées avec l'indice k) de gain $G(k,t)$ et de délai $T(k,t)$, donnés à partir de deux courbes de contrôle, l'une $g(t) \in [0;1[$ (en échelle linéaire) pour le gain et l'autre $\tau(t)$ (en s) pour le délai.

$$y\left(t + \sum_{j=1}^{k} T(k,t)\right) \quad = \quad G(k,t)\ x_{in}(t) \tag{5.43}$$

Dès lors, on peut calculer $G(k,t)$ et $T(k,t)$ de plusieurs manières. La plus simple consiste à utiliser, pour un même grain, le même gain et le même délai pour toutes les répétitions :

$$\left\{ \begin{array}{lcl} G(k,t) & = & g(t)^k \\ T(k,t) & = & k\tau(t) \end{array} \right. \tag{5.44}$$

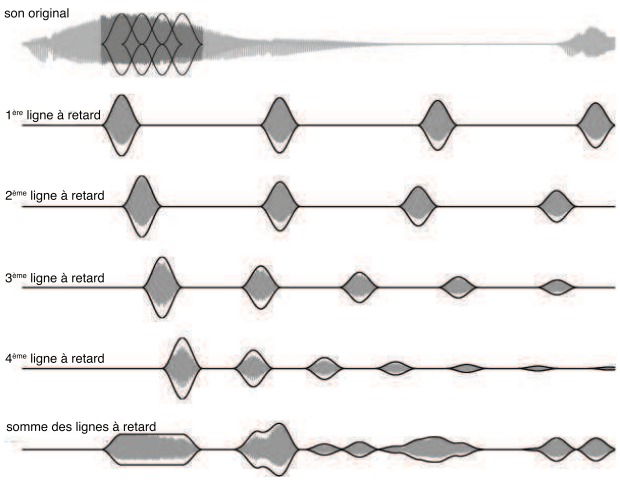


FIG. 5.30 – *Diagramme d'écho adaptatif granulaire : chaque grain est répété et retardé selon deux courbes de contrôle donnant le gain et le temps de délai.*

Un exemple est donné *fig.* 5.30 : chaque grain parcourt une ligne à retard dépendant des propriétés $g(t)$ et $\tau(t)$, contrôles provenant de descripteurs du son. Des variantes consistent à faire perdre ce synchronisme entre le grain traité et les contrôles, en lisant la courbe de contrôle aux temps de synthèse et non plus aux temps d'analyse. La première variante consiste à faire varier le gain $G(k,t)$ en fonction de la valeur de $g(t)$ à l'instant auquel le grain de synthèse est ajouté :

$$\left\{ \begin{array}{lcl} G(k,t) & = & G(k-1,t).g\left(t + \sum_{j=1}^{k} T(k,t)\right) \\ T(k,t) & = & k\tau(t) \end{array} \right. \tag{5.45}$$

De la même manière, on peut faire varier le délai $T(k,t)$ en fonction de la valeur de $\tau(t)$ à l'instant auquel le grain de synthèse est ajouté :

$$\left\{ \begin{array}{lcl} G(k,t) & = & g(t)^k \\ T(k,t) & = & T(k-1,t) + \tau\left(t + \sum_{j=1}^{k-1} T(k-1,t)\right) \end{array} \right. \tag{5.46}$$

On peut enfin combiner ces deux variantes :

$$\left\{ \begin{array}{lcl} G(k,t) & = & G(k-1,t).g\left(t + \sum_{j=1}^{k} T(k,t)\right) \\ T(k,t) & = & T(k-1,t) + \tau\left(t + \sum_{j=1}^{k-1} T(k-1,t)\right) \end{array} \right. \tag{5.47}$$

On peut enfin utiliser une pondération entre la valeur du contrôle au temps d'analyse et sa valeur au temps de synthèse :

$$\begin{cases} G(k,t) &= G(k-1,t).\frac{g(t)+g\left(t+\sum_{j=1}^{k} T(k,t)\right)}{2} \\ T(k,t) &= T(k-1,t) + \frac{\tau(t)+\tau\left(t+\sum_{j=1}^{k-1} T(k-1,t)\right)}{2} \end{cases} \tag{5.48}$$

Limitations La description de l'écho granulaire qui a été faite jusqu'ici possède une limitation : le délai doit être un nombre d'échantillons entier, et donc un multiple de $1/F_e$ s. D'autre part, la mise en œuvre hors temps-réel de ce qui a été présenté jusqu'ici ne peut se faire sans limiter le nombre de lignes à retard. En effet, si l'on considère que n'importe quel valeur de gain et n'importe quel délai de M échantillons est possible, alors il faut autant de lignes à retard que de valeurs de couples $(G(k,t), T(k,t))$ pour une mise en œuvre temps-réel, ce qui est impossible.

Solutions pour lever ces limitations La première limitation concernant la valeur des délais peut être dépassée en interpolant chaque grain aux valeurs d'échantillonnage. Soit $x_{in}(t), t = 1 : N$ un grain de N échantillons. Le retard qui y est associé se décompose comme un nombre entier D de fois $1/F_e$ plus δ une fraction de $1/F_e$.

$$\tau = \frac{D+\delta}{F_e} \tag{5.49}$$

En utilisant une interpolation cubique, le grain $x_{out}(t)$ est calculé comme $x_{out}(t+D) = x_{in}(t-\delta)$. La limitation concernant les temps de délais n'en est alors plus une.

En ce qui concerne le passage au temps-réel, la solution consiste à utiliser un nombre limité de lignes à retard traditionnels, à retard entier ou non, et à discrétiser les courbes de contrôle de gain et de retard, de manière à effectuer une sorte de triage des grains, et à faire passer chaque grain dans la ligne à retard de couple (G,T) le plus proche des valeurs données par les courbes. Ceci implique une moindre précision pour le gain et pour le retard, ce qui s'entend dès lorsque l'on ne prend pas suffisamment de lignes. Pour plus de détails, ce point est abordé dans la partie concernant les fonctions de discrétisation, *cf.* sec. 6.4.3.

Exemples sonores

La première série d'exemples que nous donnons a été obtenue à taille de grain constante (2048 échantillons) à partir de la Piste n°1 ♫♪ . Nous avons appliqué un écho dont seul le gain de réinjection est contrôlé par une courbe dérivée du RMS ($\tau = 0.3$ s), afin d'appliquer l'écho de manière prépondérante sur les attaques (Piste n°65-CD1 ♫♪) ou sur les parties harmoniques (Piste n°66-CD1 ♫♪). La forme d'onde du son original est donnée *fig.* 5.31, le RMS *fig.* 5.31 et la courbe de contrôle correspondant à un lissage du RMS après troncature *fig.* 5.32. La longueur des lignes à retard est fixe.

La deuxième série d'exemples est cette fois le délai granulaire appliqué à une voix parlée dans une langue imaginaire Piste n°10 ♫♪ . Si l'on fixe le gain à 0.4 et que le temps de délai varie entre 0.2 et 1 seconde, en utilisant comme courbe de contrôle la fréquence fondamentale, on obtient un son de synthèse granulaire dont la hauteur monte (Piste n°67-CD1 ♫♪), étiré dans le temps au fur et à mesure de ses répétitions. Si au contraire on fixe le temps de délai à 0.2 seconde et fait varier le gain de réinjection en fonction de la balance de voisement, on obtient une voix dont seules les consonnes sont répétées (Piste n°68-CD1 ♫♪).

Illustration des problèmes de mise en œuvre en temps-réel Lorsque le délai granulaire est mis en œuvre en temps réel, le nombre de lignes à retard est limité. Aussi, il faut discrétiser la courbe de contrôle (*cf.* 6.4.3). Pour de petites grilles (ie. $n_q \leq 30$), l'écho granulaire adaptatif est clairement différent de

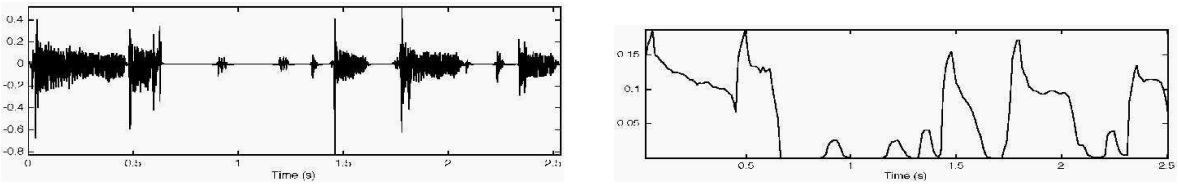


FIG. 5.31 – *A gauche : forme d'onde du son de guitare (Piste n°1 ♫♪).*
A droite : courbe de contrôle de l'écho granulaire adaptatif : énergie par RMS du son de guitare (Piste n°1 ♫♪).

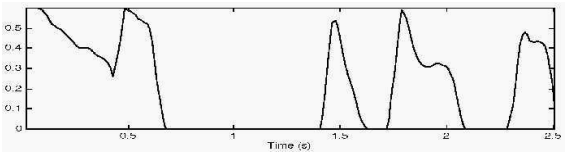


FIG. 5.32 – *Courbe de contrôle après troncature et lissage du RMS du son de guitare.*

celui obtenu par la mise en œuvre en temps différé, bloc par bloc. Par exemple, un effet persistant de filtrage en peigne va apparaître à une hauteur définie, du fait que le pas entre deux grains répétés ne peut pas prendre toutes les nuances de valeurs possibles. Aussi, nous recommandons l'utilisation d'au moins 60 lignes à retard concernant la discrétisation du temps de délai, et 20 lignes à retard concernant la discrétisation du gain. Aussi, si l'on désire moduler à la fois le délai et le gain, il faut 1200 lignes à retard : c'est bien trop pour fonctionner en temps réel, aussi, on doit choisir de moduler uniquement le gain ou uniquement le temps de délai, ou alors diminuer la discrétisation.
CD2 Les exemples sonores suivants ont été présentés à DAFx-02 Hambourg (**passer du CD1 au CD2**) :

- la Piste n°44-CD2 ♫♪ correspond à écho granulaire adaptatif obtenu à l'échantillon près (hors temps-réel), appliqué à la Piste n°11 ♫♪ ;
- la Piste n°45-CD2 ♫♪ correspond à écho granulaire adaptatif temps réel obtenu par une quantification uniforme de la courbe de contrôle, avec $n_q = 3$: le nombre de valeurs de quantification est très insuffisant.
- idem avec $n_q = 10$ (Piste n°46-CD2 ♫♪) : le nombre de valeurs de quantification est insuffisant, même si on commence à reconnaître le son obtenu.
- idem avec $n_q = 30$ (Piste n°47-CD2 ♫♪) : le son est assez ressemblant, mais il manque encore de précision.
- idem avec $n_q = 60$ (Piste n°48-CD2 ♫♪) : ne nombre de valeurs de quantification est suffisant, le son est quasiment identique.

Intérêt de cet effet Ces différents exemples de délai granulaire adaptatif montrent l'intérêt musical qu'apportent ces nouveaux effets audionumériques pour la composition et le traitement du jeu en direct. L'écho granulaire adaptatif permet de produire des échos sélectifs (avec $\tau = cte$), lorsque des portions d'intérêt du son deviennent plus ou moins présentes dans les répétitions (par exemple en ne répétant que les forts RMS, les portions harmoniques du signal ou les attaques), ce qui d'une certaine manière correspond au *morphing* de timbre dans le sens utilisé par Leigh Landy [Landy, 1991]. Cet effet permet aussi de rendre un son plus "éthérique" (avec $g = cte$), avec des modifications de la localisation temporelle des répétitions. C'est dans cette configuration que l'aspect granulaire de la méthode s'entend le plus.

5.5.2 Panoramisation adaptative (TR, NTR)

Principe Nous avons mis en œuvre une panoramisation adaptative à puissance constante, à la fois sous *Max/MSP* et sous *Matlab*. Les niveaux gauche α_G et droite α_D sont calculés à partir d'une

courbe de contrôle $\mathcal{C}_i(t)$, à partir de la loi de Blumlein ([Blauert, 1983], [Rochesso, 2002] pp.138-41), avec $\theta = \mathcal{C}(t) \in [-\frac{\pi}{4}; \frac{\pi}{4}]$ et :

$$\begin{aligned} \alpha_G(t) &= \frac{\sqrt{2}}{2}(\cos\theta(t) + \sin\theta(t)) \\ \alpha_D(t) &= \frac{\sqrt{2}}{2}(\cos\theta(t) - \sin\theta(t)) \end{aligned}$$

Déplacement rapide selon une loi non sinusoïdale On sait que lorsque le son se déplace rapidement, à des fréquences proches ou même supérieures à 20 Hz, le type de mouvement influe sur la perception de l'effet. Si le mouvement est sinusoïdal, la modulation d'amplitude sur les signaux gauche et droit est entendue dans le domaine fréquentiel (*cf.* sec. 1.2.6 et sec. 1.2.8). Par contre, si le mouvement est plus complexe qu'une simple sinusoïde, le modulation d'amplitude peut être perçue comme une ségrégation de flux auditifs, dès lors qu'à chaque position correspond un timbre particulier [Bregman, 1990]. Ainsi, si on effectue une panoramisation adaptative en fonction de paramètres du timbre (CGS, par exemple), on peut donner à un son monophonique la perception d'être composé de plusieurs sources positionnées différemment.

Exemples sonores (retour au CD1) Les premiers exemples sonores donnés sont basés sur la Piste n°4 ♫♪ monophonique, son de synthèse réalisé par John Chowning à l'aide de la synthèse FM. Ces exemples ont été réalisés en son honneur, lors de sa venu à Marseille pour recevoir le titre de Docteur *Honoris Causa* de l'Université d'Aix-Marseille II (en Novembre 2002). Le premier exemple, Piste n°69-CD1 ♫♪ illustre la perception de plusieurs flux sonores. La première trompette synthétique semble déjà "à cheval" sur les deux canaux, et lorsque la seconde trompette entre en jeu, elle semble elle aussi se déplacer, par symétrie avec la première. Le deuxième exemple, Piste n°70-CD1 ♫♪), illustre les problèmes dus à des variations trop rapides de l'angle $\theta(t)$: on entend aux débuts et fins de chaque note un effet de type *noise gate*, où le bruit de fond est supprimé brusquement. Ceci semble dû au mouvement rapide du son à cet instant où il passe d'un son harmonique de niveau fort à un son bruité de niveau faible. Un dernier exemple (Piste n°71-CD1 ♫♪) est donné à partir de la Piste n°3 ♫♪. Celui-ci illustre la façon dont cet effet peut produire des déplacements assez lents pour un source seule.


CD1

Variante Une variante consiste à utiliser deux courbes de contrôle, une pour le gain de chaque signal.

$$\begin{aligned} \alpha_G(t) &= \mathcal{C}_1(t) = g(t)\frac{\sqrt{2}}{2}(\cos\theta(t) + \sin\theta(t)) \\ \alpha_D(t) &= \mathcal{C}_2(t) = g(t)\frac{\sqrt{2}}{2}(\cos\theta(t) - \sin\theta(t)) \end{aligned}$$

Ceci ne correspond plus à une simple panoramisation. En effet, étant donnés deux gains $\alpha_G(t)$ et $\alpha_D(t)$ non corrélés, on peut calculer l'angle de panoramisation correspondant à une panoramisation à puissance constante. Il reste alors un terme de gain $g(t)$, correspondant à un changement de niveau global, appliqué aux deux canaux. Le gain $g(t)$ et l'angle $\theta(t)$ sont calculés à partir des courbes de contrôle. On a :

$$\begin{aligned} \cos\theta(t) &= \frac{\alpha_G(t) + \alpha_D(t)}{\sqrt{2}g(t)} \\ \sin\theta(t) &= \frac{\alpha_G(t) - \alpha_D(t)}{\sqrt{2}g(t)} \\ g^2(t) &= \alpha_G^2(t) + \alpha_D^2(t) \end{aligned}$$

De ces trois équations à deux inconnues, on tire les valeurs de $g(t)$ et $\theta(t)$ en fonction des niveaux :

$$\begin{aligned} g(t) &= \sqrt{\alpha_G^2(t) + \alpha_D^2(t)} \\ \theta(t) &= \tan^{-1}\left(\frac{\alpha_G(t) - \alpha_D(t)}{\alpha_G(t) + \alpha_D(t)}\right) \end{aligned}$$

ce qui peut encore s'exprimer en fonction des courbes de contrôle :

$$g(t) = \sqrt{\mathcal{C}(1,t)^2 + \mathcal{C}(1,t)^2} \tag{5.50}$$

$$\theta(t) = \tan^{-1}\left(\frac{\mathcal{C}(1,t) - \mathcal{C}(2,t)}{\mathcal{C}(1,t) + \mathcal{C}(2,t)}\right) \tag{5.51}$$

Cet effet est une combinaison du changement de distance adaptatif et de la panoramisation adaptative. En effet, comme on peut le voir avec l'écriture du gain $g(t)$, il s'ajoute un mouvement de rapprochement et d'éloignement (mis à part l'effet de salle absent) à l'effet de panoramisation.

5.5.3 Panoramisation spectrale adaptative (TR, NTR)

Principe Il s'agit d'une mise en œuvre particulière de l'équaliseur adaptatif (*cf.* sec. 5.6.1) mis en œuvre avec le filtrage par TFCT. Une forme $\alpha_G(f,t)$ contrôle le gain de chaque panier de fréquence du canal gauche, et sa forme "complémentaire" (en respectant la puissance totale du signal) contrôle le gain des paniers de fréquence du canal droit. Cette forme peut être la forme d'onde ou l'enveloppe spectrale calculée par le cepstre et normalisée entre 0 et 1 après *mapping*. La forme "complémentaire" $\alpha_D(f,t)$ est calculée de manière à varier elle aussi entre 0 et 1 et à conserver la puissance constante, pour chaque composante spectrale.

Mise en œuvre Aux équations *eq.* (3.16) et *eq.* (3.15), on a donné les formules de gain à appliquer à chaque canal pour placer le son à l'azimut θ. La courbe $\alpha_G(f,t)$ est donnée en fonction de l'angle $\theta(f,t) \in [-\pi/4; \pi/4]$ pour le canal gauche, et la courbe $\alpha_G(f,t)$ est donnée en fonction de l'angle $-\theta(f,t) \in [-\pi/4; \pi/4]$ pour le canal droit. Le paramètre de contrôle est donc $\theta(f,t)$, car à partir de lui, on peut calculer pour chaque panier de fréquence le gain du panier $\alpha_{G/D}(f,t)$ dans chaque canal :

$$\alpha_G(f,t) = \frac{\sqrt{2}}{2}\left(\cos\theta(f,t) + \sin\theta(f,t)\right) \tag{5.52}$$

$$\alpha_D(f,t) = \frac{\sqrt{2}}{2}\left(\cos\theta(f,t) - \sin\theta(f,t)\right) \tag{5.53}$$

Exemple sonore Comme on peut l'entendre avec l'exemple sonore Piste n°72-CD1 ♫♪ obtenu à partir de la piste monophonique Piste n°14 ♫♪, cet effet permet d'ajouter plusieurs plans sonores à un seul son. En effet, les composantes spectrales sont traités de manières différentes, et l'on entend le son comme provenant de plusieurs sources sonores se déplaçant selon des trajets différents. Le son n'est pas placé quelque part entre la gauche et la droite, mais c'est une modification plus subtile et plus complexe à saisir pour l'auditeur qui s'opère.

5.5.4 Spatialisation adaptative temps-réel (TR)

Principe Il s'agit ici d'utiliser les techniques de spatialisation en Ambisonic pour rendre la sensation de trajectoires sonores dans un espace muni de 4 ou 8 haut-parleurs sur un cercle autour de l'auditeur.

Mise en œuvre La mise en œuvre est décrite en détail en sec. 7.2.1. Ce projet a eu lieu durant une collaboration dans le cadre de l'ACI[1] "*Espaces Sonores*" avec Anne Sédès, Benoît Courribet et Jean-Baptiste Thiébaut, au CICM[2] de Paris VIII. La position du son est donné sur une trajectoire par la valeur d'un descripteur simple (RMS, CGS). Nous avons choisi une trajectoire elliptique car elle permet de passer continûment du cercle au segment en passant par l'ellipse : le contrôle gestuel porte notamment sur la forme de cette trajectoire. Le déplacement sur la trajectoire peut se faire en position ou ne vitesse.

Résultats Nous ne pouvons hélas pas faire entendre d'exemple sonore sans disposer d'un système quadri ou octophonique. Nous pouvons cependant donner ici quelques uns des résultats présentés en sec. 7.2.1. Concernant le déplacement du son en position, on retrouve d'une manière plus générale le résultat de la panoramisation adaptative, à savoir que des déplacements rapides non sinusoïdaux s'entendent comme une ségrégation de flux auditifs [Bregman, 1990]. Toutes les composantes du son ayant la même valeur du descripteur sont placées au même endroit du cercle de diffusion. Du fait de la concordance à la fois de la position et de l'énergie si c'est le RMS qui contrôle la position, l'oreille regroupe ces composantes en un seul et même flux. Le même phénomène apparaît s'il y a concordance entre la position et la brillance si c'est le CGS qui contrôle la position. Ainsi, pour différentes positions prépondérantes (dues à différentes valeurs du descripteur utilisé comme contrôle), on aura différents flux.

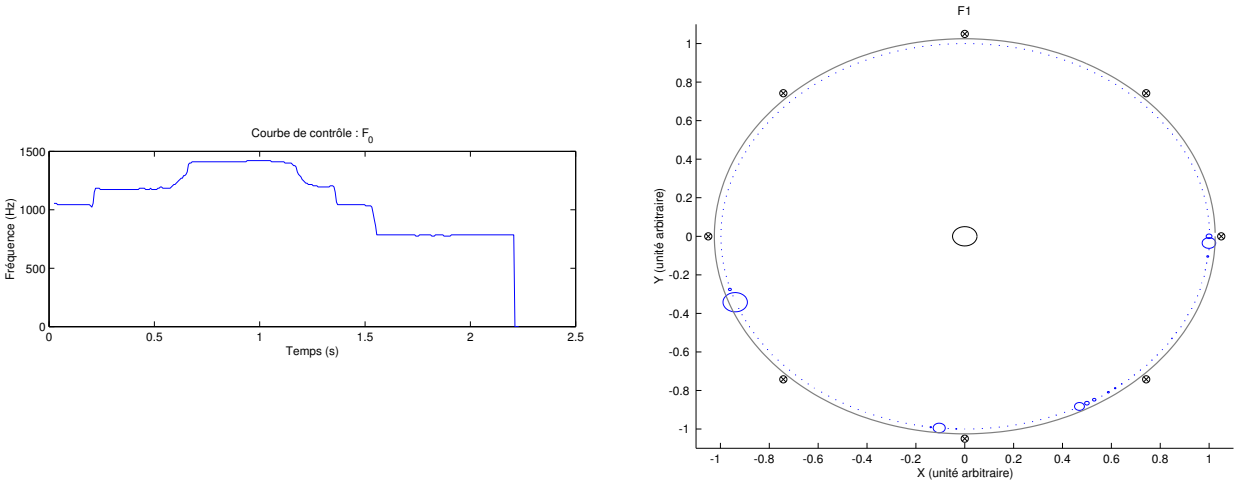


FIG. 5.33 – *A gauche : courbe de contrôle de la spatialisation adaptative : fréquence F_1 du premier harmonique.*
A droite : occupation du cercle de diffusion par la Piste n°8 ♫♪ spatialisée, contrôlé par F_1. Le grand cercle gris symbolise le cercle de diffusion, avec les 8 haut-parleurs symbolisés par les croix ; le cercle central symbolise l'auditeur. Les points et les petits cercles représentent l'occupation de l'azimuth par le son.

Nous donnons un exemple graphique du positionnement du son spatialisé avec pour contrôle la fréquence fondamentale du son Piste n°8 ♫♪ (*cf. fig.* 5.33 gauche). L'occupation du cercle de diffusion, *cf. fig.* 5.33 droite est représentée par des cercles, plus ou moins grands selon le temps total d'occupation de la zone. On voit que quatre zones principales sont occupées ; le son sera perçu comme séparé en quatre flux différents. Remarquons que sur cette représentation, il n'y a pas de notion d'ordre temporel des événements (nous vérifions cependant sur la courbe de contrôle qu'il s'agit bien que chaque groupe correspond à des notes). Pour un son variant bien plus vite, le son n'est plus séparé en notes à des positions différentes, mais c'est bien une ségrégation de flux qui a

[1] ACI : Action Concertée Incitative
[2] CICM : Centre de recherche en Informatique et en Création Musicale

lieu suite aux variations rapides de position.

Concernant le déplacement en vitesse, cet effet permet par exemple de déplacer le son uniquement pendant les silences (le *mapping* attribuera à un faible RMS une grande vitesse, et inversement), ce qui donne la sensation d'un interprète se déplaçant d'une manière très corrélée par rapport à ce qu'il interprète. Il est encore possible de ne déplacer le son que pendant que son niveau est assez fort : des accélérations du déplacement sur la trajectoire se produisent alors lorsque le niveau monte, ce qui ressemble au changement de position et de dynamique de l'effet Doppler (sans le changement de hauteur).

5.5.5 Directivité adaptative temps-réel (TR)

En janvier 2003, lors d'une collaboration avec Nicolas Misdariis à l'IRCAM, nous avons interfacé notre contrôle adaptatif avec le dispositif **LeCube** (ou "la Timée"). Ce dispositif permet de simuler le rayonnement de source acoustique à partir d'un système multi haut-parleurs, *cf.* sec. 3.5.10. Cela a été l'occasion de s'entendre confirmer tout l'intérêt que les compositeurs utilisant ce système portent au contrôle adaptatif. Nous avons utilisé trois contrôles simples : le premier en vitesse de rotation, le deuxième en position et le troisième en type de directivité. Les descripteurs utilisés étaient soit le RMS, soit le CGS.

Concernant les résultats, ils ne peuvent être reproduits sans ce système, aussi nous ne pouvons joindre aucun exemple au CD d'illustrations sonores. Pour des sons dont le descripteur utilisé varie trop rapidement, le contrôle en position ne se perçoit pas très bien, alors que le contrôle en vitesse est intéressant. Par contre, pour des contrôles à lentes variations (un filtrage passe-bas ayant été appliqué au contrôle), le contrôle en position est perceptible.

5.5.6 Remarques sur les autres effets adaptatifs sur la spatialisation

Une partie prospective serait d'investiguer dans la direction de la réverbération adaptative, du changement de distance (changement de niveau et réverbération adaptatives), qui auraient pu compléter ces effets sur la dimension spatiale du son, en prenant en compte l'effet de salle.

5.6 Effets adaptatifs sur le timbre

Les effets portant sur le spectre fréquentiel du signal (nous dirons "sur le timbre" par abus de langage) peuvent modifier l'enveloppe spectrale seule. Ils peuvent aussi modifier la structure du spectre de magnitude en préservant l'enveloppe spectrale. Ils peuvent enfin modifier à la fois l'enveloppe spectrale la structure fine du spectre, en modifier le spectre de phase ou en modifiant le spectre de magnitude. C'est selon ces quatre approches que nous allons maintenant présenter les effets sur le timbre.

5.6.1 Effets adaptatifs sur l'enveloppe (NTR)

Les effets modifiant l'enveloppe spectrale peuvent être vus comme des filtrages, qui atténuent ou amplifient certaines bandes de fréquences du spectre sans modifier la répartition fréquentielle des raies spectrales.

5.6.1.i) Filtrage adaptatif (NTR)

Lorsqu'on utilise l'expression "filtrage adaptatif", on pense tout d'abord aux méthodes d'estimation des paramètres d'un filtre [Haykin, 1996]. L'exemple des télécommunication a été donné en sec. 5.1.8. Les effets de filtrage adaptatif qui nous intéressent ont des paramètres (coefficients, bande passante, facteur de qualité, formants, etc.) qui évoluent dans le temps d'après des paramètres du

son et une mise en correspondance. On les utilise dans un but musical, et ceci n'implique pas l'utilisation des mêmes techniques. Nous allons présenter des modifications adaptatives de l'enveloppe spectrale, un changement de centroïde, un équaliseur adaptatif par banc de filtres à facteur de qualité constant (domaine temporel) et par TFCT (domaine fréquentiel), un filtre en peigne adaptatif, un compresseur spectral adaptatif, des filtres vocaliques et wha-wha adaptatifs, un changement de voyelle et une conversion de voix en temps-réel. On considère à chaque fois un signal que l'on analyse et modifie à l'aide du vocodeur de phase ou du modèle additif. On extrait l'enveloppe spectrale de chaque grain à partir de la TFCT du grain, dans le domaine fréquentiel, à l'aide de la méthode du cepstre.

5.6.1.ii) Modifications sur l'enveloppe spectrale

Les modifications de l'enveloppe spectrale ont été présentées *cf.* sec. 3.6.1. Leur version adaptative consiste à faire évoluer dans le temps les paramètres de cette modification, que ce soit un décalage, une dilatation/contraction en fréquence ou une conformation. Cela a été réalisé sous *Matlab*. Les exemples sonores qui illustrent ces effets proviennent de plusieurs sons : un glissement de doigts sur une basse électrique, Piste n°9 ♫♪, et la voix de Pierre Schaeffer, Piste n°16 ♫♪.

Décalage adaptatif de l'enveloppe spectrale (NTR) Le seul paramètre de contrôle du décalage de l'enveloppe spectrale est la valeur de ce décalage d en Hz. Cette valeur peut être positive ou négative. Un décalage renforce les composantes fréquentielles se trouvant dans la direction où ce décalage a lieu. En effet, les formants sont déplacés avec l'enveloppe, aussi un décalage vers les basses fréquences renforce celles-ci. Il en va de même pour le décalage vers les hautes fréquences. L'intérêt de rendre cet effet adaptatif est qu'on ne se limite plus à mettre en valeur les composantes basses fréquences seules ou les composantes en hautes fréquences seules. Au cours du temps, ce sont tantôt les composantes basses fréquences, tantôt les composantes hautes fréquences dont l'amplitude est modifiée. Ceci apporte une plus grande variabilité au son obtenu par le traitement, *cf.* Piste n°73-CD1 ♫♪ pour un exemple sur un son électroacoustique et Piste n°74-CD1 ♫♪ pour un exemple sur la voix.

Dilatation/contraction adaptative de l'enveloppe spectrale (NTR) La dilatation/contraction adaptative de l'enveloppe spectrale a un seul paramètre de contrôle (*cf.* sec. 3.6.1) : le facteur de changement d'échelle (dilatation/contraction). Pour une valeur constante inférieure à 1, l'effet fait ressortir les composantes en basses fréquences, puisque les formants sont contractés (et donc décalés vers les basses fréquences), *cf.* Piste n°75-CD1 ♫♪. Pour une valeur constante supérieure à 1, l'effet fait ressortir les composantes en hautes fréquences, puisque les formants sont dilatés, *cf.* Piste n°76-CD1 ♫♪. L'intérêt de rendre cet effet adaptatif est le même que précédemment, à savoir que l'on apporte une plus grande variabilité au timbre du son obtenu par le traitement, *cf.* Piste n°76-CD1 ♫♪ pour un exemple sur un son électroacoustique, Piste n°77-CD1 ♫♪ pour un exemple sur la voix et Piste n°78-CD1 ♫♪ pour un exemple sur un son instrumental (d'après la Piste n°8 ♫♪).

Conformation adaptative de l'enveloppe spectrale (NTR) Il s'agit ici de modifier l'enveloppe spectrale non plus uniquement selon son échelle ou selon son décalage, mais dans tous les sens, comme pour la conformation spectrale. Un méta-descripteur sert de fonction de conformation. Ce peut être la forme d'onde, le spectre d'amplitude, l'enveloppe spectrale, ou la somme cumulative de l'un de ces méta-descripteurs. Cette fonction de conformation $\mathcal{F}_{warp}$ donne pour chaque fréquence f la nouvelle amplitude de l'enveloppe $\mathcal{F}_{warp}(f)$ selon :

$$\mathcal{E}_{warp}(f,t) = \mathcal{E}\left(\mathcal{F}_{warp}(f,t)\right) \tag{5.54}$$

Ainsi, si $\mathcal{F}_{warp}(f,t) = f$, l'enveloppe n'est pas modifiée. Ce premier paramètre de contrôle permet de modifier dynamiquement l'enveloppe spectrale de manière inattendue. Un autre paramètre de

contrôle $\alpha(t)$ sert à modifier l'amplitude de l'effet de conformation spectrale, en utilisant comme fonction de conformation non plus $\mathcal{F}_{warp}$ mais la combinaison linéaire :

$$\mathcal{G}_{warp}(f,t) = (1-\alpha(t))\mathcal{F}_{warp}(f,t) + \alpha(t)f \tag{5.55}$$

Ceci permet d'avoir une conformation de l'enveloppe spectrale donnée par une courbe dynamique, elle-même contrôlée par un second paramètre afin d'ajouter des variations. Les sons que l'on peut obtenir sont surprenants, que ce soit pour le son électroacoustique Piste n°9 ♫♪ (*cf.* Piste n°79-CD1 ♫♪), pour la voix parlée Piste n°16 ♫♪ (*cf.* Piste n°80-CD1 ♫♪) ou pour le son instrumental Piste n°8 ♫♪ (*cf.* Piste n°81-CD1 ♫♪).

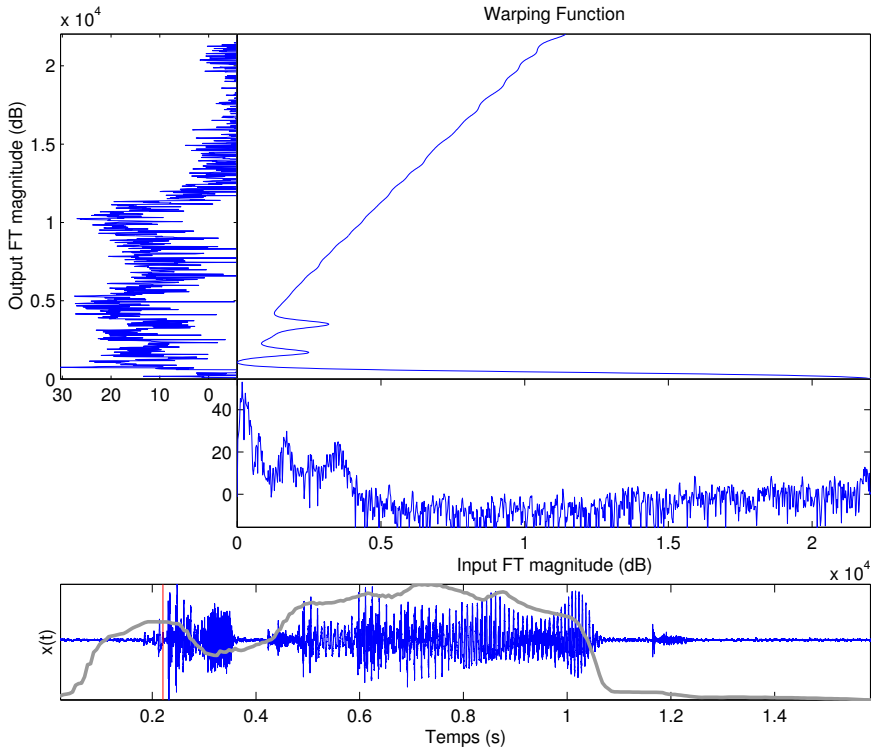


FIG. 5.34 – *Exemples de conformation adaptative de l'enveloppe spectrale. La figure inférieure présente la forme d'onde et la courbe $\alpha(t)$. La fonction de conformation $\mathcal{G}_{warp}(f)$ est donnée par l'enveloppe spectrale $\mathcal{E}(f,t)$.*

Remarque On peut faire une analogie de concept entre cet effet dans le domaine du timbre via le spectre, et la réorganisation temporelle granulaire dans le domaine du temps : on choisit de quelle manière on lit les données dans un domaine (le temps, les fréquences) afin de les réorganiser.

5.6.1.iii) Changement de centroïde (NTR)

Nous proposons d'utiliser les effets de changement adaptatif de l'enveloppe spectrale afin d'imposer une valeur $c_0(t)$ au centroïde d'un signal, et donc une brillance. Soit $x(t)$ un signal fenêtré, et $X(f)$ sa transformée de Fourier à court-terme. Rappelons l'une des formules du centroïde, ou centre de gravité spectrale :

$$\text{cgs}(X) = \frac{\sum_{f=0}^{F_e/2} f.X(f)}{\sum_{f=0}^{F_e/2} X(f)} \tag{5.56}$$

Nous proposons trois manières de transformer $X(f)$ en $\tilde{X}(f)$ de façon à ce que $\text{cgs}(\tilde{X}) = c_0$: par addition d'une forme spectrale donnée, par translation de l'enveloppe spectrale et par étirement de l'enveloppe spectrale.

Changement de centroïde par addition d'une forme spectrale donnée Soit $E(f)$ une enveloppe spectrale donnée, par exemple la courbe d'un équaliseur. Soit α un facteur multiplicatif. Nous cherchons α permettant d'imposer que $\text{cgs}(\tilde{X}) = c_0(t)$ avec :

$$\tilde{X}(f) = X(f) + \alpha E(f) \tag{5.57}$$

Le centroïde de $\tilde{X}$ s'exprime :

$$\text{cgs}(\tilde{X}) = \frac{\sum_{f=0}^{F_e/2} f.\tilde{X}(f)}{\sum_{f=0}^{F_e/2} \tilde{X}(f)}$$

$$\alpha\ \text{cgs}(\tilde{X}) \sum_{f=0}^{F_e/2} E(f) + \text{cgs}(\tilde{X}) \sum_{f=0}^{F_e/2} X(f) = \sum_{f=0}^{F_e/2} fX(f) + \alpha \sum_{f=0}^{F_e/2} fE(f)$$

D'où finalement :

$$\text{cgs}(\tilde{X}) = \frac{\text{cgs}(X) \sum_{f=0}^{F_e/2} X(f) + \text{cgs}(E)\alpha \sum_{f=0}^{F_e/2} E(f)}{\sum_{f=0}^{F_e/2} X(f) + \alpha \sum_{f=0}^{F_e/2} E(f)} \tag{5.58}$$

Les spectres d'amplitude $X(f)$ et $E(f)$ sont connus, donc les sommes $\sum_{f=0}^{F_e/2} X(f)$ et $\sum_{f=0}^{F_e/2} E(f)$ sont connues, ainsi que les centroïdes $\text{cgs}(X)$ et $\text{cgs}(E)$. On peut alors modifier α par itérations afin de faire coïncider la valeur du centroïde $\text{cgs}(\tilde{X})$ avec $c_0(t)$ dans l'équation *eq.* (5.58).

Changement de centroïde par translation de l'enveloppe spectrale Le spectre d'amplitude modifié par translation en fréquences d'une valeur f_t est donné par :

$$\tilde{X}(f) = X(f_t + f)\mathbb{1}_{(f_t+f)\in[0;F_e/2]} \tag{5.59}$$

Nous pouvons alors calculer son CGS :

$$\begin{aligned}
\text{cgs}(\tilde{X}) &= \frac{\sum_{f=0}^{F_e/2} f.X(f_t + f)\mathbb{1}_{(f_t+f)\in[0;F_e/2]}}{\sum_{f=0}^{F_e/2} X(f_t + f)\mathbb{1}_{(f_t+f)\in[0;F_e/2]}} \\
&= \frac{\sum_{f=0}^{F_e/2} (f_t + f).X(f_t + f)\mathbb{1}_{(f_t+f)\in[0;F_e/2]} - \sum_{f=0}^{F_e/2} f_t.X(f_t + f)\mathbb{1}_{(f_t+f)\in[0;F_e/2]}}{\sum_{f=0}^{F_e/2} X(f_t + f)\mathbb{1}_{(f_t+f)\in[0;F_e/2]}}
\end{aligned}$$

D'où la formulation finale :

$$\text{cgs}(\tilde{X}) = \text{cgs}\left(X(f_t + f)\mathbb{1}_{(f_t+f)\in[0;F_e/2]}\right) - f_t \tag{5.60}$$

Changement de centroïde par étirement de l'enveloppe spectrale Le spectre d'amplitude modifié par translation en fréquences d'une valeur f_t est donné par :

$$\tilde{X}(f) = X(\gamma f)\mathbb{1}_{\gamma f\in[0;F_e/2]} \tag{5.61}$$

Nous pouvons alors calculer son CGS :

$$\begin{aligned} \text{cgs}(\tilde{X}) &= \frac{\sum_{f=0}^{F_e/2} f.X(\gamma f)\mathbb{1}_{(\gamma f)\in[0;F_e/2]}}{\sum_{f=0}^{F_e/2} X(\gamma f)\mathbb{1}_{(\gamma f)\in[0;F_e/2]}} \\ &= \frac{1}{\gamma}\frac{\sum_{f=0}^{F_e/2} \gamma f.X(\gamma f)\mathbb{1}_{(\gamma f)\in[0;F_e/2]}}{\sum_{f=0}^{F_e/2} X(\gamma f)\mathbb{1}_{(\gamma f)\in[0;F_e/2]}} \end{aligned}$$

D'où la formulation finale :

$$\text{cgs}(\tilde{X}) = \frac{1}{\gamma}\text{cgs}_\gamma\left(X(\gamma f)\mathbb{1}_{(\gamma f)\in[0;F_e/2]}\right) \tag{5.62}$$

avec cgs_γ la fonction cgs calculée aux points de fréquence $\gamma f = \gamma\frac{k}{N}F_e$ $k \in \{0, ..., N/2\}$.

5.6.1.iv) Equaliseur adaptatif (TR)

L'équaliseur permet de modifier l'enveloppe du spectre (*cf.* sec. 3.6.1). Nous avons utilisé une version par TFCT de l'équaliseur afin d'en réaliser une version adaptative sous *Max/MSP* (*cf.* vidéo B.3). Les filtres ne sont pas paramétriques, et le banc de filtres n'est pas à facteur de qualité constant. Ceci dit, cette mise en œuvre est plus souple pour permettre le contrôle adaptatif.

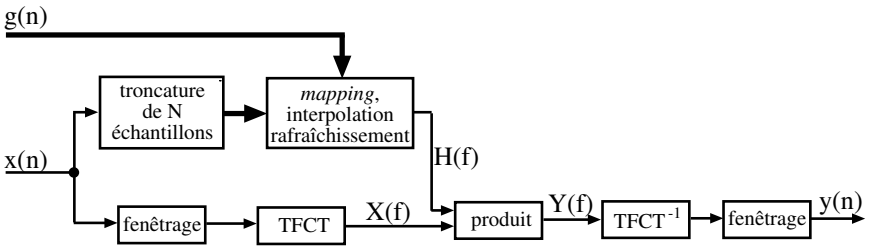


FIG. 5.35 – *Diagramme de l'équaliseur adaptatif dans le domaine fréquentiel, contrôlé par le geste.*

Le principal problème à résoudre consiste à trouver suffisamment de paramètres du son pour piloter cet effet. Dans notre cas (application du filtre dans le domaine fréquentiel), pour une TFCT de N_{FFT} points, il nous faut $\frac{N_{FFT}}{2}$ valeurs de contrôle. Nous avons donc pensé utiliser comme méta-descripteur les échantillons du signal d'entrée (*cf. fig.* 5.36). Par exemple, les $\frac{N_{FFT}}{2}$ premiers échantillons du signal $x(t) \in [-1; 1]$ sont transformées en valeurs entre 0 et 1 en module. Ainsi, pour une valeur nulle, le panier de fréquence correspondant ne contiendra plus la composante sinusoïdale originale ; pour une valeur nominale 1, le panier de fréquence correspondant conservera à l'identique sa composante. Un autre méta-descripteur possible est le spectre d'amplitude, ou son enveloppe. Ceux-ci sont déjà très proches du spectre à filtre : ils permettront de mettre encore plus en évidence les formants et la structure harmonique du son. La forme d'onde est quant à elle de forme bien différente du spectre, et permet des modifications moins directement compréhensibles pour l'utilisateur.

Le méta-descripteur est utilisé comme le module de la fonction de transfert du filtre à appliquer. Il est modifié (rafraîchit) à intervalles temporels réguliers. Ceci dit, on peut vouloir que le taux de rafraîchissement de la forme de l'équaliseur soit inférieur à celui de calcul de la TFCT : en effet, pourquoi changer à chaque itération temporelle la forme de l'équaliseur ? Notons de plus que faire évoluer rapidement le filtrage revient à effectuer des modulations d'amplitude rapides à la composante sinusoïdale de chaque panier de fréquence, avec les risques déjà évoqués précédemment (*cf.* sec. 1.2.6 et sec. 1.2.8). On utilise alors une interpolation entre une forme d'équaliseur source $H_s(f)$ et une forme cible $H_c(f)$. Pour un taux de rafraîchissement de T_{raf} *Hz* et un taux de calcul

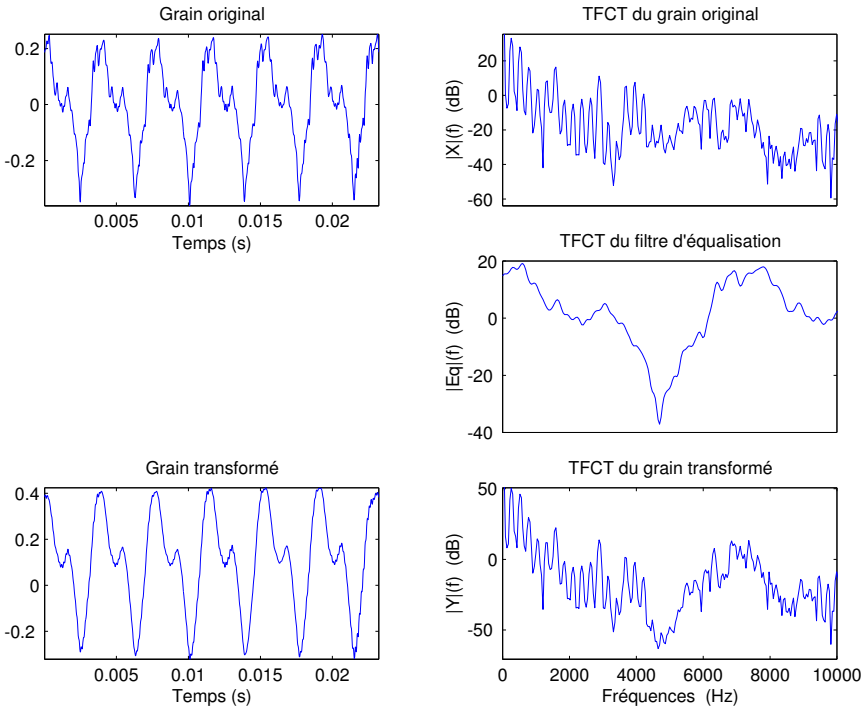


FIG. 5.36 – *Illustration de l'équaliseur adaptatif.*

de TFCT de $T_{TFCT} = \frac{F_e}{N_{TCFT}}$ Hz, on utilisera la forme donnée par $\alpha H_s(f) + (1-\alpha)H_c(f)$ avec $\alpha = \frac{\Delta t}{T_{raf}}$ et Δt l'intervalle de temps écoulé depuis le dernier rafraîchissement.

Le contrôle gestuel peut ensuite s'effectuer soit sur une modification de la forme de la fonction de transfert du filtre (par translation vers le haut ou le bas, par étirement vertical) afin d'amplifier les variations, soit sur le rapport entre taux de rafraîchissement de la forme de l'équaliseur et taux de rafraîssement de la fonction de transfert. Cette dernière façon de contrôler l'équaliseur adaptatif permet de donner une variabilité à la vitesse d'évolution du filtrage ; on peut aussi contrôler cette vitesse d'évolution par le son, par exemple par le RMS, pour que les sons globalement fort évoluent plus vite que les sons faibles, ou inversement.

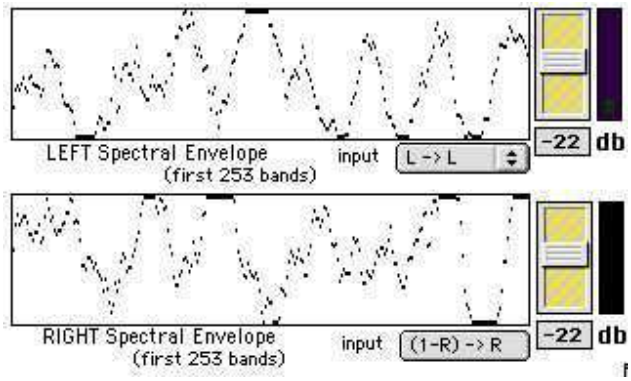


FIG. 5.37 – *Equaliseur adaptatif stéréophonique sous* Max/MSP *avec deux fonctions de transfert non corrélées.*

Pour des signaux stéréophoniques, on peut utiliser deux équaliseurs différents, un par canal, *cf.* Piste n°82-CD1 ♫♪ obtenu à partir de Piste n°6 ♫♪. Dans ce cas, le contrôle des équaliseurs peut être indépendant, croisé ou encore corrélé. Le contrôle indépendant signifie que chaque voix

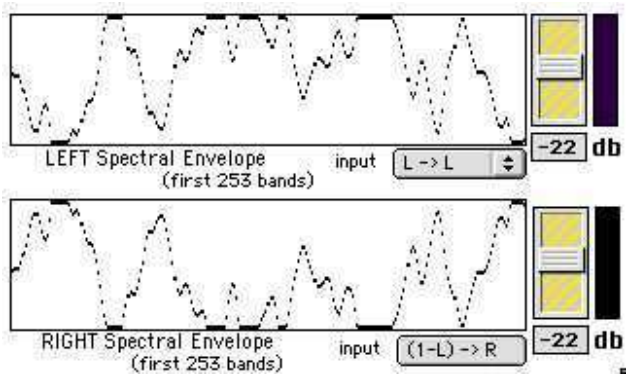


FIG. 5.38 – *Equaliseur adaptatif stéréophonique sous* Max/MSP *avec deux fonctions de transfert corrélées : un effet de panoramisation spectrale apparaît.*

contrôle son propre équaliseur (*cf. fig.* 5.37). Le contrôle croisé consiste à faire en sorte que le signal du canal gauche contrôle l'équaliseur de droite, et réciproquement. Enfin, le contrôle corrélé consiste à donner deux formes complémentaires aux filtres gauche et droit à partir d'une seule forme d'onde (*cf. fig.* 5.38). De cette manière, la répartition totale en fréquence est partagée entre le canal de gauche et le canal de droite : on modélise un déplacement des composantes sinusoïdales de gauche à droite selon une loi linéaire. Il en résulte un effet de panoramisation spectrale adaptative, panier de fréquence par panier de fréquence.

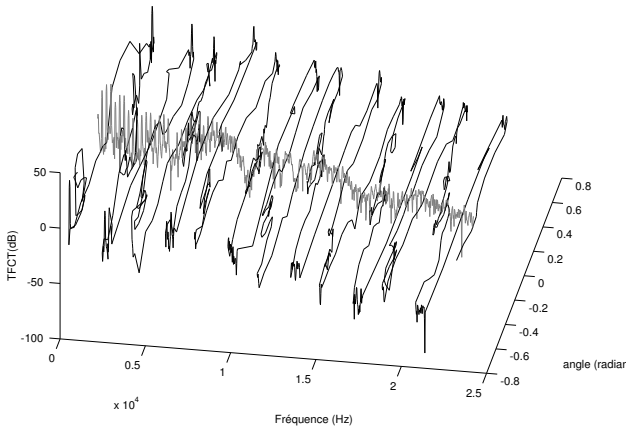


FIG. 5.39 – *Représentation Fréquence-Espace d'une panoramisation spectrale adaptative : les paniers de fréquence de la TFCT (en gris) d'un signal sont positionnés différemment sur le plan spatial (en noir).*

Instrument "NoiSonic" : il s'agit d'un instrument numérique que nous avons développé, basé sur l'équaliseur adaptatif et la synthèse croisée. On utilise un signal comme forme d'onde alimentant l'équaliseur adaptatif. D'autre part, on utilise un bruit (rose ou blanc) comme signal à filtrer. La forme de l'équaliseur est tronquée, de manière à ne laisser passer qu'une petite partie du spectre. Concernant le contrôle, il porte sur le temps d'interpolation, l'étirement vertical et la translation verticale de la forme du filtre. On obtient alors un instrument de synthèse de bruits fort sympathique, pouvant simuler le bruit de la mer, le vent, des cliquetis rapides. Un exemple sonore est donné Piste n°83-CD1 ♫♪. On peut aussi accompagner un son filtré par une sorte de double bruité (son ombre ?).

Vidéo Une démonstration vidéo préparée pour la soutenance est disponible (*cf.* B.3).

5.6.1.v) Filtres en peigne universel adaptatif (NTR, TR)

Nous rappelons la formule de la fonction de transfert du filtre en peigne universel :

$$H(z) = \frac{\beta + g_x \ z^{-L}}{1 - g_y \ z^{-L}} \tag{5.63}$$

Selon les valeurs d'amplitude des coefficients β, g_x et g_y, le filtre en peigne est à réponse impulsionnelle infinie ($g_y = 0$) ou finie ($g_x = 0$). Les paramètres que l'on contrôle de manière adaptative sont le temps de délai L et les trois coefficients d'amplitude β, g_x and g_y. La mise en œuvre temps-réel a été réalisée à l'aide de *Max/MSP*, sous forme d'un patch autonome, après avoir effectué des tests avec l'objet `vst~` et le filtre en peigne des *GRM Tools* [Favreau, 2001]. Nous avons réalisé deux filtres en peigne d'ordre supérieur à 1 (allant jusqu'à 3) : en parallèle et en cascade. Celui en parallèle met en œuvre le filtre en peigne traditionnel, d'ordre 2 ou 3, en utilisant 3 filtres en peigne, de longueurs L, $2L$ et $3L$, et des coefficients bien choisis : (β, g_x, g_y, L) pour le premier filtre, $(0, g_x, 0, L)$ pour le deuxième et le troisième filtre. L'équation aux récurrences correspondante est donc :

$$y_p(n) = \beta x(n) + g_x x(n-L) + (g_x)^2 x(n-2L) + (g_x)^3 x(n-3L) + g_y y_p(n-L) \tag{5.64}$$

avec la fonction de transfert $H_1(z) = \frac{\beta + g_x \ z^{-L} + (g_x)^3 \ z^{-2L} + (g_x)^3 \ z^{-3L}}{1 - g_y \ z^{-L}}$. Dans le cas des filtres en peigne en cascade, si l'on utilise les même coefficients pour les trois filtres, cela revient à utiliser pour fonction de transfert :

$$H_2(z) = \left(\frac{\beta + g_x \ z^{-L}}{1 - g_y \ z^{-L}} \right)^3 \tag{5.65}$$

qui est bien différent de $H_1(z)$, pour preuve son équation aux récurrences :

$$\begin{aligned} y_c(n) &= \beta^3 x(n) + 3\gamma^2 g_x x(n-L) + 3\gamma (g_x)^2 x(n-2L) + (g_x)^3 x(n-3L) \\ &+ 3 g_y y_c(n-L) - 3 (g_y)^2 y_c(n-2L) + (g_y)^3 y_c(n-3L) \end{aligned} \tag{5.66}$$

Concernant les termes en $x(n-kL)$, leurs coefficients sont différents. De plus, il existe des termes en $y_c(n-kL)$ pour la version en cascade. Ce qu'il en résulte, au niveau sonore, c'est que la version cascade fait beaucoup mieux ressortir les harmoniques de la fondamentale du filtre en peigne, avec un son bien plus cinglant, tandis que la version en parallèle est plus supportable. Notons un réglage amusant : en utilisant la version en cascade, avec $g_y > 0.9$, un temps de l'ordre de 50 ms et un contrôle adaptatif, on obtient alors un pseudo-moustique (*cf.* Piste n°84-CD1 ♫♪, d'après la référence Piste n°6 ♫♪).

5.6.1.vi) Compresseur spectral adaptatif

Le principe du compresseur spectral est de se placer dans le domaine fréquentiel pour appliquer une modification d'amplitude aux composantes fréquentielles en fonction de leurs amplitudes. C'est un ensemble de compresseurs, chacun appliqué à un groupe de paniers de fréquence de la TFCT. Le *plug-in* **Contrast** des *GRM Tools* [Favreau, 2001] en est le premier exemple : il modifie le spectre après l'avoir séparé en trois composantes. Deux seuils d'amplitude (en dB) sont fixés par l'utilisateur, et le spectre est séparé en composantes hautes (supérieures au seuil supérieur), moyennes (entre les deux seuils) et basses (en dessous du seuil inférieur). L'utilisateur contrôle le gain appliqué à chaque groupe, indépendamment. Le gain et les seuils sont donc constant pour toutes les fréquences, mais variable dans le temps. Nous proposons plusieurs généralisations de cet effet, sous le nom de compresseur spectral adaptatif :

– la première généralisation est l'utilisation de plus de deux seuils (et donc plus de 3 composantes), ou d'un seul, *cf. fig.* 5.40 ;
– la deuxième généralisation consiste à faire varier automatiquement le gain ou les seuils dans le temps, en fonction de descripteurs $\mathcal{D}_i(t)$;
– la troisième généralisation consiste à utiliser des courbes de seuils et des courbes de gain qui ne soient plus constants en fréquences. On utilise dans ce cas un méta-descripteur, tel que l'enveloppe spectrale *cf. fig.* 5.41.
– la quatrième généralisation est l'utilisation d'une fonction de conformation plus ou moins abrupte : la courbe de changement de gain est lissé par une fonction de forme "pyramide" (segments de transition), de forme "colline" (sinusoïde de transition) ;
– la cinquième généralisation est l'utilisation d'une combinaison linéaire de seuils fixes et de seuils dépendant de l'enveloppe spectrale, avec un coefficient $\gamma = C_c(t)$, *cf. fig.* 5.43.

Cet effet permet d'effectuer un filtrage modifiant le poids des composantes spectrales de manière fine.

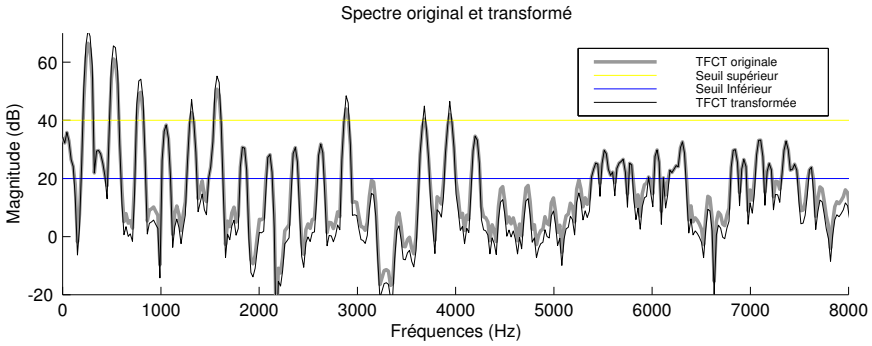


FIG. 5.40 – *Modification du spectre par le compresseur spectral : les seuils sont donnés à* 10 *et* 40 dB*, avec un gain de* 4.5 dB *au-dessus du seuil supérieur, un gain nul entre les deux seuils, et un gain de* -4.5 dB *en dessous du seuil inférieur.*

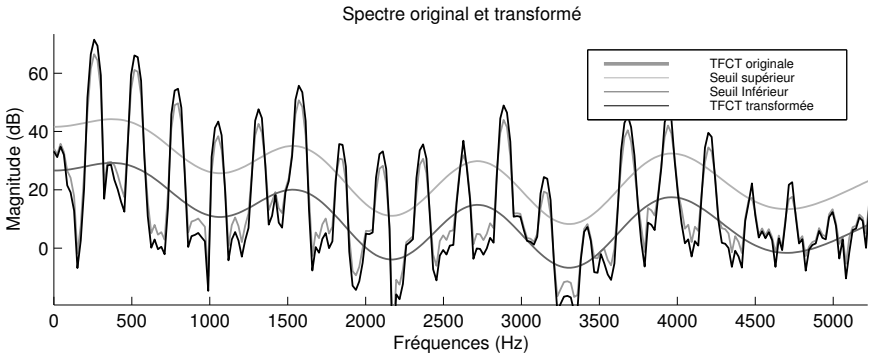


FIG. 5.41 – *Modification du spectre par le compresseur spectral : les seuils sont donnés par translation de l'enveloppe spectrale (calculée à partir du Cepstre) de* -5 *et* $+10$ dB*, avec un gain de* 5 dB *au-dessus du seuil supérieur, un gain nul entre les deux seuils, et un gain de* -5 dB *en dessous du seuil inférieur.*

5.6.1.vii) Wha-wha et autres filtres vocaliques adaptatifs (TR)

Nous avons expérimenté des filtres adaptatifs à partir de filtres vocaliques, d'un filtre photosonique (*cf.* [Arfib *et al.*, 2002a] pour une description de ces filtres), en temps réel sous *Max/MSP*. Pour éviter des phénomènes de parasites et de saturation due à l'utilisation d'interpolation de filtres

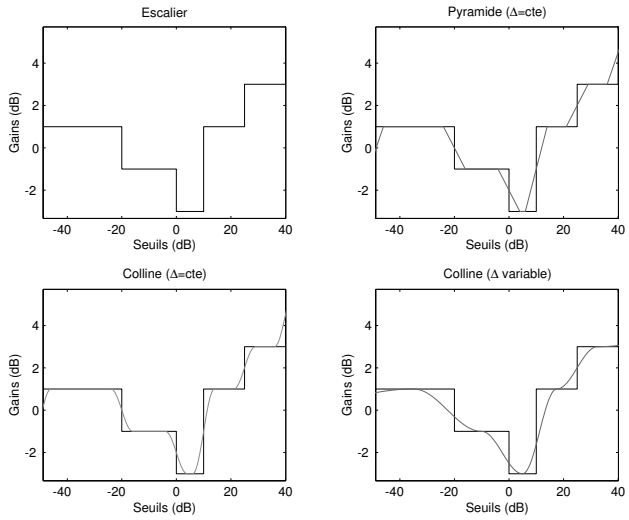


FIG. 5.42 – *Modification du spectre par le compresseur spectral : fonctions de conformation du gain en fonction des seuils, en escalier, en pyramide, en "colline" avec Δ constant et variable ($1/N$ fois la demi largeur entre deux seuils, de chaque côté variable, avec $N \geq 2$).*

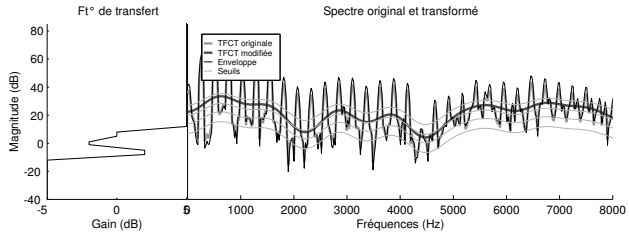


FIG. 5.43 – *Modification du spectre par le compresseur spectral : fonction de conformation du gain en fonction des seuils, eux-mêmes moyenne entre une constant et l'enveloppe spectrale ($\gamma = 0.7513$).*

sans précautions pour assurer leur stabilité, il faut utiliser par exemple des filtres en treillis. D'une certaine manière, ce filtrage adaptatif est une généralisation de la wha-wha sensitive (*cf.* sec. 3.6.1), qui est, rappelons-le, un effet wha-wha déclenché par une détection d'attaque. Pour ces différents filtres vocaliques adaptatifs, la généralisation vient de deux aspects : tout d'abord, le paramètre de contrôle n'est pas forcément la détection d'une attaque ; ensuite, on ne se limite pas à l'effet wha-wha, mais on peut utiliser d'autres voyelles, à l'aide de la représentation du timbre vocal (via ses formants) dans le triangle vocalique [Slawson, 1985].

Remarque : l'effet auto-wha, puisqu'il est piloté par une fonction sinusoïdale, peut lui aussi devenir adaptatif, tout comme le vibrato adaptatif (*cf.* sec. 3.6.3) et le trémolo adaptatif (*cf.* sec. 3.2.5).

5.6.1.viii) Changement de couleur (NTR)

L'effet de changement de couleur (ou de voyelle) consiste à reconnaître dans l'enveloppe spectrale d'une voyelle, et à la remplacer par une autre (*cf. fig.*5.44). On utilise le modèle source-filtre pour le calcul de l'enveloppe spectrale par la technique du cepstre. Cet effet présenté dans l'article [Verfaille and Arfib, 2001]. Il nécessite deux blocs fonctionnels : le premier reconnaît la voyelle, par des calculs de corrélation entre l'enveloppe spectrale et des enveloppes spectrales de référence, et le second applique une enveloppe spectrale de référence à la source du son analysé.

Les premiers résultats obtenus avec un filtre formantique (plusieurs filtres résonants en cascade)

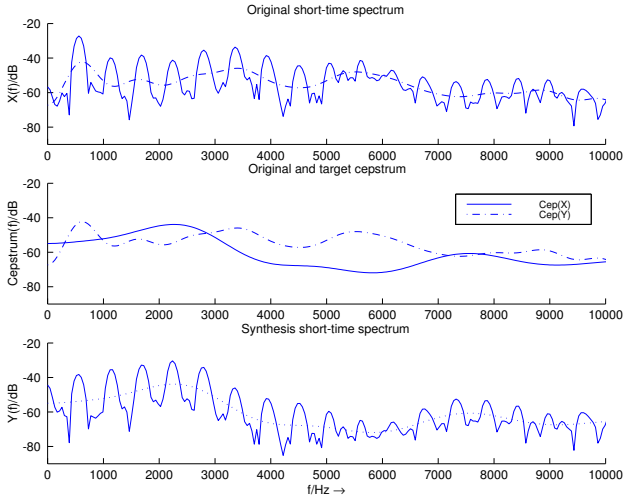


FIG. 5.44 – *Changement de voyelle : TFCT originale, enveloppe spectrale originale, enveloppe spectrale "cible" et TFCT de synthèse.*

constant appliqué sur la source et sur toute la durée de la voyelle sont mauvais, du fait que le filtrage est constant, et s'entend alors comme un filtrage de la source et non comme l'application d'une voyelle. Il faut absolument tenir compte de l'évolution des formants des voyelles lors de la co-articulation consonne–voyelle ou voyelle–consonne (on parle alors de "diphones" [Lienard *et al.*, 1977]). Nous avons donc manuellement segmenté plusieurs signaux de références, de consonne et de voyelle connue et indiquée au programme (sous *Matlab*).

Lorsque l'effet est appliqué avec des filtres variables, le résultat est bien meilleur car on réalise alors une sorte de synthèse croisée entre le son original et les voyelles de référence. L'idée de cet effet nous est venue du jeu qui existe sur la chanson "*Buvons un coup ma serpette est perdu*, ***mais le manche est revenu***". Le but du jeu est de chanter en remplaçant toutes les voyelles par des [a], puis par des [e], et ainsi de suite[3]. Nous avons donc utilisé comme point de départ "*ma la mach' a ravana*" Piste n°13 ♫♪, et avons remplacé les [a] par des suites de voyelles arbitraires. Deux exemples sonores sont donnés, Piste n°85-CD1 ♫♪ pour la suite {[o],[u],[é],[a],[o],[u],[a]}, et Piste n°86-CD1 ♫♪ pour la suite {[i],[o],[a],[e],[u],[é],[i]}.

Nous ne somme pas très satisfait du résultat, notamment sur les consonnes et la coarticulation. Nous pensons que cet effet aurait besoin de nombreuses améliorations. Cela dit, plusieurs mois auraient été nécessaires pour ce seul effet, et le problème ne vient pas tant ici du contrôle que du traitement. Nous avons donc décidé de nous concentrer sur le contrôle automatique et gestuel des effets, au risque de ne pas développer plus cet effet.

5.6.1.ix) Voice Impersonator (TR)

Le système de conversion de voix chantée en temps-réel développé au *Music Technology Group* à Barcelone [Amatriain *et al.*, 2001; Amatriain *et al.*, 2002] réalise la synthèse croisée de deux voix chantées en temps-réel, en conservant des attributs de chaque voix : la dimension temporelle et la hauteur du chanteur de karaoké, et l'enveloppe spectrale (le timbre) du "chanteur idéal". Cette application est l'un des meilleurs exemples d'effets adaptatifs, du fait que les traitements sont très complexes, efficaces et fonctionnent en temps-réel.

[3]Certains spécialistes (des amis que je ne citerai pas) arrivent à utiliser le [ouin] !

5.6.2 Effets adaptatifs sur la phase

5.6.2.i) Chorus et *flanger* adaptatif (TR)

Les effets de chorus et de ***flanger*** se basent sur la modulation de lignes à retard. La modulation se fait selon une loi sinusoïdale : la version adaptative de ces effets peut donc se faire en modulant très légèrement la fréquence de modulation de l'effet, afin de rendre le son un peu plus chaud et l'effet moins systématique, homogène. C'est ce qui est implémenté dans certains réglages de ***Sfx Machine RT***, développé par ***Sound Guys*** [The Sound Guys, 2003]. Dans ce cas, on utilise le RMS comme contrôle de la modulation de fréquence.

5.6.2.ii) Chuchotement adaptatif (NTR)

Nous avons mise en œuvre le chuchotement adaptatif avec le vocodeur de phase (*cf.* sec. 11), en contrôlant la taille du grain d'entrée. Du fait que l'on utilise des grains de taille variable avec un pas constant, le recouvrement des grains de synthèse n'est pas normalisé. il faut alors calculer l'enveloppe de puissance, puis renormaliser échantillon par échantillon le signal de synthèse. De même que pour la robotisation adaptative (*cf.* sec. 5.7.1), il est intéressant d'utiliser un échelle exponentielle et non linéaire pour le contrôle de la taille du grain.

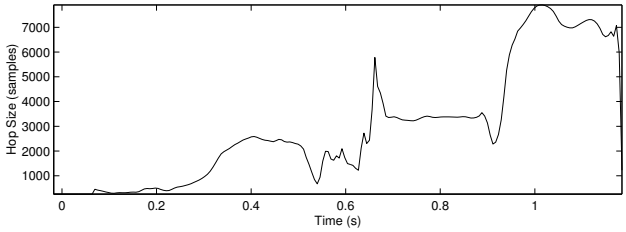


FIG. 5.45 – *Courbe de contrôle pour un chuchotement adaptatif appliqué en* fig. 5.46.

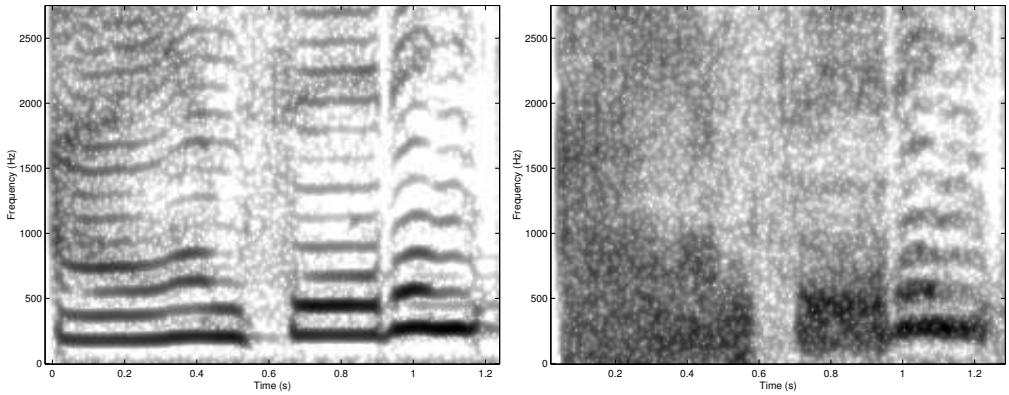


FIG. 5.46 – *Sonagrammes d'une voix chantée (fig. gauche) et de ce son après chuchotement adaptatif (fig. droite) contrôlé par F_0. Pour les plus basses notes, les phases sont aléatoires (petit grain), alors qu'elles sont bien respectées pour les hautes notes (grands grains).*

L'exemple donné utilise pour contrôle la fréquence fondamentale, *cf. fig.* 5.45. La taille de grain varie de 64 à 4096 échantillons. La figure *fig.* 5.46 gauche correspond au sonagramme avant traitement (Piste n°19 ♫♪), et la figure droite au sonagramme après traitement (Piste n°87-CD1 ♫♪) : remarquons que de petites valeurs de la courbe de contrôle, associées à de petites valeurs de la taille du grain, correspondent aux zones où le sonagramme est aléatoire, ce qui implique que le son devienne chuchoté. De grandes valeurs du contrôle correspondent aux parties où l'harmonicité du son réapparaît : le son est bien plus proche de l'original. Cet effet est l'un des trois premiers

que nous avons développé, avec la robotisation adaptative et la dilatation/contraction temporelle adaptative.

5.6.3 Effets adaptatifs sur la structure du spectre

5.6.3.i) Vibrato adaptatif

On considère deux manières d'appliquer un vibrato adaptatif : la première consiste à ajouter un vibrato (dont les contrôles varient) à un son n'en possédant pas, la seconde consiste à modifier les paramètres d'un vibrato existant dans un son.

Ajouter un vibrato (NTR) Le vibrato adaptatif correspond à une transposition modulée par une sinusoïde (modulation de fréquence) avec une fréquence $f_{vib}(k) \in [4;8]$ *Hz* et une profondeur $d_{vib}(t) \in [-\frac{1}{2};\frac{1}{2}]$ ton (au maximum 75 cents, ie. 75% d'un demi-ton), données par deux courbes de contrôles. Il peut se déclencher par exemple en fonction de l'indice de voisement (pour un vibrato automatique) ou en fonction de la fréquence fondamentale (pour un vibrato moins systématique). De même que pour le trémolo adaptatif (*cf.* sec. 5.2.4), on utilise la modulation de phase pour donner la valeur de contrôle. Le facteur de transposition $\rho(k)$ est donné par :

$$\varphi(k) = \varphi(k-1) + 2\pi\ f_{vib}(k)\Delta t \tag{5.67}$$

$$\rho(k) = 2^{d_{vib}(k)\sin\varphi(k)} \tag{5.68}$$

avec la nouvelle hauteur :

$$\mathcal{P}(k) = \rho(k).\mathcal{H}_0(k) \tag{5.69}$$

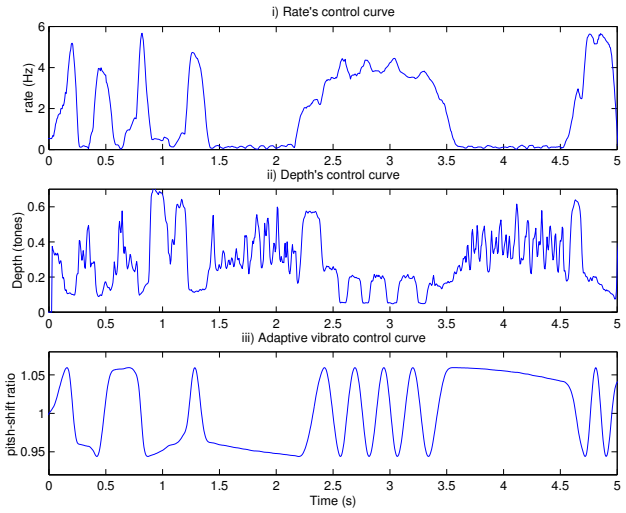


FIG. 5.47 – *Courbes de contrôle d'un vibrato adaptatif : i) fréquence $f_{vib}(k)$, ii) profondeur $d_{vib}(k)$ et iii) le facteur de transposition $\gamma(t_k)$ correspondant.*

Un exemple est donnée Piste n°88-CD1 ♫♪ (avec pour son de référence la Piste n°8 ♫♪), où la profondeur du vibrato dépend de la fréquence fondamentale (seule la note la plus aigue a un vibrato perceptible). On peut construire un vibrato adaptatif plus complexe en utilisant un détecteur de transition [Rossignol *et al.*, 1998b], afin de n'appliquer le vibrato qu'à partir du moment où le son est stable (portions stationnaires du signal).

5.6.3.ii) Changement de vibrato (NTR)

Afin de modifier un vibrato, il faut procéder en deux phases : d'abord le soustraire au son de départ, puis appliquer le nouveau vibrato (*cf.* sec. 5.3.4). Ceci est très utile, non seulement pour modifier le vibrato, mais aussi pour pouvoir réaliser une dilatation/contraction temporelle en préservant le vibrato [Arfib and Delprat, 1998]. Nous avons donc besoin d'extraire les paramètres du vibrato, à savoir la fréquence $f_{vib}(t)$, la profondeur d_{vib} et la phase à l'origine $\alpha(t)$. Cela peut se faire de plusieurs manières [Rossignol *et al.*, 1998a; Herrera and Bonada, 1998; Arfib and Delprat, 1998]. Ensuite, on applique une transposition inverse pour "effacer" le vibrato, puis le nouveau vibrato peut être appliqué comme vu précédemment, à l'aide de la transposition. Une perspective serait d'automatiser le procédé, et de pouvoir appliquer un vibrato adaptatif.

5.6.3.iii) Dual detune (NTR)

Nous avons appelé *dual detune* l'effet qui consiste à ajouter deux versions transposées du signal, l'un vers les hautes fréquences, l'autre vers les basses fréquences. Le facteur de transposition est petit, ce qui fait l'effet Detune, et contrôlé par le son, ce qui rend cet effet adaptatif. Nous l'avons développé sous *Matlab*. L'amplitude de la transposition est donnée en demi-tons. Un exemple est donné Piste n°89-CD1 ♫♪ (avec pour référence la appliqué à la Piste n°8 ♫♪), avec un ambitus de transposition de ± 1 ton, et contrôlé par le RMS.

5.6.4 Effets adaptatifs sur le spectre et l'enveloppe

Conformation spectrale adaptative

On utilise la conformation spectrale présentée en *cf.* sec. 3.6.4, avec les contrôles présentés avec la conformation adaptative de l'enveloppe spectrale en *cf.* sec. 5.6.1. Le contrôle de cet effet est donné par un méta-descripteur du signal Piste n°8 ♫♪ , par exemple (Piste n°90-CD1 ♫♪) la somme cumulative de l'enveloppe spectrale. Les transformations du son peuvent être telles que cet effet pourrait être classé parmi ceux qui modifient plusieurs paramètres perceptifs du son ; en effet, on peut ne reconnaître ni le timbre, ni la hauteur, ni même parfois la durée du son (qui est objectivement conservée), lorsque les changements sont trop importants.

Conservation de l'enveloppe Remarquons que si l'on utilise une fois de plus le modèle source-filtre, on peut conserver l'enveloppe spectrale en appliquant la conformation spectrale adaptative sur la source puis en appliquant le filtre au résultat. Ceci permet d'obtenir des sons dont seul le spectre est modifié par conformation, ce qui ne rend pas le son forcément plus reconnaissable, *cf.* Piste n°91-CD1 ♫♪ , d'après le son de référence appliqué à la Piste n°8 ♫♪ .

Décorrélation enveloppe/spectre : On peut appliquer l'un des traitements de décalage, de changement d'échelle, de conformation aussi bien sur le spectre que sur l'enveloppe. Toutes les combinaisons sont possibles. Le fait de travailler sur les effets de synthèse croisée, de *morphing*, d'interpolation spectrale nous a donné beaucoup d'idées de transformations, comme l'illustrent les effets de conformation de l'enveloppe spectrale d'une part, du spectre d'autre part, et des deux à la fois. Nous avons trouvé là une source d'inspiration, très productive en terme d'outils de transformations musicales.

5.7 Effets adaptatifs portant sur plusieurs paramètres

Dans les sections précédentes, nous avons présenté des effets modifiant principalement une dimension perceptive. Certains, tel le *dual detune* ou la conformation spectrale adaptative modifient aussi la hauteur dans certains cas. Nous présentons maintenant des effets qui modifient

systématiquement plusieurs dimensions perceptives : la robotisation adaptative, le ré-échantillonnage adaptatif, le brassage adaptatif, la martianisation, le changement de prosodie, la modulation en anneau adaptative avec et sans préservation de l'enveloppe spectrale, la transposition adaptative sans conservation des formants et enfin la panoramisation-octaviation adaptative.

5.7.1 Robotisation adaptative (NTR)

La robotisation a été présentée en sec. 3.7.1. Rappelons qu'elle consiste à transformer une voix parlée ou chantée en un train d'ondes, ceci à l'aide d'une approche granulaire où chaque grain voit ses phases mises à zéro, et les échantillons entre deux grains mis à zéros. Les deux contrôles dont on dispose sont la taille du grain L_W et le pas de synthèse R_S. Nous illustrons le contrôle de la hauteur (à l'aide de R_S) par des exemples sonores basés sur les Piste n°15 ♫♪ et Piste n°10 ♫♪. On peut modifier totalement l'intonation et plus généralement la prosodie (*cf.* Piste n°21-CD2 ♫♪, et Piste n°22-CD2 ♫♪ pour une version harmonisée) et changer ainsi l'expressivité de la voix, ou conserver la hauteur originale (*cf.* Piste n°23-CD2 ♫♪ et *fig.* 5.48). C'est d'ailleurs lors de l'étude de la robotisation adaptative que nous avons pris conscience de l'intérêt d'un effet de changement de prosodie (*cf.* sec. 5.7.5), lorsque nous avons utilisé une courbe de contrôle différente de la fréquence fondamentale (*cf.* Piste n°24-CD2 ♫♪).

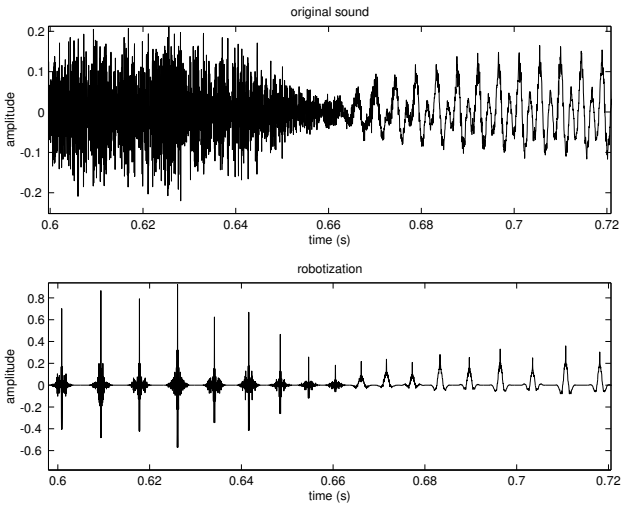


FIG. 5.48 – *Robotisation adaptative contrôlée par F_0 avec un grain de 256 échantillons : forme d'onde d'origine (*fig. *inférieure) et forme d'onde du son après une robotisation ayant conservé la hauteur (*fig. *supérieure).*

Ces possibilités de modification de la prosodie ne sont strictement vraies que pour de petites fenêtres. En effet, nous avons montré (*cf.* sec. 3.7.1) que pour de grandes fenêtres, un second effet apparaît : la hauteur originale réapparaît, du fait des pics secondaires, en plus de la hauteur imposée par le contrôle (pics primaires). Le résultat ressemble à une synthèse croisée entre le son original et une version robotisée adaptative. On peut alors aussi jouer sur cette taille de grain, de manière à retrouver par moments la voix originale en plus de la robotisation.

5.7.2 Ré-échantillonnage adaptatif (NTR)

Le ré-échantillonnage d'un signal permet d'obtenir un signal de longueur et de hauteur différentes (*cf.* sec. 5.7.2). Si l'on raccourcit le signal d'un facteur 2, on transpose aussi d'un facteur 2, soit d'une octave vers le haut. Si maintenant on utilise un facteur de ré-échantillonnage variant dans le temps, on va alternativement allonger et raccourcir le signal, comme lors du changement d'échelle temporelle adaptatif (*cf.* sec. 5.3), mais en modifiant simultanément la hauteur du signal. Cet

effet, dans une version non adaptative contrôlée gestuellement, est couramment utilisé notamment en techno et en rap, par la lecture à vitesse variable d'un disque vinyle, et plus récemment d'un CD. Avec un contrôle adaptatif, la manière de procéder est la suivante : à partir du signal de contrôle de fréquence d'échantillonnage F_c inférieure à F_e, celle du signal sonore, on calcule par interpolation cubique un signal de contrôle à la même fréquence d'échantillonnage que le signal sonore F_e. Ensuite, on calcule un second signal résultant de la somme des échantillons du signal de contrôle $c(t)$ comme suit :

$$c_2(t) = \sum_{k=1}^{t} c(k) \tag{5.70}$$

Ce nouveau signal de contrôle correspond aux nouveaux instants auxquels les échantillons de $x(t)$ se trouvent, et remplacent les temps d'échantillonnage originaux $n/F_e, n \in [1; t]$. On calcule enfin le signal $y(t)$ résultant de ce ré-échantillonnage, par interpolation cubique, du signal $x(t)$ avec pour temps initiaux $c_2(t)$ et pour temps finaux $n/F_e, n \in [1; t_{fin}]$.

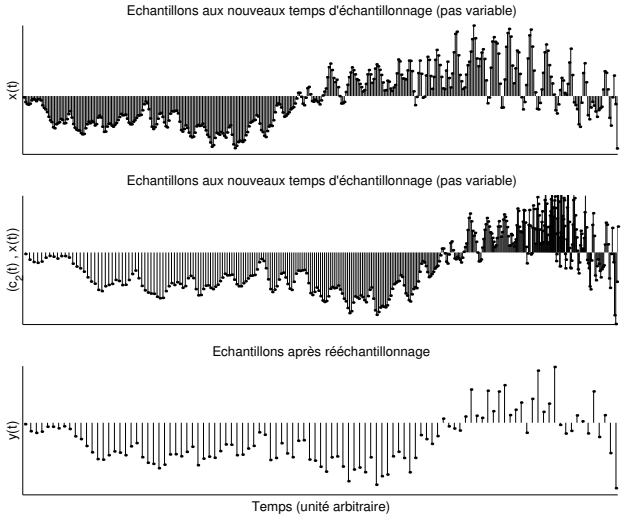


FIG. 5.49 – *Ré-échantillonnage adaptatif : signal original $x(t)$, signal après changement des temps $(c_2(t), x(t))$, signal ré-échantillonné $y(t)$.*

Trois exemples sont donnés à partir de la voix de Pierre Schaeffer Piste n°15 ♫♪, le premier (Piste n°25-CD2 ♫♪) avec le RMS conformé en facteur de ré-échantillonnage variant entre 0.5 et 1.4, le deuxième (Piste n°26-CD2 ♫♪) variant entre 0.25 et 2, et le troisième (Piste n°27-CD2 ♫♪) variant entre 0.24 et 4.

5.7.3 Brassage adaptatif – réorganisation temporelle

Le brassage adaptatif est basé sur des méthodes de synthèse granulaire. Il est légèrement différent de l'écho granulaire adaptatif avec un temps de retard variable (*cf.* sec. 5.5.1), puisque la courbe de contrôle donne la position du grain à lire. Ainsi, le son n'est pas répété, mais lu en avant et en arrière, à vitesse variable, et sans doute plusieurs fois si la courbe monte et descend plusieurs fois. Il s'agit d'une sorte de balayage temporel adaptatif. On peut réorganiser temporellement le son avec une méthode temporelle, granulaire (habituelle), ou par le vocodeur de phase afin de conserver le déroulement de phase et éviter la modulation d'amplitude due au fait que les grains ne sont pas ajoutés-superposés de manière synchrone à la hauteur. Deux exemples sont données, l'un sur la voix Piste n°16 ♫♪ (*cf.* Piste n°28-CD2 ♫♪), un autre sur un son de flûte Piste n°7 ♫♪ (*cf.* Piste n°29-CD2 ♫♪) et un dernier sur le son électroacoustique de glissé de cordes Piste n°9 ♫♪ (*cf.* Piste n°30-CD2 ♫♪).

5.7.4 Martianisation (NTR)

La martianisation est un effet transformant une voix humaine en une suggestion de ce que pourrait être une voix de martien. Le principe consiste à utiliser l'algorithme de vibrato adaptatif en dehors de ses plages de valeurs habituelles, avec une fréquence $f_{vib} \in [0; 14]$ Hz et une profondeur maximale de l'ordre de ± 1 octave au lieu de $\pm 1/2$ ton. Cet effet donne de tellement grandes variations de hauteur que le système auditif n'arrive plus à les suivre, et compose plutôt un timbre étrange, que nous qualifions de "martien" en référence aux voix habituellement utilisées dans les films de science fiction. Pour des variations trop grandes ou trop rapides de hauteur, on peut perdre complètement le sens du message original (Piste n°15 ♫♪ , *cf.* Piste n°31-CD2 ♫♪ .

5.7.5 Changement de prosodie / d'intonation (NTR)

Principe de l'effet La manière de modifier la prosodie (*cf.* sec. 1.3.3) consiste à segmenter le signal en phonèmes, et à altérer la hauteur, l'intensité et la durée des différents phonèmes. Pour la suite de l'exposé de cet effet, nous nous concentrerons sur les modifications de hauteur, ou d'intonation (par soucis de simplicité et non de réalisme), sachant que la même démarche peut être appliquée sur l'intensité (*cf.* la modulation d'amplitude adaptative sec. 5.2.1) et la durée (*cf.* dilatation/contraction temporelle en sec. 5.3.1).

La segmentation du signal peut se faire selon des méthodes usuelles (*cf.* sec. 4.3.6). Ceci dit, nous avons ici utilisé une segmentation manuelle, pour bien distinguer les différents phonèmes, ne disposant pas de système de segmentation parfaitement adapté à la voix parlée. La fréquence fondamentale est extraite pour chaque segment, et analysée en terme de variations (moyenne, variance). Le son est transposé d'un facteur calculé de manière à modifier l'intonation (amplifier ou diminuer les micro-variations).

Représentation de l'intonation Puisque la hauteur du son est calculée et puisque le signal est segmenté, les parties non voisées de la voix que les silences ou bruits ne sont pas pris en compte, et l'on peut calculer la valeur moyenne $< F_0 >$ de F_0 sur chaque segment, ainsi que l'écart à cette moyenne $\delta F_0(t) = F_0(t) - < F_0 > (t)$. Puis on calcule la fréquence fondamentale moyenne $\overline{F_0}$ sur l'ensemble du son par la formule de la moyenne pondérée par l'indicateur d'harmonicité ($\mathbb{1}_{F_0>0}$) :

$$\overline{F_0} = \frac{\sum_{t=1}^{T} F_0 \mathbb{1}_{F_0>0}}{\sum_{t=1}^{T} \mathbb{1}_{F_0>0}} \tag{5.71}$$

La décomposition de la fréquence fondamentale décrivant l'intonation peut donc s'écrire :

$$F_0 = < F_0 > + \delta F_0 \tag{5.72}$$

On calcule ensuite des approximations de ces courbes par des polynômes de degré 1 à 5, qui servent au contrôle gestuel. En effet, si l'on incorpore la fréquence fondamentale moyenne $\overline{F_0}$ et que l'on ajoute deux coefficients compris dans l'intervalle $[-1; 1]$ à l'équation *eq.* (5.72), ils peuvent être contrôlés par le geste pour modifier l'intonation [Arfib and Verfaille, 2003] (*cf.* vidéo DAFx-03 en B.2) :

$$F_0 = \overline{F_0} + \alpha(< F_0 > - \overline{F_0}) + \beta \delta F_0 \tag{5.73}$$

Modification de l'intonation On peut maintenant appliquer des modifications de l'intonation de manière locale, segment par segment, ou de manière globale, ie/ de manière identique pour tout le son. Aplatir l'intonation consiste à supprimer les variations (modulations) de hauteur, au niveau local (sur chaque segment, *cf. fig.* 5.51 gauche) ou global (sur l'ensemble du son, *cf. fig.* 5.51 droite). Ainsi, on obtient des voix plus monotones que l'originale.

Les micro-modulations de hauteur peuvent aussi être renforcées, ou remplacées par d'autres courbes, afin de renforcer la variation (donner plus d'ambitus à la hauteur et donc obtenir une

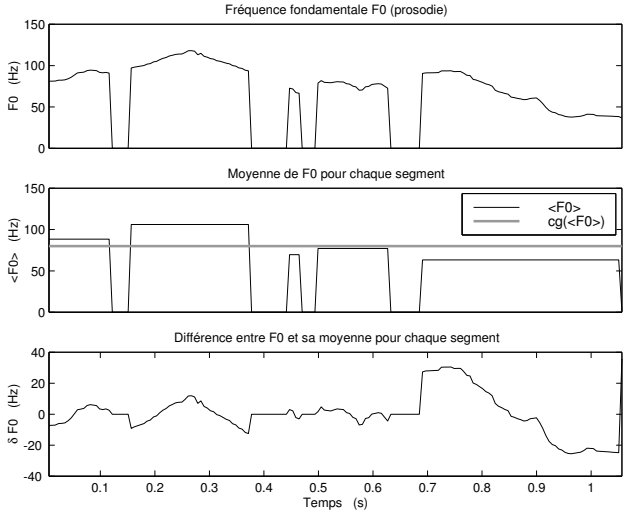


FIG. 5.50 – *Décomposition de l'intonation (en haut) en valeur moyenne $< F_0 >$ (au milieu) et écart à la moyenne δF_0 (en bas).*

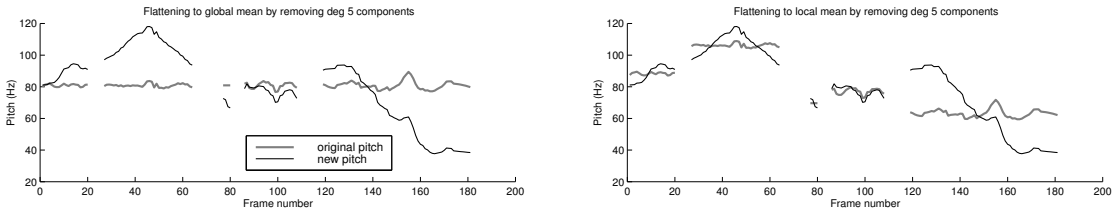


FIG. 5.51 – *A gauche : aplatissement global de l'intonation. A droite : aplatissement local de l'intonation.*

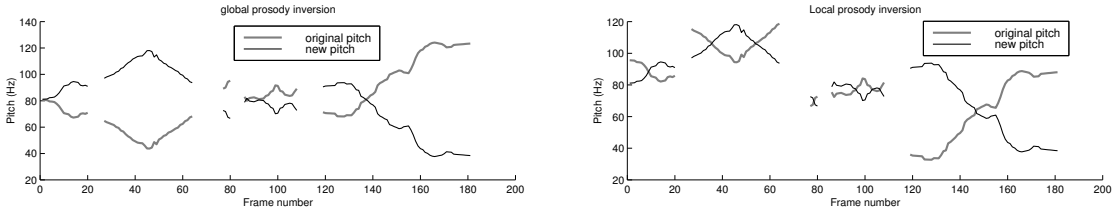


FIG. 5.52 – *A gauche : inversion globale de l'intonation. A droite : inversion locale de l'intonation.*

voix plus "chantante") ou de la rendre complètement différente. Par exemple, inverser le profil d'intonation (inverser la forme de la courbe) sur chaque segment donne des résultats surprenants (au niveau global, *cf. fig.* 5.52 gauche, ou au niveau local *fig.* 5.52 droite). Remplacer la courbe de variation δF_0 par le RMS ou le CGS donne aussi des changements d'intonation intéressants (*cf. fig.* 5.53).

5.7.6 Modulation en anneau adaptative (NTR, TR)

Principe La modulation en anneau consiste à multiplier un signal $x(t)$ par un signal de modulation $x_{mod}(t) = \sin(f_{mod}t)$ constitué d'une seule sinusoïde. Dans le cas où le contrôle de cet effet est adaptatif, c'est la fréquence de modulation qui varie. Ainsi, pour des sons harmoniques, ayant une hauteur donnée, le seul moyen d'obtenir un son modulé ayant lui aussi une hauteur consiste à conserver une relation de fraction entière entre la fréquence fondamentale du signal d'origine et la fréquence de modulation, comme expliqué précédemment en 2.3.1. Par exemple, $f_{mod} = F_0$ ou encore $f_{mod} = \frac{M}{N}F_0$ fonctionne bien. C'est alors la fréquence fondamentale qui contrôle l'effet

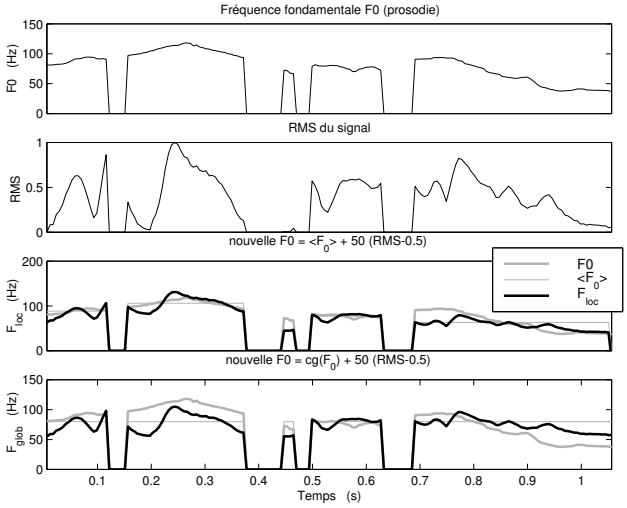


FIG. 5.53 – *Modification de l'intonation : F_0 remplacée par le RMS.*

[Dutilleux, 1991].

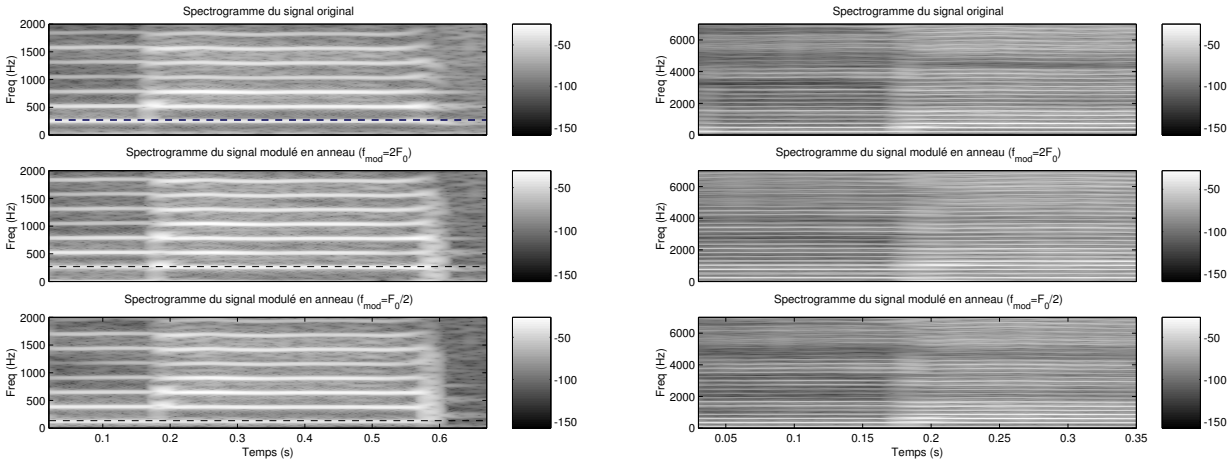


FIG. 5.54 – *Modulation en anneau adaptative pilotée par F_0, $2F_0$ et $F_0/2$. A gauche : zoom sur les premières harmoniques. A droite : vue globale de l'enveloppe spectrale.*
En haut : sonagramme du signal original
Au milieu : modulation en anneau adaptative pilotée par F_0.
En bas : modulation en anneau adaptative pilotée par $f_{mod} = F_0/2$. La hauteur baisse d'une octave, avec des harmoniques paires inexistantes).

Les exemples donnés *fig.* 5.54 correspondent à la modulation en anneau de Piste n°12 ♫♪ ;
- avec $\frac{M}{N} = \frac{1}{2}$: la hauteur change d'octave (*fig.* 5.54, Piste n°32-CD2 ♫♪, la hauteur baisse d'une octave, Piste n°33-CD2 ♫♪, la hauteur baisse de deux octaves) ;
- avec $\frac{M}{N} = k \in \mathbb{N}$ (*fig.* 5.54 et Piste n°34-CD2 ♫♪ pour $k = 1$, Piste n°35-CD2 ♫♪ pour $k = 2$- : la hauteur ne change pas. Par contre, de légères modifications de l'enveloppe spectrale surviennent (*cf. fig.* 5.54).

Dans le cas où l'on utilise un autre paramètre de contrôle, on prend le risque de passer de sons harmoniques à des sons inharmoniques, lorsque l'on passe (par hasard) d'un relation de fraction entière à une relation autre. Dans ce cas, des sons de type cloches peuvent se prêter beaucoup mieux à cet effet : une fois traités, ces sons donneront des sons de cloches dont la hauteur varie au cours du temps !

Modulation adaptative avec conservation des formants On utilise la méthode du cepstre pour extraire l'enveloppe spectrale et la source, afin d'appliquer la modulation en anneau adaptative sur la source uniquement. Ainsi, pour une voix par exemple, on préserve une partie du timbre. la voix traitée est alors plus ressemblante à la voix originale (Piste n°12 ♫♪), puisque les formants sont respectés. Elle est par contre rendu inharmonique si ce n'est pas la fondamentale qui guide la modulante. Dans les deux séries d'exemples que nous donnons, on applique d'abord une modulation adaptative conduite par F_0 sans conservation des formants (Piste n°36-CD2 ♫♪ pour le premier exemple et Piste n°38-CD2 ♫♪ pour le second), puis avec conservation des formants : Piste n°37-CD2 ♫♪ pour le premier exemple, avec une quéfrence de coupure de 538 ; Piste n°39-CD2 ♫♪ pour le second exemple avec une quéfrence de 538 et Piste n°40-CD2 ♫♪ avec une quéfrence de 1077, qui respecte mieux le timbre, du fait que la courbe d'enveloppe utilisée est très proche de la forme du spectre de raies.

5.7.7 Modulation (en anneau) spectrale et adaptative

On applique le principe du trémolo spectral adaptatif (*cf.* sec. 5.2.6) avec des fréquences du domaine audible. Chaque panier de fréquence se voit donc appliquer une modulation en anneau, de fréquence différente et variable dans le temps. Il est ici important d'appliquer une fenêtre à la synthèse du fait que le traitement est fortement non linéaire. En effet, la somme des modulations en anneau des paniers de fréquences de la TFCT ne forme pas un grain valant 0 aux extrémités, ce qui engendre des hautes fréquences lors de l'ajout-superposition. Le résultat sonore est aussi intéressant que celui du trémolo spectral adaptatif, *cf.* Piste n°41-CD2 ♫♪, et très coûteux lui aussi en temps de calcul (Piste n°12 ♫♪ de référence).

5.7.8 Transposition adaptative sans conservation des formants (NTR)

Il s'agit du traditionnel effet de transposition, par exemple par vocodeur de phase, lorsqu'on n'apporte pas de correction des formants par dilatation/contraction de l'enveloppe spectrale. C'est un artefact peu intéressant si on veut conserver un timbre proche de l'original, ou au contraire très intéressant si on veut s'en éloigner, surtout pour de forts taux de transposition.

5.7.9 Panoramisation et octaviation (NTR)

Nous avons combiné la panoramisation adaptative avec un octaveur. Cet effet est particulier, que nous appelons "pan-octaviation", car il transpose le son (1 ou 2 octaves en dessus ou en dessous) puis effectue une balance entre le son original et le son transposé. Ainsi, l'intensité sonore reste la même, et le son fait des va-et-vient entre le son original et le son transposé. De plus, la panoramisation est adaptative, et les deux sons (original et octavié) sont placés en miroir. Autrement dit, soit $\alpha(t) \in [0;1]$ le contrôle de cet effet : le signal de synthèse comprendra le son original modulé en amplitude par $\alpha(t)$ et placé en $\theta(\alpha(t))$; et le son transposé modulé en amplitude par $1-\alpha(t)$ et placé en $-\theta(\alpha(t))$. Le résultat sonore est très intéressant, comme l'attestent les exemples Piste n°42-CD2 ♫♪ (à partir de la Piste n°16 ♫♪) et Piste n°43-CD2 ♫♪ (à partir de la Piste n°19 ♫♪). Pour ces exemples, nous avons choisi d'utiliser un seul contrôle pour l'octaviation et la panoramisation. Toutefois, on peut utiliser deux contrôles distincts, effectuer plusieurs octaviations, déplacer indépendamment le son original et le son octavié, etc. Le champ de possibilités offertes est évidemment bien plus grand que le champ que nous avons investigué.

5.8 Eléments de réflexion sur les effets adaptatifs

5.8.1 Représentation des liens traitements–perception

Pour représenter les relations complexes entre les traitements appliqués au son et les modifications de la perception de ces sons, nous avons réalisé un diagramme heuristique [Buzan and Buzan,

2003] regroupant la plupart des informations données dans la présentation des effets (*cf. fig.* 5.55). Les liens en rouge représentent les dimensions principales de la perception, les liens en bleu les modifications analytiques induites par le traitement, et les liens en bleu les modifications perceptives induites par le traitement. Les liens en vert concernent les effets adaptatifs, et montrent l'apport (involontaire) de notre démarche dans la complexification de ces liens.

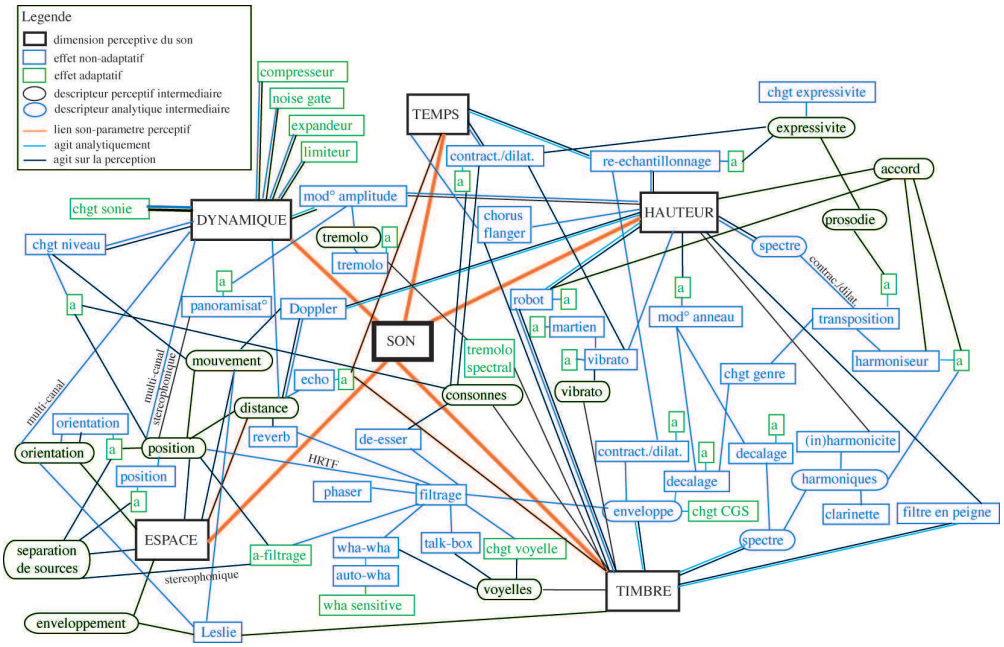


FIG. 5.55 – *Schéma heuristique des effets audionumériques et de leur liens analytiques et perceptifs avec les dimensions du son.*

5.8.2 Remarques sur le contrôle et les descripteurs utilisés

Nous avons volontairement utilisé des contrôles simples et des descripteurs faciles à extraire, tels que l'enveloppe par RMS, le centre de gravité spectrale, la fréquence fondamentale. Les illustrations sonores à partir de ces configurations de contrôle simples révèlent déjà un très grand potentiel des effets audionumériques adaptatifs.

Pour notre utilisation et dans le cadre de développement d'un ensemble logiciel plus complet pour la composition de musique électroacoustique, nous utilisons l'ensemble des descripteurs présentés au chap. 4, ainsi que des contrôles (*cf.* chap. 6) nettement plus complexes et plus fins que ceux utilisés pour les exemples.

5.8.3 Remarques sur les grandes variations des contrôles

Il faut noter que lors des expériences réalisées au cours du développement des effets adaptatifs, nous avons vérifié plusieurs faits perceptifs avérés :

- de grandes variations non sinusoïdales de hauteur provoquent des changements de perception du timbre (martien, *cf.* sec. 5.7.4) ;
- de grandes variations de panoramisation ségrèguent le flux auditif (*cf.* sec. 5.5.2) ;
- de grandes variations de gain modifient le rythme perçu du son (*cf.* sec. 5.2.1) ;
- de grandes variations du spectre (enveloppe, composition spectrale) impliquent des effets de *phasing*, des changements de timbre et de texture.

Ceci nous a conforté dans notre approche. Le fait de proposer un contrôle adaptatif impose de tenir compte de la perception.

Chapitre 6

Contrôles adaptatif et gestuel de l'effet

La loi de l'évolution est la plus importante de toutes les lois du monde parce qu'elle a présidé à notre naissance, qu'elle a régi notre passé, et dans une large mesure, elle contrôle notre avenir.
Yves Coppens [Coppens, 2001]

Sommaire

6.1 Structure du *mapping* : un double contrôle, automatique et gestuel

Comme nous l'avons expliqué dans le chapitre introductif aux effets adaptatifs (chap. 5.1), nous proposons un double contrôle sur les effets audionumériques, à la fois automatique par des descripteurs du son (fréquences audio et sub-audio) et gestuel via un transducteur (fréquences sub-audio). Il nous a semblé utile de hiérarchiser le contrôle en différents niveaux et différentes étapes, afin de pouvoir dire précisément qui contrôle quoi. En effet, le contrôle peut se faire directement sur l'effet, mais aussi sur le *mapping*, c'est-à-dire sur la manière de combiner des descripteurs pour en faire une valeur de contrôle.

Nous pensons que le contrôle de l'effet se prête bien à des variations rapides, comme c'est le cas pour les descripteurs de signal et perceptifs. D'autre part, on considère le *mapping* entre descripteurs et contrôle de l'effet comme un pré-réglage qui donne une sonorité propre à l'effet. Nous pensons que le contrôle offert par les modifications du *mapping* est du domaine sub-audio et se prête bien

mieux à un contrôle gestuel qu'à un contrôle automatique. Cette distinction sous-tend la structure du *mapping* utilisée.

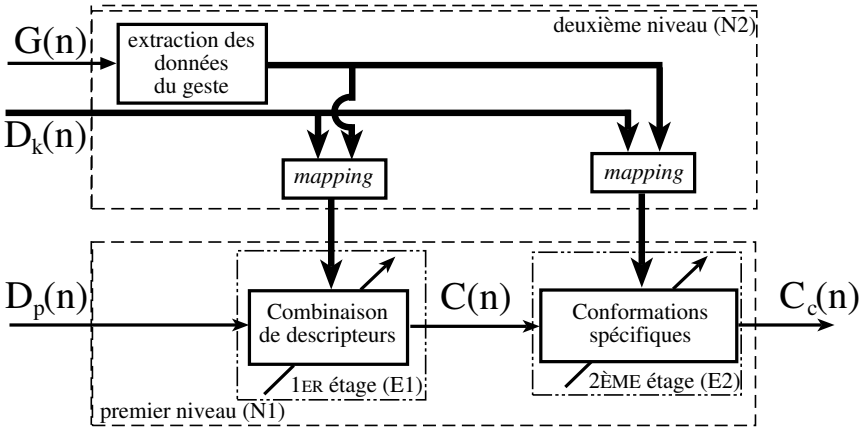


FIG. 6.1 – *Diagramme du* mapping *explicite 1 vers 1 entre descripteurs sonores et contrôles de l'effet.*

Nous proposons de considérer deux niveaux dans le *mapping* global de contrôle (*cf. fig.* 6.1) :
- le premier niveau (N1) concerne les descripteurs, leurs transformations et leur combinaison pour en faire des contrôles (contrôle à des fréquences audio) ;
- le second niveau (N2) concerne la modification des fonctions de conformation et de combinaison, et est modifié principalement par le geste (fréquences sub-audio).

Le premier niveau du *mapping* est constitué de deux étapes : la première étape (E1) consiste à combiner les descripteurs, la seconde étape (E2) consiste à conformer le paramètre obtenu par combinaison afin qu'il respecte des critères de contrôle spécifiques.

6.2 Premier niveau de *mapping* (N1) : contrôle de l'effet par les descripteurs du son

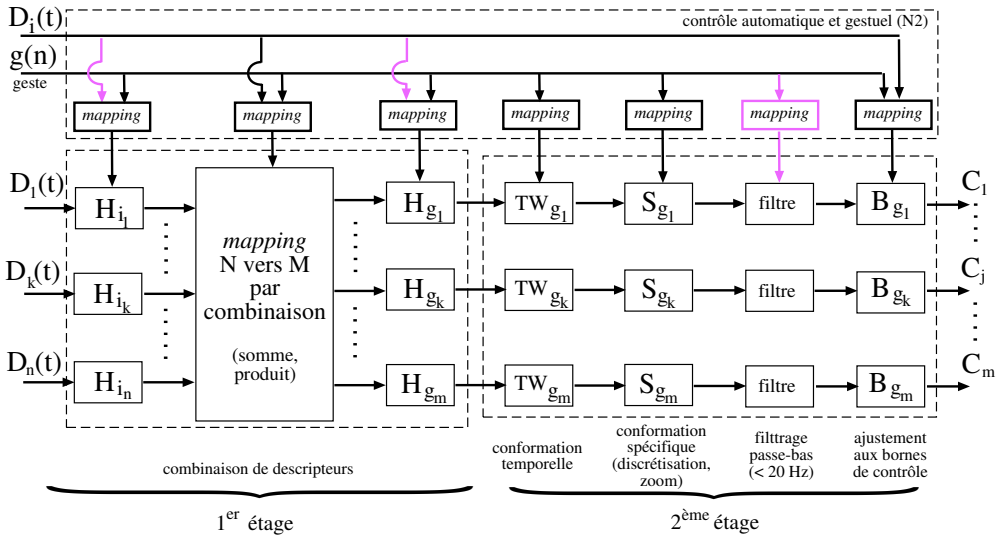


FIG. 6.2 – *Diagramme du* mapping *explicite M vers N entre descripteurs sonores et contrôles de l'effet.*

L'évolution des descripteurs du son est donnée à l'utilisateur par le biais de courbes. Ensuite, différentes transformations sont appliquées à ces courbes pour les transformer en contrôles de l'effet

(différentes fonctions de correspondance), selon deux concepts : le type de connexion et la loi de connexion. Une fois exposés ces concepts, la mise en correspondance entre descripteurs sonores et contrôles de l'effet est décrite, comme la succession de deux étapes :

- **N1-E1** la première étape consiste à transformer N descripteurs en M contrôles (combinaison) ;
- **N1-E2** la seconde étape à ajuster chaque contrôle à des critères (zoom, quantification, ajustement de bornes).

Les types de connexion entre descripteurs sonores et paramètres de contrôles de l'effet peuvent être simples (de 1 vers 1, de 2 vers 2) ou complexes (de N vers M). La loi de connexion peut être linéaire ou non linéaire pour chacune des connexions. Ceci signifie que l'on peut combiner dans un même *mapping* des lois linéaires et des lois non linéaires. Nous utilisons un *mapping* N-vers-M explicite divisé en deux étages principaux (*cf. fig.* 6.1), ceci donnant un ensemble de M contrôles à partir de N descripteurs.

En général, nous avons utilisé pour le premier étage plusieurs *mappings* 3 vers 1 (un pour chaque contrôle) avec plusieurs lois (linéaires et non linéaires) appliquées avant et après la combinaison. La combinaison elle-même est linéaire (pondération puis somme) ou non linéaire (pondération puis produit, division).

Le second étage quant à lui consiste à faire correspondre chaque contrôle à des critères prédéfinis : type de variations (sinusoïdale, filtrage passe-bas), zoom, quantification, ajustement aux bornes. On obtient alors le diagramme général plus détaillé donné en *fig.* 6.2. Dans les sections 6.3 et 6.4, nous entrons dans le détail des des étapes que nous venons de nommer.

6.3 Premier étage (N1-E1) : combinaison de descripteurs

Le premier étage de la mise en correspondance consiste à normaliser chaque descripteur, puis à leur appliquer une série de lois non-linéaires, ensuite à les combiner, puis enfin à appliquer au résultat une autre série de lois non linéaires (*cf. fig.* 6.3).

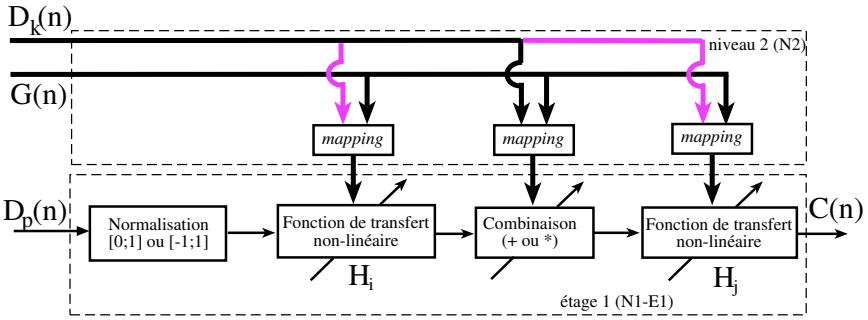


FIG. 6.3 – *Diagramme du premier étage de* mapping *: combinaison de descripteurs.*

6.3.1 Première étape : normalisation

Soient k descripteurs, notés $\mathcal{F}_k(t)$, $t = 1, \ldots, NT$. Nous commençons par les normaliser. Notons $\mathcal{M}^+ = \max$ la fonction "maximum" et $\mathcal{M}^- = \min$ la fonction "minimum", et les extrema du descripteur $\mathcal{F}_k^M = \mathcal{M}^+_{t\in[1;NT]}\mathcal{F}_k(t)$ et $\mathcal{F}_k^m = \mathcal{M}^-_{t\in[1;NT]}\mathcal{F}_k(t)$. On considère deux manières de normaliser les descripteurs : l'une pour pour les descripteurs de signe constant, l'autre pour les descripteurs pouvant changer de signe. La première normalisation consiste à donner pour bornes au descripteur celles de l'intervalle $[0;1]$, par translation et homothétie : on obtient la courbe $\mathcal{N}_k^1(t)$. De cette manière, on s'assure de la plage de variation du paramètre, et on peut donc définir les

autres lois (non linéaires) en fonction de cet intervalle.

$$\mathcal{N}_1(k,t) = \mathcal{N}_1\left(\mathcal{F}_k(t)\right) = \frac{\mathcal{F}_k(t) - \mathcal{F}_k^m}{\mathcal{F}_k^M - \mathcal{F}_k^m} \tag{6.1}$$

La seconde manière de normaliser un descripteur consiste à diviser chaque valeur par le maximum en valeur absolue. On obtient alors la courbe $\mathcal{N}_k^2(t)$ par homothétie. De cette manière, le paramètre conserve le signe du descripteur, ce qui peut être utile. Cependant, contrairement à la normalisation entre 0 et 1, on ne peut s'assurer que le paramètre normalisé atteigne ses deux bornes -1 et 1 (sauf dans le cas où le descripteur vérifie $|\mathcal{F}_k^m| = \mathcal{F}_k^M$).

$$\mathcal{N}_2(k,t) = \mathcal{N}_2\left(\mathcal{F}_k(t)\right) = \frac{\mathcal{F}_k(t)}{\max_{t\in[1;NT]}|\mathcal{F}_k(t)|} \tag{6.2}$$

6.3.2 Deuxième étape : modification par application d'un ensemble de lois

Les modifications que l'on applique aux courbes des descripteurs peuvent porter en amplitude seulement (temps-réel et hors temps-réel) et de manière statique ou dynamique, en temps seulement (hors temps-réel) ou à la fois en amplitude et en temps.

6.3.2.i) Modifications statiques en amplitude

La deuxième étape consiste à appliquer à chaque paramètre normalisé une loi (ou fonction de conformation) non linéaire $\mathcal{H}_{i_k}$ statique. L'indice i_k définit quelle fonction est utilisée, et l'indice i correspond à 1 ou 2, selon la normalisation appliquée (*cf.* sec. 6.3.1).

$$\mathcal{J}(k,i,i_k,t) = \mathcal{H}_{i_k}\left(\mathcal{N}_i(k,t)\right) = \mathcal{H}_{i_k}\left(\mathcal{N}_i\left(\mathcal{F}_k(t)\right)\right) \tag{6.3}$$

Les fonctions de conformation trouvent une place tout à fait logique dans cette étude ; elles servent à modéliser certains comportements physiques, telle la distorsion du signal appliquée par les transistors, comme expliqué en sec. 3.6.4. C'est aussi le principe à la base de méthodes de synthèse, telle la distorsion non-linéaire [Arfib, 1979], mais aussi plus généralement lors de synthèse par modèles physiques de systèmes non-linéaires (telles les non-linéarités dues aux turbulences dans un jet d'air sur un biseau de flûte, par exemple). Ainsi, les effets de compresseur, d'expandeur, de distorsion, entre autres, utilisent des lois non-linéaires, et peuvent être qualifiés d'adaptatifs.

Nous utilisons plusieurs fonctions de conformation $\mathcal{H}_{nom}()$ données par des fonctions affines (*cf. tab.* 6.1 pour un résumé). La première fonction est simple et linéaire (c'est en fait la seule) :

$$\mathcal{H}_{lin}\left(\mathcal{X}(t)\right) = \mathcal{X}(t) \tag{6.4}$$

La deuxième fonction est basée sur la fonction sinusoïde (*cf. fig.* 6.4) ; elle est utilisée pour renforcer la proximité aux bornes :

$$\mathcal{H}_{sine,1}\left(\mathcal{X}(t)\right) = \frac{1+\sin(\pi(\mathcal{X}(t)-0.5))}{2} \tag{6.5}$$

$$\mathcal{H}_{sine,2}\left(\mathcal{X}(t)\right) = \sin(\pi\mathcal{X}(t)/2) \tag{6.6}$$

On peut l'appliquer à la puissance p, ou convoluée à elle-même c fois (*cf. fig.* 6.5 et *fig.* 6.6).
On peut aussi utiliser la fonction réciproque $\sin^{-1}$, afin de rapprocher les valeurs du milieu du segment au lieu de les rapprocher des extrémités :

Il peut être utile de tronquer une courbe (*cf. fig.* 6.8), à l'aide de cette fonction :

$$\mathcal{H}_{trunc}\left(\mathcal{X}(t)\right) = \frac{t_m \mathbb{1}_{\mathcal{X}(t)<t_m} + t_M \mathbb{1}_{\mathcal{X}(t)>t_M} + \mathcal{X}(t)\mathbb{1}_{t_m<\mathcal{X}(t)<t_M}}{t_M - t_m} \tag{6.7}$$

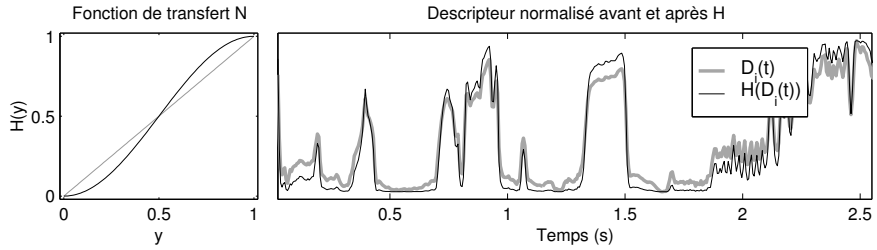


FIG. 6.4 – *Fonction de conformation de type sinusoïde.*

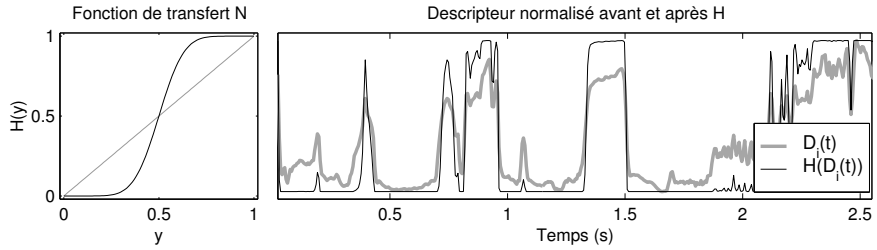


FIG. 6.5 – *Fonction de conformation de type sinusoïde convoluée 3 fois à elle-même.*

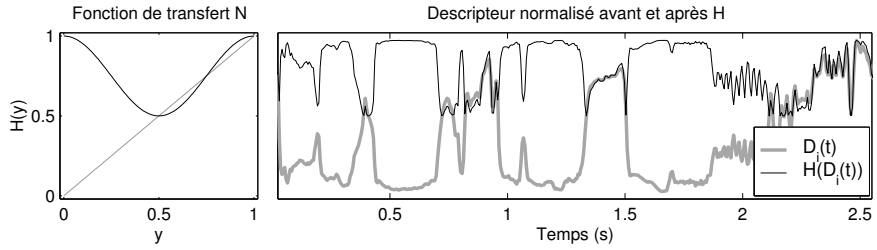


FIG. 6.6 – *Fonction de conformation de type sinusoïde à la puissance 2.*

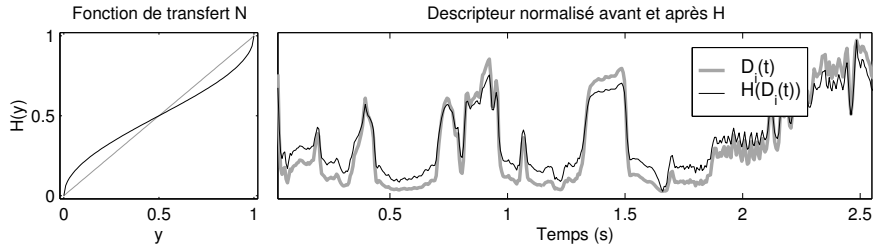


FIG. 6.7 – *Fonction de conformation de type sinusoïde inverse.*

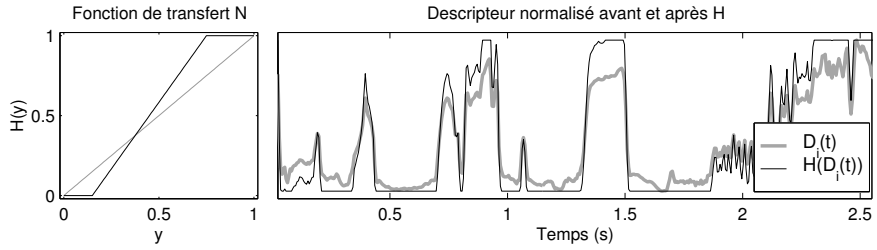


FIG. 6.8 – *Fonction de conformation de type troncature.*

où $\mathbb{1}_a$ est la fonction de Heaveside, ou fonction indicatrice (dont la valeur est 1 si le résultat du test a est Vrai, 0 si le résultat du test est Faux) et $[t_m; t_M] \in [0; 1]$ ou $[t_m; t_M] \in [-1; 1]$ selon la normalisation utilisée au départ. La fonction de troncature permet de sélectionner une portion d'intérêt d'une courbe.

Nous avons utilisé trois fonctions en deux morceaux, principalement pour la dilatation/contraction temporelle.

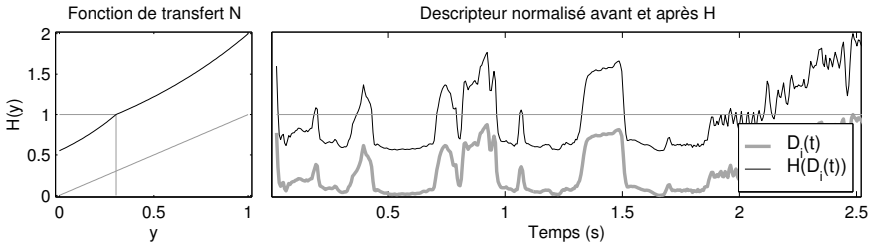


FIG. 6.9 – *Fonction de conformation $\mathcal{H}_{Xpow}$ de type $\rho_{\mathcal{X}(t)}$, en deux zones ($\alpha = 0.3$).*

La première est la fonction puissance (*cf. fig.* 6.9), où s_m (resp. s_M) est porté à une puissance $p = f(\mathcal{X}(t))$:

$$\mathcal{H}_{Xpow}\left(\mathcal{X}(t), \alpha, s_m, s_M\right) = s_m{}^{\left(\frac{\alpha - \mathcal{X}(t)}{\alpha}\right)} \mathbb{1}_{\mathcal{X}(t) \leq \alpha} + s_M{}^{\left(\frac{\alpha \mathcal{X}(t)}{1-\alpha}\right)} \mathbb{1}_{\mathcal{X}(t) > \alpha} \tag{6.8}$$

avec $s_m \leq 1$ le facteur de contraction ou de compression et $s_M \geq 1$ le facteur de dilatation ou d'étirement. Cette fonction est conçue en deux morceaux, spécialement pour les effets de changement d'échelle, afin de pouvoir choisir quelle partie est dilatée, quelle partie est contractée. Le facteur $\alpha \in [0 : 1]$ (resp. $[-1; 1]$) divise le segment $[0; 1]$ (resp. $[-1; 1]$) en deux parties : la partie inférieure $[0; \alpha]$ (resp. $[-1; \alpha]$) qui est contractée, et la partie supérieure $[\alpha; 1]$ qui est dilatée.
La deuxième fonction en deux morceaux est constituée à partir de fonctions puissances, mais cette fois-ci, c'est $\mathcal{X}(t)$ qui est porté à la puissance p :

$$\begin{aligned} \mathcal{H}_{powP}\left(\mathcal{X}(t), \alpha, s_m, \gamma, s_M, p_-, p_+\right) &= \left[s_m + \frac{\gamma - s_m}{\alpha^{p_-}} \left[\mathcal{X}(t)\right]^{p_-}\right] \mathbb{1}_{\mathcal{X}(t) \leq \alpha} \\ &+ \left[\gamma + \frac{s_M - \gamma}{(1-\alpha)^{p_+}} [\mathcal{X}(t) - \alpha]^{p_+}\right] \mathbb{1}_{\mathcal{X}(t) > \alpha} \end{aligned} \tag{6.9}$$

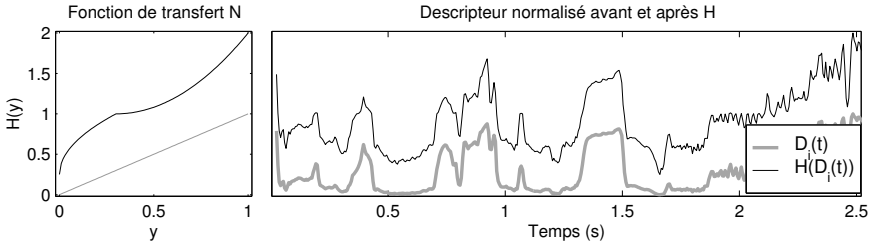


FIG. 6.10 – *Fonction de conformation $\mathcal{H}_{powP}$ de type $\mathcal{X}(t)^p$, en deux zones ($\alpha = 0.3$, $\gamma = 1$, $p_- 0.5 =$, $p_+ = 2$).*

Comme on peut le voir à l'aide des figures *fig.* 6.10, *fig.* 6.11, *fig.* 6.12, *fig.* 6.13 mais aussi avec la comparaison *fig.* 6.14, les puissance supérieures à 1 rapprochent la partie supérieure de son maximum (ie. du maximum de dilatation) et la partie inférieure de 1 (ie. sans dilatation/contraction) alors que les puissances inférieures rapprochent la partie supérieure de 1 (ie. sans dilatation/contraction) et la partie inférieure de son minimum (ie. du maximum de contraction).

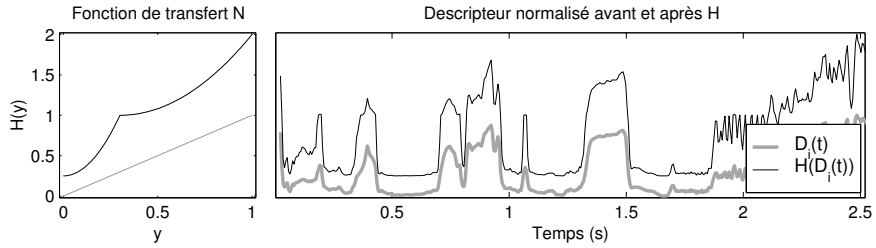


FIG. 6.11 – *Fonction de conformation $\mathcal{H}_{powP}$ de type $\mathcal{X}(t)^p$, en deux zones ($\alpha = 0.3$, $\gamma = 1$, $p_- = 2$, $p_+ = 2$).*

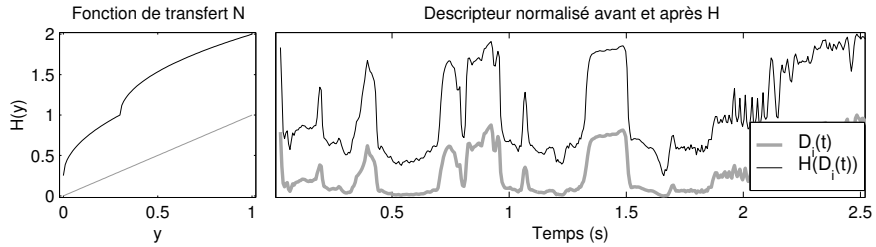


FIG. 6.12 – *Fonction de conformation $\mathcal{H}_{powP}$ de type $\mathcal{X}(t)^p$, en deux zones ($\alpha = 0.3$, $\gamma = 1$, $p_- = 0.5$, $p_+ = 0.5$).*

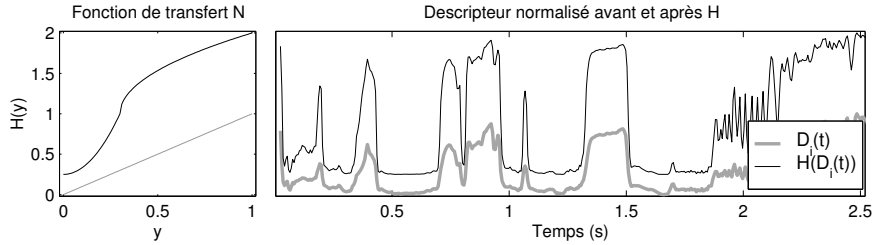


FIG. 6.13 – *Fonction de conformation $\mathcal{H}_{powP}$ de type $\mathcal{X}(t)^p$, en deux zones ($\alpha = 0.3$, $\gamma = 1$, $p_- = 2$, $p_+ = 0.5$).*

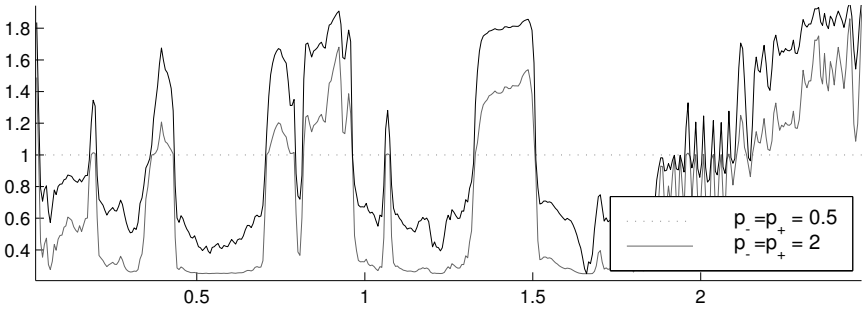


FIG. 6.14 – *Fonctions de conformation $\mathcal{H}_{powP}$, en deux zones ($\alpha = 0.3$, $\gamma = 1$, $p_- = 0.5$, $p_+ = 0.5$) et ($p_- = 2$, $p_+ = 2$).*

La troisième fonction est constituée à partir de fonction sinusoïdes :

$$\mathcal{H}_{bisin}\left(\mathcal{X}(t), \alpha, \gamma, s_m, s_M\right) \quad = \quad \left[s_m + \frac{\gamma - s_m}{2}\left(1 + \cos\left[\left(\frac{\mathcal{X}(t)}{\alpha} + 1\right)\pi\right]\right)\right] \mathbb{1}_{\mathcal{X}(t) \leq \alpha}$$

$$+ \quad \left[\gamma + \frac{s_M - \gamma}{2}\left(1 + \cos\left[\left(\frac{\mathcal{X}(t) - \alpha}{s_M - \alpha} + 1\right)\pi\right]\right)\right] \mathbb{1}_{\mathcal{X}(t)>\alpha} \quad (6.10)$$

Cette fonction bi-sinusoïdale peut-être utilisée pour la dilatation/contraction (*cf. fig.* 6.15), mais aussi pour les autres effets, en changeant les bornes s_m et s_M (*cf. fig.* 6.16). Son grand intérêt est qu'elle définit trois zones d'attraction : pour des valeurs de $\mathcal{X}(t) \in [0; \alpha/2]$, les valeurs de contrôle sont attirées vers s_m. Pour des valeurs de $\mathcal{X}(t) \in [\alpha/2; \frac{\alpha+1}{2}]$, les valeurs de contrôle sont attirées vers γ. Pour des valeurs de $\mathcal{X}(t) \in [\frac{\alpha+1}{2}; 1]$, les valeurs de contrôle sont attirées vers s_M. Autrement dit, pour une dilatation/contraction, on définit ainsi la zone qui n'est pas trop modifiée (autour de α, pour $\gamma = 1$), les autres zones étant dilatées/contractées au maximum.

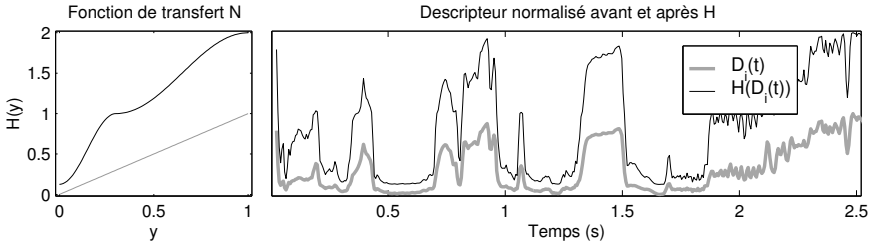


FIG. 6.15 – *Fonction de conformation $\mathcal{H}_{bisin}$ de type bi-sinusoïdale ($\alpha = 0.3$, $s_m = 0.125$, $\gamma = 1$, $s_M = 2$).*

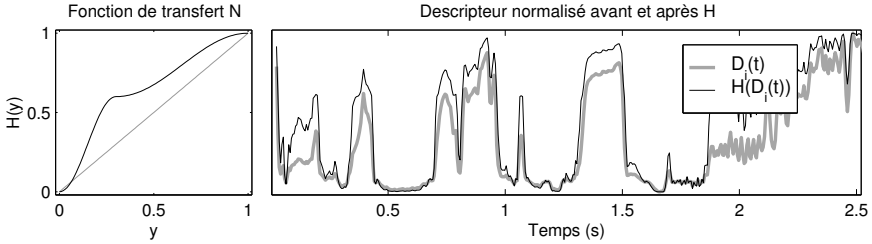


FIG. 6.16 – *Fonction de conformation $\mathcal{H}_{bisin}$ de type bi-sinusoïdale ($\alpha = 0.3$, $s_m = 0$, $\gamma = 0.6$, $s_M = 1$).*

Nous utilisons aussi les fonctions logarithme (*cf. fig.* 6.17) et exponentielle (*cf. fig.* 6.18) :

$$\mathcal{H}_{log}\left(\mathcal{X}(t)\right) = \log_{10}(\alpha + \mu\mathcal{X}) \quad (6.11)$$

$$\mathcal{H}_{exp}\left(\mathcal{X}(t)\right) = 10^{\mu\ (\mathcal{X}-\alpha)} \quad (6.12)$$

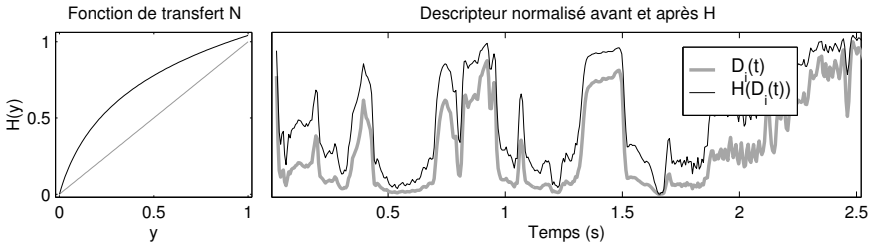


FIG. 6.17 – *Fonction de conformation de type logarithme.*

Le compresseur et l'expandeur sont les exemples les plus anciens d'effet adaptatif, aussi il convient de ne pas oublier de donner leurs fonctions de conformation (*cf. fig.* 6.19 et *fig.* 6.20) :

$$\mathcal{H}_{comp}\left(\mathcal{X}(t)\right) = 10^{\left(\overline{\mathcal{X}}(t).\mathbb{1}_{\overline{\mathcal{X}}(t)<T_c} + (T_c + p.(\overline{\mathcal{X}}(t) - T_c))\mathbb{1}_{\overline{\mathcal{X}}(t)\geq T_c}\right)/20} \quad (6.13)$$

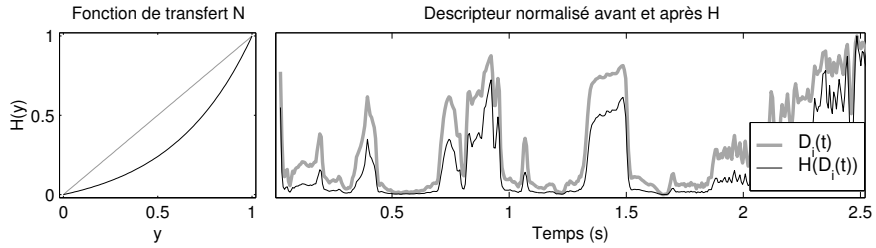


FIG. 6.18 – *Fonction de conformation de type exponentielle.*

avec $\overline{\mathcal{X}}(t) = 20\log_{10}(\mathcal{X}(t))$ le passage en décibels, pour l'application du gain.

$$\mathcal{H}_{expan}\left(\mathcal{X}(t)\right) = 10^{\left(\overline{\mathcal{X}}(t).\mathbb{1}_{\overline{\mathcal{X}}(t)\geq T_c} + (T_c + p.(\overline{\mathcal{X}}(t) - T_c))\mathbb{1}_{\overline{\mathcal{X}}(t)<T_c}\right)/20} \tag{6.14}$$

où T_c et le seuil de compression, T_e le seuil d'expansion, p la pente de compression ($p < 1$) ou d'expansion ($p > 1$).

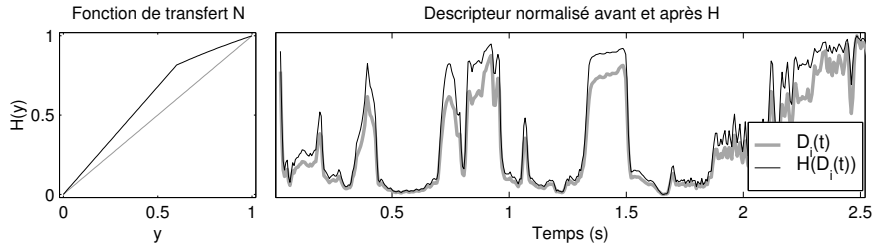


FIG. 6.19 – *Fonction de conformation de type compresseur.*

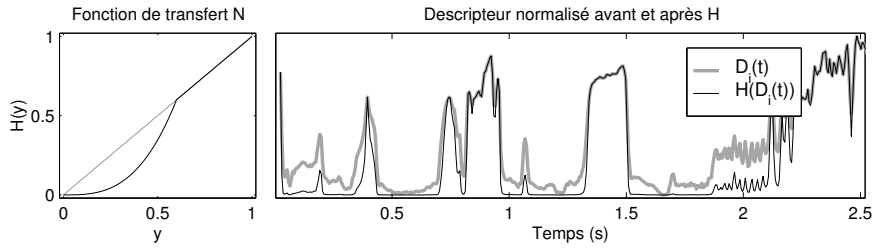


FIG. 6.20 – *Fonction de conformation de type expandeur.*

Enfin, le lissage par moyenne sur une fenêtre de longueur $2o + 1$ est utile (voir *fig.* 6.21 pour $o = 3$ et *fig.* 6.22 pour $o = 10$) :

$$\mathcal{H}_{sm}\left(\mathcal{X}(t)\right) = \frac{\sum_{k=t-o}^{t+o}\mathcal{X}(k)}{2o+1} \tag{6.15}$$

Ces fonctions de lissage sont des systèmes à mémoire, qui permettent de supprimer les trop fortes variations des descripteurs, et peuvent faire double emploi avec le filtre passe-bas appliqué en fin de *mapping* (*cf.* sec. 6.4.4).

On peut aussi vouloir dériver (*cf. fig.* 6.23) à l'aide des différences finies :

$$\mathcal{H}_{deriv}\left(\mathcal{X}(t)\right) = \mathcal{X}(t) - \mathcal{X}(t-1) \tag{6.16}$$

Le descripteur obtenu ne varie plus dans $[0;1]$ (resp $[-1;1]$), aussi il faut le renormaliser : l'intervalle $[-1;1]$ semble le plus approprié, car il concerne le signe. On peut aussi intégrer (*cf. fig.* 6.24) une

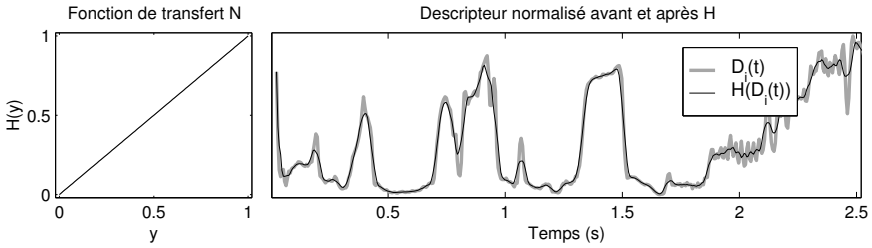


FIG. 6.21 – *Fonction de conformation de type lissage avec une fenêtre de 3 échantillons.*

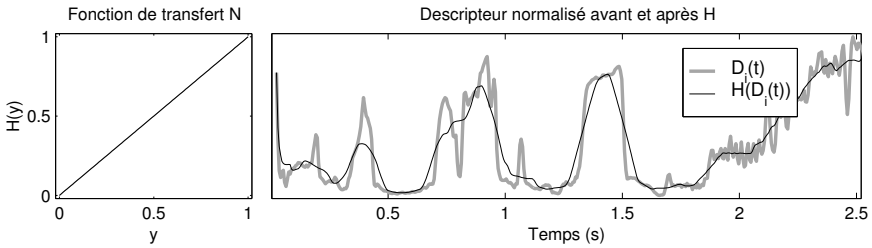


FIG. 6.22 – *Fonction de conformation de type lissage avec une fenêtre de 10 échantillons.*

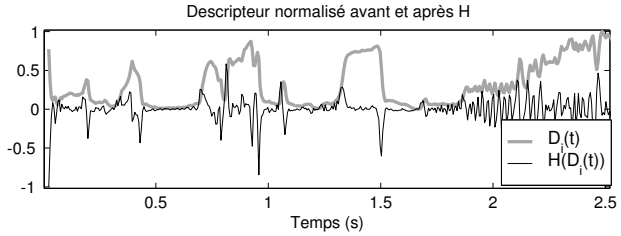


FIG. 6.23 – *Fonction de conformation de dérivation.*

courbe à l'aide de la fonction de somme cumulative :

$$\mathcal{H}_{int}\left(\mathcal{X}(t)\right) = \sum_{k=0}^{t} \mathcal{X}(t) \tag{6.17}$$

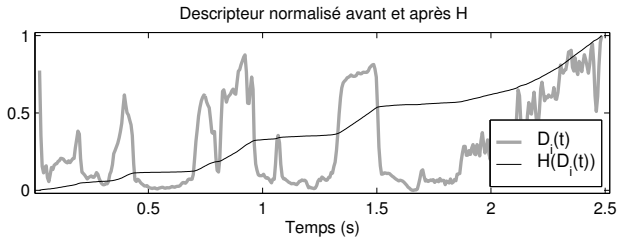


FIG. 6.24 – *Fonction de conformation de type intégration (somme cumulative).*

Une variante consiste à utiliser l'écart à la variation affine :

$$\mathcal{H}_{int,2}\left(\mathcal{X}(t)\right) = \sum_{k=0}^{t} \mathcal{X}(t) - (at + b) \tag{6.18}$$

L'intérêt de cette variante est qu'en choisissant bien la pente de la droite, on obtient un paramètre dont la courbe temporelle n'est plus monotone, mais reste bien corrélé avec le descripteur (ce qui

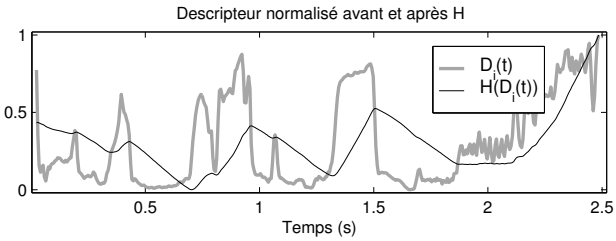


FIG. 6.25 – *Fonction de conformation de type écart à la variation affine.*

est pratique pour que l'effet soit synchronisé avec ce que l'on perçoit des changements dans le son original). On peut aussi appliquer la valeur absolue (*cf. fig.* 6.24) sur une courbe variant dans $[-1;1]$ ou dans un autre intervalle :

$$\mathcal{H}_{abs}\left(\mathcal{X}(t)\right) = |\mathcal{X}(t)| \tag{6.19}$$

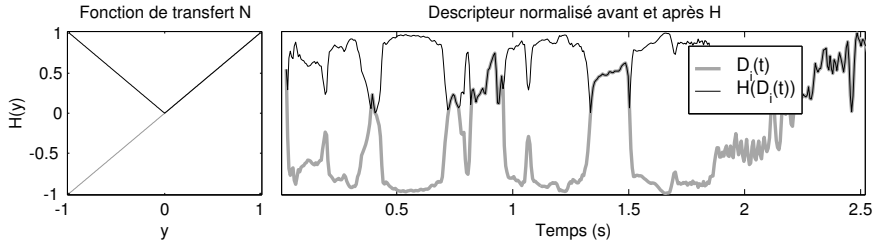


FIG. 6.26 – *Fonction de conformation de type valeur absolue.*

On peut enfin appliquer d'autres fonctions de conformation :

- la combinaison de plusieurs fonctions présentée, par exemple une troncature puis une fonction sinusoïde ;
- une fonction composée comme une bi-courbe : par exemple, sur l'intervalle $[0;0.5]$, on applique la fonction sinusoïde variant dans $[0;0.5]$, et sur l'intervalle $[0.5;1]$ la fonction sinusoïde variant dans $[0.5;1]$.
- une fonction de conformation arbitraire, donnée à la main ;

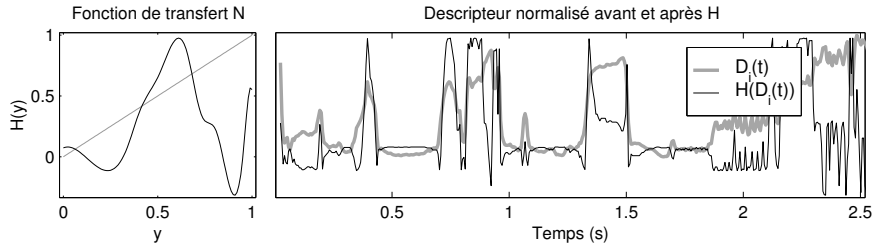


FIG. 6.27 – *Fonction de conformation arbitraire, donnée via une Interface Graphique Utilisateur (premier exemple).*

- une fonction donnée par un méta-descripteur : la forme d'onde, le spectre d'amplitude, l'enveloppe spectrale, etc.

Remarques :

1. pour toutes ces fonctions de conformation, si le résultat n'est pas normalisé dans $[0;1]$ ou $[-1;1]$, il faut appliquer ensuite une normalisation adaptée, excepté quand le résultat est utilisé pour le changement d'échelle temporelle (par exemple la fonction puissance) ;

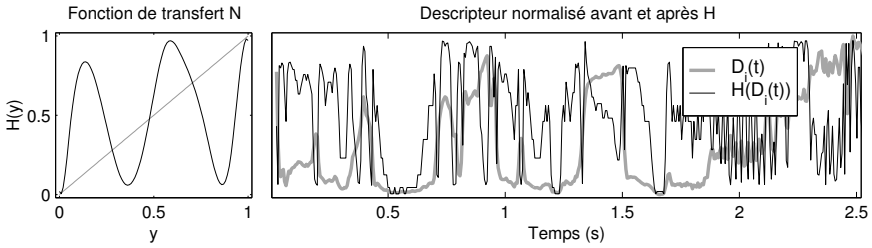


FIG. 6.28 – *Fonction de conformation arbitraire, donnée via une Interface Graphique Utilisateur (second exemple).*

2. ces modifications d'amplitude peuvent se faire en temps réel (dans ce cas, il faut modifier la fonction de lissage pour la rendre causale).

Nom	Notation	Fig. page	Nb Param.	Remarque
linéaire	$\mathcal{H}_{lin}$		0	contrôle direct par le descripteur
sinusoïde	$\mathcal{H}_{sine,i}, i = 1, 2$	*fig.* 6.4 p.217	2	2 zones d'attraction
sinusoïde réciproque	$\mathcal{H}_{\sin^{-1}}$	*fig.* 6.7 p.217	2	1 zone d'attraction
troncature	$\mathcal{H}_{trunc}$	*fig.* 6.8 p.217	2	sélection d'une portion d'intérêt
puissance X	$\mathcal{H}_{Xpow}$	*fig.* 6.9 p.218	2	dilatation/contraction temporelle
X puissance p	$\mathcal{H}_{XpowP}$	*fig.* 6.10 p.218	2	dilatation/contraction temporelle
bisinusoïdale	$\mathcal{H}_{bisin}$	*fig.* 6.15 p.220	2	3 zones d'attaction
logarithmique	$\mathcal{H}_{\log}$	*fig.* 6.17 p.220	2	
exponentielle	$\mathcal{H}_{\exp}$	*fig.* 6.18 p.221	2	
compresseur	$\mathcal{H}_{comp}$	*fig.* 6.19 p.221	2	
expandeur	$\mathcal{H}_{expan}$	*fig.* 6.20 p.221	2	
lissage	$\mathcal{H}_{sm}$	*fig.* 6.21 p.222	1	système à mémoire
dérivée	$\mathcal{H}_{deriv}$	*fig.* 6.23 p.222	0	différences finies
intégrale	$\mathcal{H}_{int}$	*fig.* 6.24 p.222	0	somme cumulative
écart	$\mathcal{H}_{int,2}$	*fig.* 6.25 p.223	1	écart à la fonction affine
valeur absolue	$\mathcal{H}_{abs}$	*fig.* 6.26 p.223	0	
arbitraire	$\mathcal{H}_{arb}$	*fig.* 6.27 p.223	0	donnée graphiquement
descripteur	$\mathcal{H}_{desc}$		0	donnée par un descripteur

TAB. 6.1 – *Récapitulatif des fonctions de conformation.*

6.3.2.ii) Modifications dynamiques en amplitude (hystérésis)

Les effets sur la dynamique que sont le compresseur et l'expandeur utilisent un temps de montée τ_m et un temps de descente τ_d pour éviter les sauts brusques lorsqu'on passe du régime sans correction au régime avec correction de dynamique, et réciproquement. Cela signifie qu'il y a une différence de comportement au moment où l'évolution du paramètre change de sens, et aussi selon si la valeur du paramètre augmente ou diminue. Nous nous sommes inspiré de ce principe pour concevoir une fonction de conformation de type hystérésis. Le principe est que nous avons deux fonctions de conformation, l'une correspondant à la fonction de conformation $\mathcal{H}_m(t)$ lorsque la valeur à transformer augmente, l'autre correspondant à la fonction de conformation $\mathcal{H}_d(t)$ lorsque la valeur à transformer diminue. On utilise un temps de montée et un temps de descente pour passer d'une courbe à l'autre.

$$\mathcal{H}_{hyst}(\mathcal{D}_i(t)) \quad = \quad \begin{cases} e^{(-t_0/\tau_m)}\ \mathcal{H}_m(\mathcal{D}_i(t)) + \left(1 - e^{(-t_0/\tau_m)}\right)\ \mathcal{H}_d(\mathcal{D}_i(t)) & \text{si } \mathcal{D}_i(t) \text{ augmente} \\ \left(1 - e^{(-t_0/\tau_d)}\right)\ \mathcal{H}_m(\mathcal{D}_i(t)) + e^{(-t_0/\tau_d)}\ \mathcal{H}_d(\mathcal{D}_i(t)) & \text{si } \mathcal{D}_i(t) \text{ diminue} \end{cases} \tag{6.20}$$

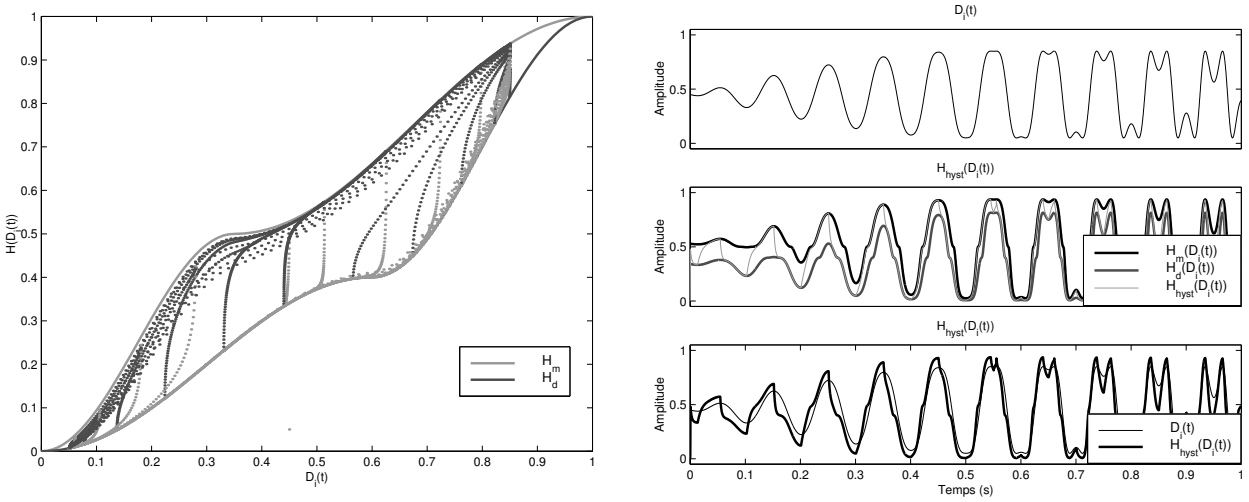


FIG. 6.29 – *A gauche : fonction et cycles d'hystérésis pour une modification dynamique de l'amplitude. A droite : modification dynamique de l'amplitude par fonction de conformation de type hystérésis.*

Un exemple de deux fonctions de conformation de type $\mathcal{H}_{bisin}$ est donné *fig.* 6.29. On voit des cycles d'hystérésis sur cette figure, pour les valeurs de la courbe $\mathcal{D}_i(t)$ donné en *fig.* 6.29 sup. Comme nous le voyons sur les figures suivantes, la courbe $\mathcal{D}_i(t)$ transformée par la fonction de conformation $\mathcal{H}_{hyst}$ passe de l'un des deux courbes $\mathcal{H}_m(\mathcal{D}_i(t))$ et $\mathcal{H}_d(\mathcal{D}_i(t))$ à l'autre, avec un temps de monté différent du temps de descente. Cette modification est bien différente des modifications en amplitude présentées précédemment. En effet, cette modification avec hystérésis est dynamique, elle dépend du temps et de l'évolution de la courbe à transformer. Ce contrôle est différent de l'utilisation de la dérivée d'un descripteur : l'utilisation de la dérivée permet des comportements différents selon le sens de variation de la courbe à transformer, mais ne permet pas de prendre en compte cette notion de temps de montée et de descente, c'est-à-dire la manière de converger plus ou moins rapidement vers la fonction de conformation.

6.3.2.iii) Modifications en temps

Les descripteurs peuvent aussi être modifiés en temps. Cependant, contrairement aux modifications en amplitude, la plupart des traitements ne peuvent s'appliquer en temps-réel. C'est donc plutôt la mise en œuvre sous *Matlab* qui est concernée. Les modifications en temps peuvent être des décalages (retard, ou avance), des inversions (temps différé), des dilatations ou contractions. On parlera de conformation temporelle (*time warping*).

Décalage Appliquer un retard peut se faire en temps-réel. Lorsqu'on l'effectue en temps-différé, on peut utiliser une permutation circulaire (*cf. fig.* 6.30) : les dernières valeurs sont utilisées en premier), ce qui permet d'avoir des valeurs non nulles au départ. Par contre, décaler la courbe pour que les valeurs de contrôle soient en avance sur le son ne peut se faire qu'en temps-différé. On peut aussi utiliser un retard variable (*cf. fig.* 6.31).

Inversion du déroulement temporel On peut aussi inverser le cours du temps de la courbe du descripteur (*cf. fig.* 6.32).

Conformation temporelle On peut ensuite étirer ou contracter des portions de la courbe. Pour ce faire, il existe deux manières :

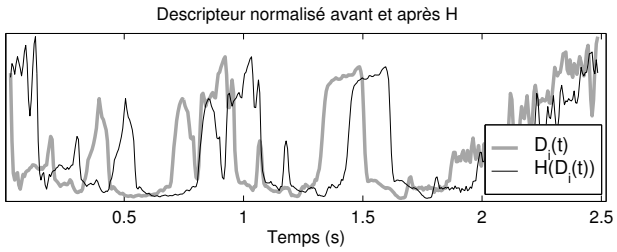


FIG. 6.30 – *Modification du temps par décalage, en temps différé.*

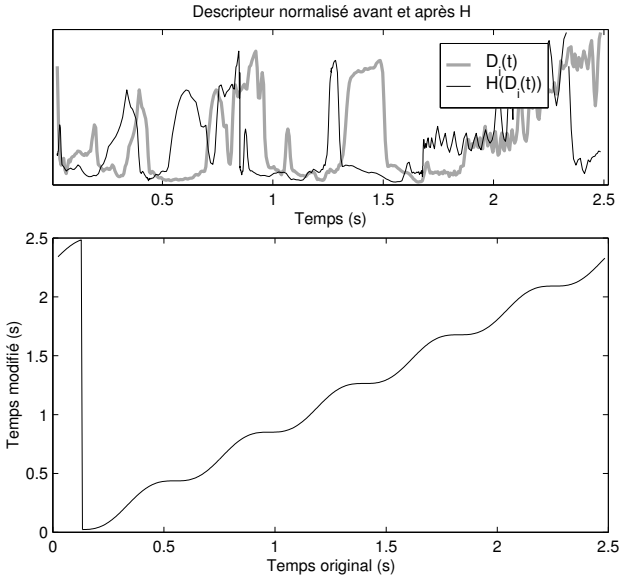


FIG. 6.31 – *Modification du temps par décalage temporel variable.*

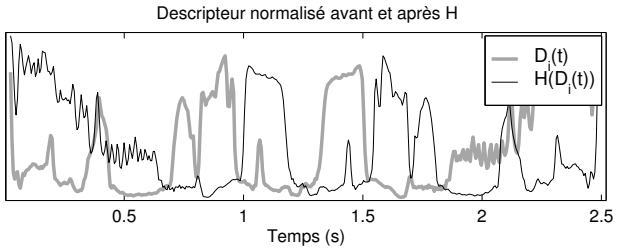


FIG. 6.32 – *Modification du temps par fonction de conformation de type sinusoïde.*

– la première manière consiste à utiliser les fonctions de modification présentées pour l'amplitude (elles doivent être monotones, ce qui réduit l'ensemble des possibilités), adaptées au fait que le temps ne se déroule pas dans l'intervalle $[0;1]$ mais plutôt $[t_-;t_+]$ (*cf. fig.* 6.33) :
– la seconde manière consiste à utiliser une fonction de la valeur de la courbe par utilisation de la somme cumulative du descripteur (*cf. fig.* 6.34).

Ces fonctions de modification en temps permettent :
– de conserver la corrélation entre le son et le contrôle de l'effet à l'aide du décalage (la courbe de contrôle dépend du son), tout en évitant le parfait synchronisme entre le contrôle et le son ;
– de ne pas conserver le synchronisme ni la corrélation entre le contrôle et le son, ce qui permet d'avoir un contrôle variable dont on ne soupçonne pas la provenance ;
– de modifier la saillance des pics, en les rendant plus ou moins large lorsque le temps est donné

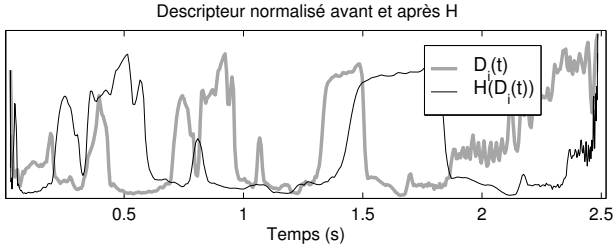


FIG. 6.33 – *Modification du temps par fonction de conformation de type sinusoïde.*

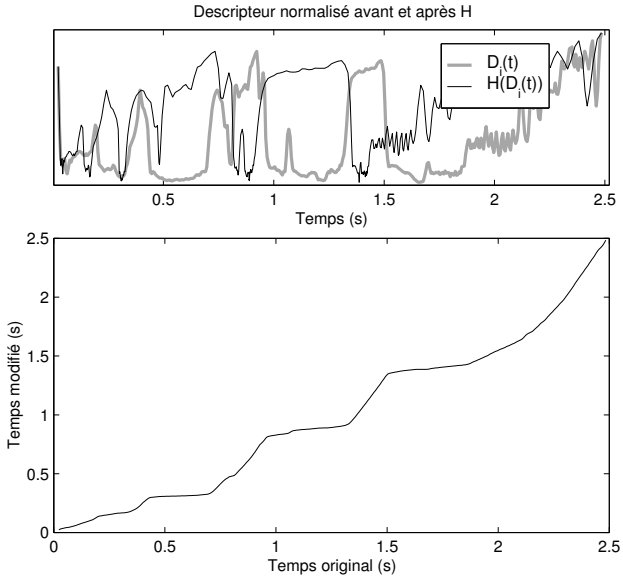


FIG. 6.34 – *Modification du temps par fonction de conformation de type somme cumulative.*

en fonction de la courbe de contrôle ;
– de proposer deux contrôles différents (avec deux conformations temporelles différentes) pour des traitements stéréophoniques.

6.3.2.iv) Modifications en amplitude et en temps

On peut enfin appliquer à la fois une modification en amplitude et une modification en temps. Dans ce cas, on peut obtenir un bien plus grand nombre de possibilités de transformation, mais au prix d'une lisibilité de moins en moins grande dans le rapport existant entre le son et le contrôle de l'effet qu'on y applique.

Quelle fonction choisir ?

Le choix du *mapping* se fait par expérimentations et écoute. En effet, une multitude de combinaisons de ces fonctions non linéaires peut être réalisée, et seule l'écoute permet de savoir ce qui est pertinent musicalement. Cela dépend de la sensibilité de chacun.

Si l'on cherche une grande cohérence entre le son et son contrôle, il vaudrait mieux éviter les complications introduites par une double modification en temps et en amplitude. Par contre, si l'on cherche un contrôle variable dont on n'arrive pas à comprendre, à l'écoute, le lien exact avec le son, alors cette double modification est exactement le genre de transformations qu'il nous faut.

6.3.3 Troisième étape : combinaison de M vers N

L'étape suivante du *mapping* consiste à effectuer une combinaison de M paramètres pour en obtenir N. Par exemple, on peut vouloir obtenir un seul paramètre à partir de deux paramètres du même son, ou au contraire obtenir deux paramètres à partir d'un seul. Il existe plusieurs méthodes pour passer de :

- de M vers N : par analyse en composantes principales (ACP ou *PCA : Principal Component Analysis*), par réseau de neurones, comme pour la synthèse imitative [Drame *et al.*, 1998] ;
- de M vers 1 : par combinaison linéaire ou non (somme, produit) ;
- de 1 vers N : en appliquant N fonctions de conformation en amplitude et temps différentes.

Malgré tout l'intérêt que présentent l'ACP et les réseaux de neurones, nos investigations dans ce domaine se sont limitées à la compréhension de la synthèse imitative (dans le cadre du stage de DEA d'Alexandre Morier en 2002). Nous le mentionnons donc uniquement à titre prospectif.

Nous avons utilisé en parallèle N combinaisons de M vers 1.

6.3.3.i) Combinaison linéaire de descripteurs

Pour commencer par le plus simple, la combinaison linéaire des descripteurs modifiés $\mathcal{J}(k,i,i_k,t)$ s'obtient par leur somme pondérée :

$$\mathcal{L}_a(t) \quad = \quad \sum_{k=1}^{K} \frac{a_k}{\sum_{k=1}^{K} a_k} \mathcal{J}(k,i,i_k,t) \tag{6.21}$$

avec a_k le poids du $k^{\text{ème}}$ descripteur. Les valeurs de pondération sont données par l'utilisateur, de manière intuitive, ou d'après une analyse en composantes principales (ACP). Par exemple, l'utilisateur dessine avec la souris la courbe qu'il désire obtenir ; il choisit ensuite un ensemble de descripteurs qui lui semblent appropriés, puis il effectue une ACP afin d'obtenir sa courbe à partir des descripteurs choisis.

Puisque $\mathcal{J}(k,i,i_k,t) \in [0;1]\ \forall t$ (resp. $[-1;1]$), on est sûr que la combinaison $\mathcal{L}_a(t) \in [0;1]\ \forall t$ (resp. $[-1;1]$) respecte l'intervalle initial.

6.3.3.ii) Combinaison non-linéaire de descripteurs par multiplication

Une autre manière de combiner les descripteurs consiste à les multiplier (combinaison non linéaire) :

$$\mathcal{L}_m(t) \quad = \quad \prod_{k=1}^{K} a_k \mathcal{J}(k,i,i_k,t) \tag{6.22}$$

Puisque $\mathcal{J}(k,i,i_k,t) \in [0;1]\ \forall t$ (resp. $[-1;1]$), on est sûr que la combinaison $\mathcal{L}_m(t) \in [0;1]\ \forall t$ (resp. $[-1;1]$) respecte l'intervalle initial.

6.3.3.iii) Descripteur croisé d'après deux descripteurs de longueurs différentes

Un problème se pose lorsque les descripteurs que l'on combine proviennent de jeux de descripteurs différents (ie. de sons différents) : leurs longueurs sont différentes. On désire en extraire des descripteurs croisés, et donc de longueur celle du plus long son. On peut réaliser cela de plusieurs manières :

- en ajoutant une valeur constante au descripteur le plus court, pour qu'il ait finalement la même durée que le descripteur le plus long, avant de procéder au calcul du descripteur croisé ;
- en répétant en boucle le descripteur le plus court, avant de procéder au calcul du descripteur croisé ;

– en étirant le descripteur le plus court pour lui faire avoir la même longueur que le plus long, avant de procéder au calcul du descripteur croisé.

Descripteur croisé obtenu par ajout d'une constante On dispose de deux sons de longueurs différentes, et de jeux de descripteurs différents. Lorsque le son à traiter est le plus court, un descripteur calculé à partir de deux descripteurs provenant chacun d'un jeu différent sera de la longueur du jeu de descripteurs le plus court. Dans ce cas, le calcul du descripteur "croisé" se fait uniquement sur la portion de temps commune aux deux jeux. Ceci dit, on peut vouloir traiter le son le plus long. Pour y parvenir, on propose de prolonger la dernière valeur du descripteur du son le plus court, jusqu'à la fin du descripteur le plus long. On peut ensuite calculer une version combinée des deux descripteurs (*cf. fig.* 6.35).

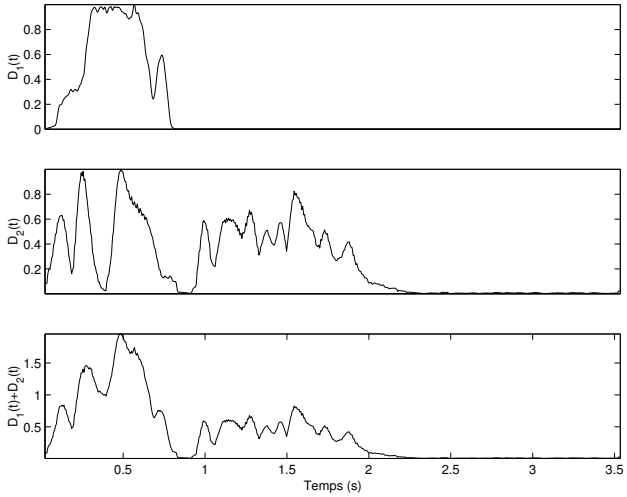


FIG. 6.35 – *Combinaison de deux descripteurs de longueurs différents sans conformation temporelle.*

Descripteur croisé obtenu par répétition périodique Une autre solution consiste à répéter la courbe du descripteur le plus court autant de fois que nécessaire. On peut ensuite calculer le descripteur combiné (*cf. fig.* 6.36). Dans ce cas, la courbe répétée n'est plus synchrone avec le signal d'origine à partir de la seconde répétition, puisque le signal plus court est terminé. Ceci peut cependant induire un contrôle intéressant, selon l'utilisation que l'on en fait.

Descripteur croisé obtenu par étirement d'un descripteur Une troisième solution consiste à rallonger le descripteur le plus court en l'étirant temporellement (*cf. fig.* 6.37). Dans ce cas, la courbe étirée n'est bien évidemment plus synchrone avec le signal d'origine. Ceci peut induire un contrôle intéressant, selon l'utilisation qu'il en est faite.

6.3.3.iv) Combinaison pondérée par des descripteurs

On peut se poser la question de savoir à quoi correspond le fait d'effectuer une combinaison linéaire, par exemple de deux descripteurs, avec des poids donnés par des descripteurs. Soient $\mathcal{X}_1(t)$ et $\mathcal{X}_2(t)$ les descripteurs modifiés à combiner. Leur combinaison linéaire s'écrit :

$$\mathcal{C}(t) = \alpha\mathcal{X}_1(t) + \beta\mathcal{X}_2(t) \tag{6.23}$$

Si on remplace maintenant les poids α et β par $\alpha\mathcal{X}_3(t)$ et $\beta\mathcal{X}_4(t)$, la combinaison linéaire devient :

$$\mathcal{C}(t) = \alpha\mathcal{X}_1(t)\mathcal{X}_3(t) + \beta\mathcal{X}_2(t)\mathcal{X}_4(t) \tag{6.24}$$

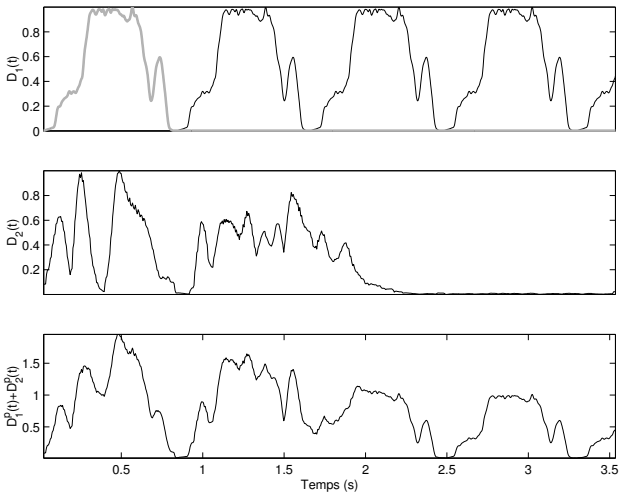


FIG. 6.36 – *Combinaison de deux descripteurs de longueurs différents par périodisation du plus court.*

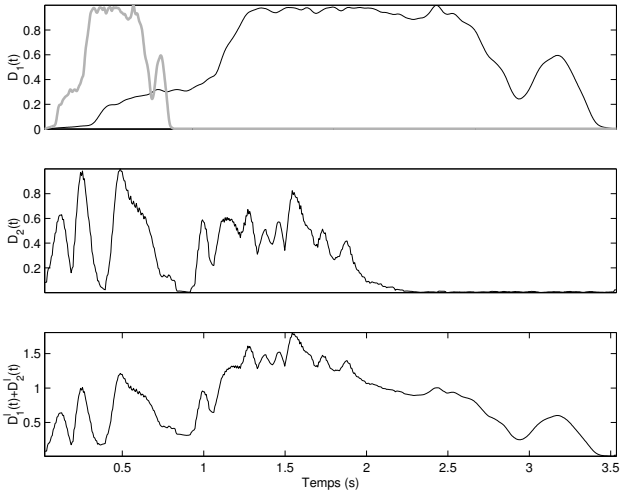


FIG. 6.37 – *Combinaison de deux descripteurs de longueurs différents par dilatation temporelle du plus court.*

Il s'agit alors d'une combinaison non linéaire ! De même, si on remplace maintenant les poids α et β par $\alpha\mathcal{X}_1(t)$ et $\beta\mathcal{X}_2(t)$, la combinaison linéaire devient :

$$\mathcal{C}(t) = \alpha\mathcal{X}_1^2(t) + \beta\mathcal{X}_2^2(t) \tag{6.25}$$

et pour α et β remplacés par $\beta\mathcal{X}_2(t)$ et $\alpha\mathcal{X}_1(t)$:

$$\mathcal{C}(t) = 2\alpha\beta\mathcal{X}_1(t)\mathcal{X}_2(t) \tag{6.26}$$

On voit donc bien que si l'on applique ici une pondération de la combinaison des descripteurs par des descripteurs, on peut obtenir un niveau supérieur de contrôle (par exemple le CGS provient d'une conbinaison d'autres descripteurs), mais on peut aussi obtenir une autre fonction de combinaison. Ceci renforce ma conviction qu'un contrôle sub-audio (gestuel ou descripteur de haut-niveau) est souvent plus adapté pour manipuler la combinaison des descripteurs qu'un contrôle à fréquence audio (descripteur de bas niveau).

6.3.4 Quatrième étape : application d'un ensemble de lois

Enfin, on applique à nouveau l'une des lois non-linéaires $\mathcal{H}_i$ proposées en sec. 6.3.2. La courbe obtenue est notée $\mathcal{C}_j(t) = \mathcal{H}_{i_j}\left(\mathcal{L}_{k_j}(t)\right)$ avec $k_j \in \{1, 2\}$ le type de combinaison appliquée. Cette courbe de contrôle n'a plus qu'à être ajustée aux critères de contrôle imposés soit par l'effet, soit par l'utilisateur, soit par les deux. Il s'agit du deuxième du premier niveau de *mapping*, l'étage de conformation spécifique.

6.4 Second étage (N1-E2) : ajustements des contrôles aux critères

Le second étage de la mise en correspondance entre descripteurs et contrôles de l'effet concerne la conformation des paramètres aux exigences de contrôle de l'effet. Le diagramme *fig.* 6.38 en indique les composantes. Les trois premières fonctionnalités que sont la conformation temporelle, le zoom et la quantification, n'ont pas d'ordre particulier. Par contre, les deux dernières fonctionnalités que sont le filtrage passe-bas et l'ajustement aux bornes de variation doivent absolument figurer dans cet ordre, en dernier. Elles sont donc présentées comme des "étapes", pour indiquer cet idée d'ordre.

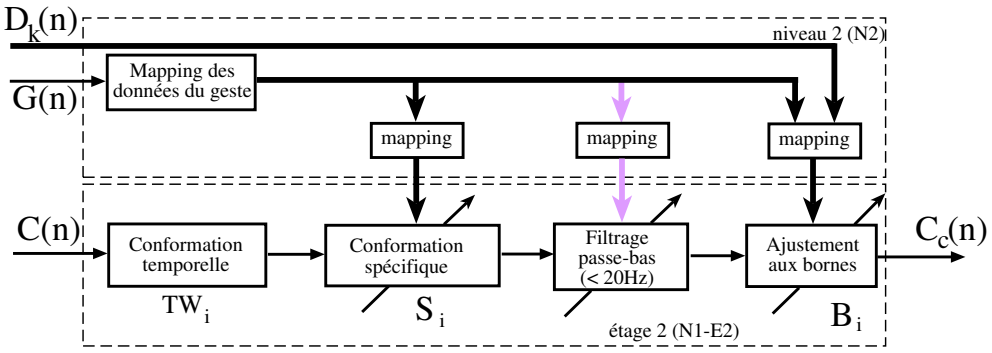


FIG. 6.38 – *Diagramme du second étage de* mapping *: conformations spécifiques (en amplitude, temporelle et filtrage passe-bas) et ajustement aux bornes de valeurs de contrôle.*

6.4.1 Première fonctionnalité : conformation temporelle

Concernant la conformation temporelle, on se reportera à l'exposé qui en a été fait auparavant en sec. 6.3.2. Cette conformation peut en fait intervenir à n'importe quel endroit du *mapping*.

6.4.2 Deuxième fonctionnalité : zoom

Principe

Si l'on utilise une courbe de contrôle (monovaluée à l'instant t) dont on modifie la plage de variation, on obtient un effet de *zoom* : on peut faire en sorte que l'effet se focalise sur de petites variations du paramètre de contrôle pour les amplifier. Ces petites variations peuvent être par exemple celles d'un descripteur d'un son de type nappe, à lente évolution. Ce peut aussi être celles d'un geste momentanément confiné dans une petite zone de l'espace gestualisable.

Mise en œuvre

Soit la courbe de contrôle $\mathcal{C}(t)$. Nous allons donner plusieurs fonctions de zoom ("loupes", ou *scaling*), notées $\mathcal{Z}_i$, définies par la formule générale :

$$\mathcal{Z}_i(t) = \mathcal{Z}_i(\mathcal{C}(t)) = \frac{\mathcal{C}(t) - \mathcal{Y}_i^-(t)}{\mathcal{Y}_i^+(t) - \mathcal{Y}_i^-(t)} \tag{6.27}$$

avec les bornes de normalisation $\mathcal{Y}_i^+(t) = \mathcal{Y}_i^+(\mathcal{C}(t))$ (borne supérieure) et $\mathcal{Y}_i^-(t) = \mathcal{Y}_i^-(\mathcal{C}(t))$ (borne inférieure) définies différemment selon la fonction de loupe. Définissons d'abord les extrema locaux et le moyenne locale :

$$\begin{aligned} \mathcal{C}_T^-(t) &= \min_{k\in\{t-T\ldots t\}} \mathcal{C}(k) \qquad \text{minimum local} \\ \mathcal{C}_T^+(t) &= \max_{k\in\{t-T\ldots t\}} \mathcal{C}(k) \qquad \text{maximum local} \\ <\mathcal{C}(t)>_T &= \frac{1}{T}\sum_{k=t-T}^{t} \mathcal{C}(k) \qquad \text{moyenne locale} \end{aligned}$$

La première fonction loupe $\mathcal{Z}_1^{\pm}$ est définie par les bornes suivantes :

$$\mathcal{Y}_1^{\pm}(t) = \mathcal{C}_T^{\pm}(t) \tag{6.28}$$

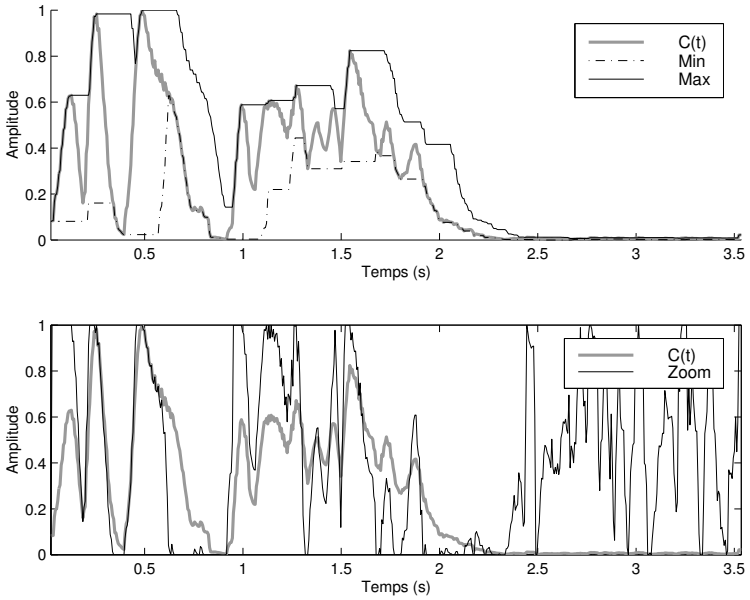


FIG. 6.39 – *Zoom avec la fonction $\mathcal{Z}_1(t)$ (bornes : extrema locaux).*

Avec la loupe $\mathcal{Z}_1^{\pm}(t)$, les limites sont données par les extrema locaux sur une fenêtre glissante. Cette fonction de loupe permet de conserver l'intervalle de variations une durée égale à la longueur temporelle de la fenêtre de moyenne, avant de se réduire, jusqu'à ce que la courbe ait un nouvel extremum local.
On utilise les notations suivantes pour le maximum et me minimum entre deux valeurs :

$$\mathcal{M}^+ = \max, \quad \mathcal{M}^- = \min$$

La loupe $\mathcal{Z}_2^{\pm}(t)$ est définie à partir des bornes :

$$\mathcal{Y}_2^{\pm}(t) = \mathcal{M}^{\pm}\left(\mathcal{C}(t), \frac{\mathcal{C}_T^{\pm}(t) + <\mathcal{C}(t)>_T}{2}\right) \tag{6.29}$$

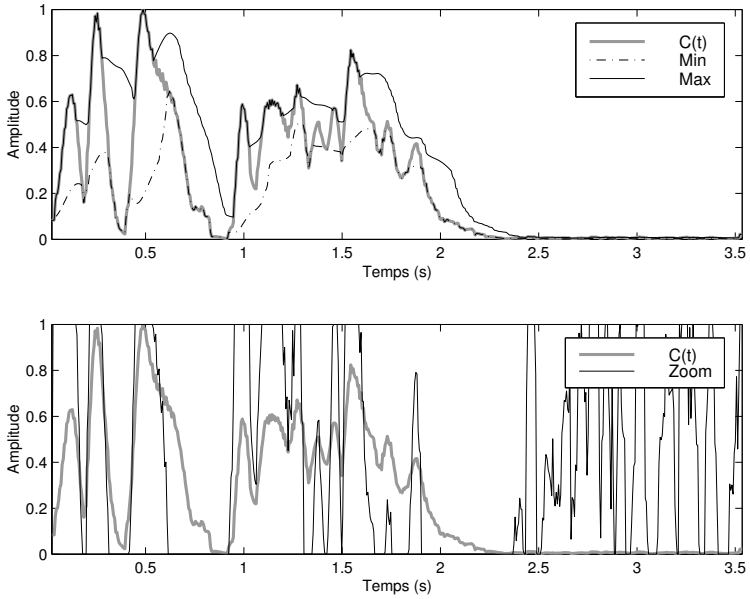


FIG. 6.40 – *Zoom avec la fonction $\mathcal{Z}_2(t)$ (borne : moyenne entre extrema locaux et valeur moyenne).*

Avec la loupe $\mathcal{Z}_2^{\pm}(t)$, on tient compte à la fois des extrema locaux et de la valeur moyenne de la courbe sur la même fenêtre. Les pics sont donc suivis avec plus de précision, et on évite les paliers présents avec la loupe $\mathcal{Z}_1^{\pm}(t)$.

$$\mathcal{Y}_3^{\pm}(t) \quad = \quad \mathcal{M}^{\pm}\left(\mathcal{C}(t), \frac{\mathcal{C}_T^{\pm}(t) + \mathcal{C}(t-1)}{2}\right) \tag{6.30}$$

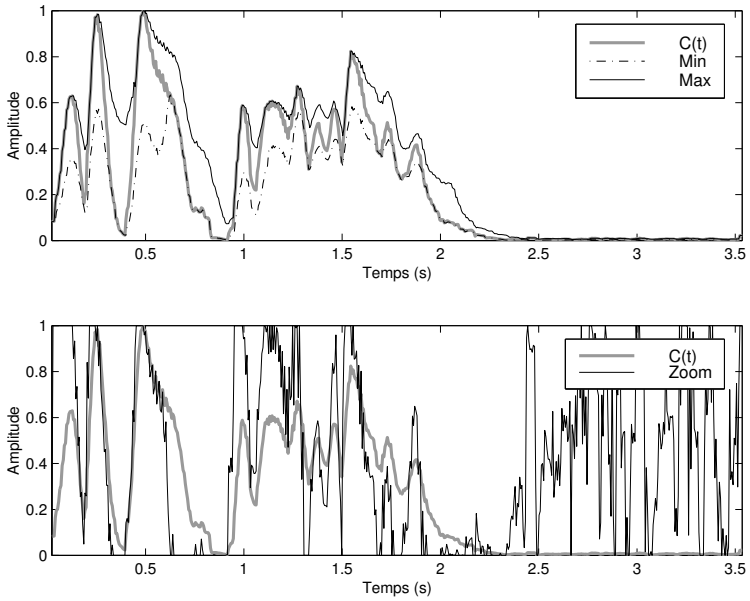


FIG. 6.41 – *Zoom avec la fonction $\mathcal{Z}_3(t)$ (bornes : moyenne entre extrema locaux et valeur locale).*

Avec la loupe $\mathcal{Z}_3^{\pm}(t)$, on utilise les extrema locaux et la dernière valeur. Les variations sont donc plus grandes qu'avec la loupe $\mathcal{Z}_2^{\pm}(t)$, puisqu'il y a moins de filtrage.

$$
\begin{aligned}
\mathcal{D}_i^{\pm}(t) &= \delta^{\pm}\mathcal{C}(t)\ \left[|\mathcal{Y}_i^{+}(t-1) - \mathcal{Y}_i^{+}(t-2)|\right] \\
\mathcal{E}^{\pm}(t) &= \beta^{\pm}\left[1 - e^{\alpha\left(t-t_a^{\pm}\right)}\right] \\
\mathcal{G}_i^{\pm}(t) &= \gamma^{\pm}\mathcal{C}(t)\ \left[\mathcal{Y}_i^{+}(t-1) - \mathcal{Y}_i^{-}(t-1)\right] \\
\mathcal{Y}_4^{\pm}(t) &= \mathcal{M}^{\pm}\left(\mathcal{C}(t), \mathcal{C}_a^{\pm} + \mathcal{D}_4^{\pm}(t) \mp \mathcal{E}^{\pm}(t) + \mathcal{G}_4^{\pm}(t)\right) \\
\text{avec} \quad \text{si} \quad & \mathcal{Y}_4^{\pm}(t) = \mathcal{C}(t), \text{alors } \mathcal{C}_a^{\pm} = \mathcal{C}(t), t_a^{\pm} = t
\end{aligned}
\tag{6.31}
$$

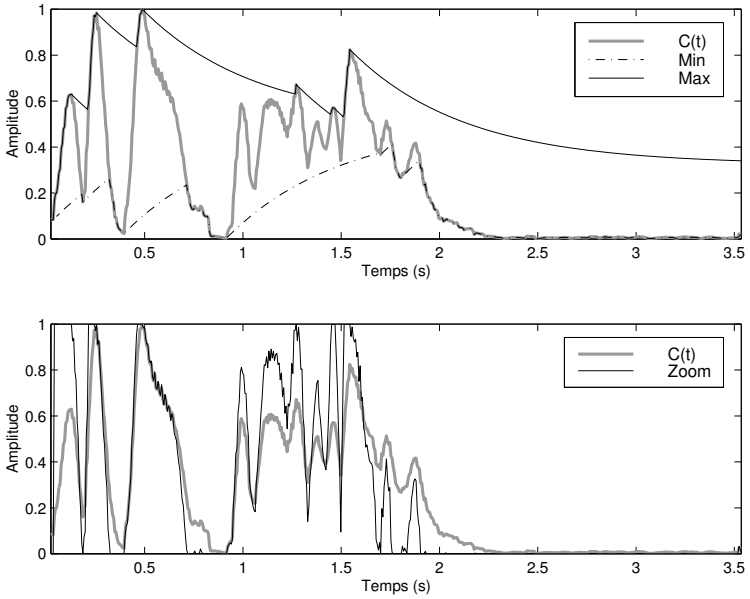


FIG. 6.42 – *Zoom avec la fonction $\mathcal{Z}_4(t)$ pour $\alpha = -0.01$, $\beta = 0.5$ et $\delta = 0$.*

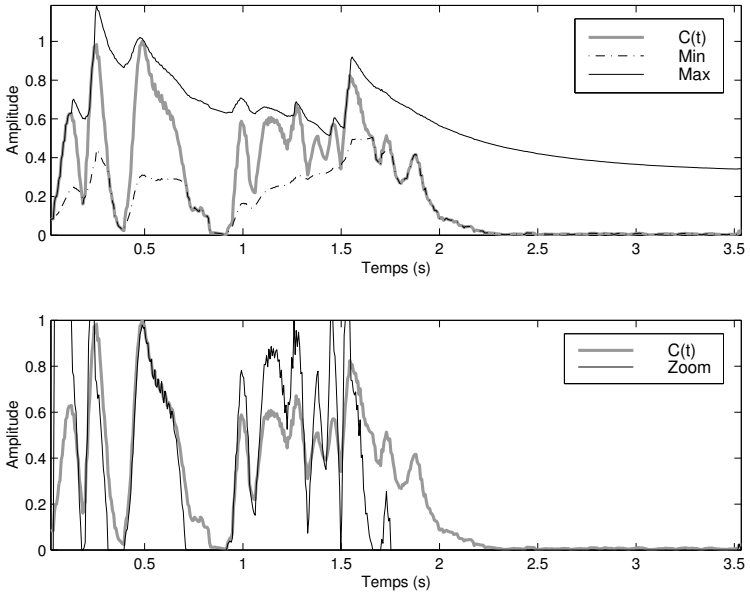


FIG. 6.43 – *Zoom avec la fonction $\mathcal{Z}_4(t)$ pour $\alpha = -0.01$, $\beta = 0.5$ et $\delta = 0.3$.*

Avec la loupe $\mathcal{Z}_4^{\pm}(t)$, on utilise deux courbes exponentielles $\mathcal{E}^{\pm}$ pour tendre petit à petit vers le milieu de l'intervalle. On utilise aussi la dérivée $\mathcal{D}$ de la courbe de contrôle et la largeur du dernier intervalle $\mathcal{G}$. Selon les valeurs de $\delta^{\pm}$, $\beta^{\pm}$ et $\gamma^{\pm}$, on pondère ces trois fonctions, ce qui permet d'obtenir des zooms bien différents. L'intérêt de l'exponentielle est que le zoom se resserre progressivement sur la courbe de contrôle ; la courbe obtenue après changement des bornes voit donc ses faibles variations devenir plus grandes de la manière la plus progressive. En ajustant la forme des exponentielle, on définit la vitesse de zoom.

Les exemples de courbes qui sont donnés ici montrent que les différentes loupes ont des comportements différents : ce n'est pas la valeur de la courbe en soi mais son évolution qui est intéressante. Remarquons que ces fonctions de zoom peuvent autant servir pour modifier des descripteurs que pour modifier des contrôles gestuels (les programmes sous *Max/MSP* correspondants ont été réalisés).

Quelques exemples sonores

Les exemples que nous donnons ont été en partie présentés lors de la conférence *DAFx-02*[1] à Hambourg [Verfaille and Arfib, 2002]. ils ont été obtenus en temps différé, avec *Matlab*.

Le premier exemple porte sur la transposition adaptative. La Piste n°16 ♫♪est le son de référence. La courbe de contrôle sans loupe donne le son Piste n°49-CD2 ♫♪. Lorsqu'on utilise la loupe $\mathcal{Z}_2(t)$, (*cf.* Piste n°50-CD2 ♫♪) certaines portions du son ont un gain et un délai différent qui les met plus en valeur, notamment sur la fin du son


CD2

Le deuxième exemple Piste n°51-CD2 ♫♪ porte sur la transposition adaptative. L'utilisation de la loupe $\mathcal{Z}_1(t)$ permet de donner plus d'ambitus, notamment sur la fin du son où le paramètre de contrôle varie peu, avant application de la loupe (*cf.* Piste n°52-CD2 ♫♪ avec la loupe $\mathcal{Z}_2(t)$ et *cf.* Piste n°53-CD2 ♫♪ avec la loupe $\mathcal{Z}_4(t)$).

Le troisième et dernier exemple porte sur la dilatation/contraction temporelle adaptative. Le son de référence est la Piste n°54-CD2 ♫♪. Lorsqu'on utilise la loupe $\mathcal{Z}_1(t)$, le son est plus dilaté et contracté aux passages où le contrôle variait peu avant application de la loupe, notamment sur la fin (*cf.* Piste n°55-CD2 ♫♪).

6.4.3 Troisième fonctionnalité : discrétisation

La discrétisation d'une courbe est utilisée lorsque la précision du paramètre de contrôle est très supérieure à celle utilisée par l'effet. On discrétise alors la valeur de contrôle sur une échelle de valeurs prédéfinies. Ce peut être une échelle de hauteur variable (*cf.* sec. 5.4.4). Dans ce cas, l'échelle est définie par l'utilisateur et le contrôle adaptatif ; elle respecte des critères musicaux (les intervalles entres hauteurs sont spécifiques à l'échelle). Ce peut aussi être une échelle utilisée pour un effet dont le nombre de valeurs de contrôle possible est fini. C'est le cas de l'écho granulaire adaptatif (*cf.* sec. 5.5.1), qui ne peut se mettre en œuvre en temps-réel qu'avec une précision limitée en terme de délai, puisqu'on ne dispose que d'un nombre fini de lignes à retard. Dans ce cas, il faut trouver des manières de discrétiser "au mieux" la courbe de contrôle (nous verrons plus loin quels sens on peut donner à l'expression "au mieux", c'est-à-dire selon quels critères). Ainsi, disposant de N_{del} lignes à retard, on associe à chacune des propriétés de longueur et de gain de réinjection en fonction d'une discrétisation. Une fois qu'un ligne à retard est vide, c'est à dire que son énergie est inférieure à un seuil, on lui attribue de nouvelles valeurs de longueur et gain de réinjection, de manière à mieux correspondre avec la discrétisation "idéale" du moment. On dit dans ce cas que la ligne à retard est re-allouée. C'est à nouveau une forme d'adaptation.

Les discrétisations que nous présentons ici ont été choisies pour l'écho granulaire adaptatif, mais peuvent aussi s'utiliser pour tous les autres effets. La courbe discrétisée est l'énergie du son Piste n°7 ♫♪. Plusieurs mise en œuvre ont été réalisées : la quantification uniforme, la quantification non uniforme par centroïde, la quantification non uniforme pondérée, la quantification non uniforme pondérée avec prise en compte des extrema.

[1] *DAFx : Digital Audio Effects Workshop*

Quantification uniforme

La courbe de contrôle doit être quantifiée, selon une grille de $n_q = 20$ ou 30 valeurs, par exemple. La solution la plus simple consiste à utiliser une grille uniforme [Zoelzer, 1997]. Soit la courbe de contrôle $\mathcal{C}$ bornée en amplitude par l'intervalle $[\Delta_m; \Delta_M]$. Les segments de quantification :

$$\mathcal{I}(n, n_q) = [i_u(n, n_q); i_u(n+1, n_q)] \tag{6.32}$$

ont pour milieu $s_u(n, n_q) = \Delta_m + \frac{n-1/2}{n_q}(\Delta_M - \Delta_m)$ et pour extrémités $i_u(n, n_q) = \Delta_m + \frac{n-1}{n_q}(\Delta_M - \Delta_m)$. La fonction de quantification uniforme est :

$$\mathcal{Q}_u(t, n_q) = s_u\left(\arg \min_{n \in \{1 \ldots n_q\}} |\mathcal{C}(t) - s_u(n, n_q)|, n_q\right) \tag{6.33}$$

Une variante consiste à utiliser un intervalle de variation élargie de manière à ce que les deux marques de quantifications aux extrémités soient les extrema de la fonction (*cf. fig.* 6.44 droite), ce qui n'était pas le cas avec la quantification uniforme standard.

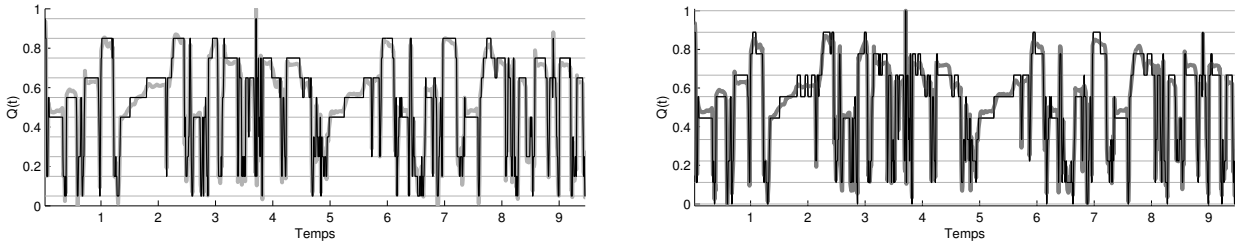


FIG. 6.44 – *Quantification uniforme et uniforme avec intervalle de variation élargi.*

Quantification non uniforme par moyenne locale

L'utilisation de la quantification uniforme ne tient pas compte des spécificités de variation de la courbe discrétisée. Une première solution consiste à prendre non pas le milieu $s_u(n, n_q)$ du segment $\mathcal{I}(n, n_q)$, mais la moyenne locale $c(n, n_q) =$ des valeurs de la courbe dans cet intervalle, donné par :

$$c(n, n_q) = \frac{\sum_{k=1}^{N_T} \mathcal{C}(k) \mathbb{1}_{\mathcal{C}(k) \in \mathcal{I}(n, n_q)}}{\sum_{k=1}^{N_T} \mathbb{1}_{\mathcal{C}(k) \in \mathcal{I}(n, n_q)}} \tag{6.34}$$

La fonction de quantification non-uniforme par moyenne locale est :

$$\mathcal{Q}_c(t, n_q) = c\left(\arg \min_{n \in \{1 \ldots n_q\}} |\mathcal{C}(t) - c(n, n_q)|, n_q\right) \tag{6.35}$$

On voit *fig.* 6.45 que l'on prend un peu mieux compte des zones où la courbe a la plus de valeurs, mais uniquement de manière locale.

Quantification non uniforme par moyenne locale et itérations

On applique maintenant la quantification non uniforme par moyenne locale en itérant le processus. Les segments de l'itération k ont pour extrémités les milieux entre marques de quantification de l'itération $k-1$. On itère le processus de recherche des valeurs de quantification jusqu'à ce que les marques convergent vers une valeur fixe. On obtient alors une quantification non uniforme par moyenne locale et itérations. Comme nous pouvons le voir *fig.* 6.46, les zones où la courbe a la plus de valeurs sont prises en compte de manière prépondérante : elles ont plus de valeurs de discrétisation.

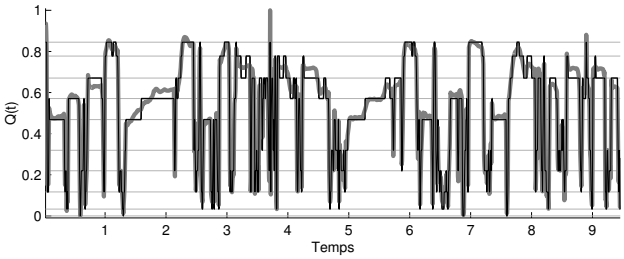


FIG. 6.45 – *Quantification non uniforme par moyenne locale.*

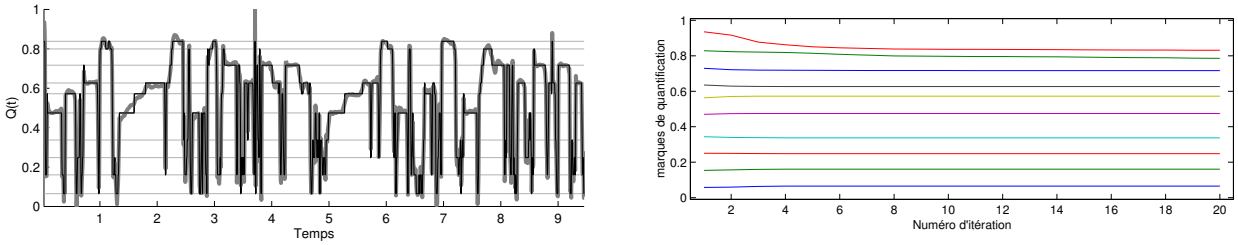


FIG. 6.46 – *A gauche : application de la quantification non uniforme par moyenne locale et itérations. A droite : marques de la quantification non uniforme par moyenne locale et itérations, au fur et à mesure des itérations.*

Quantification non uniforme pondérée

La **quantification non-uniforme pondérée** utilise les n_q maxima de l'histogramme de la courbe à discrétiser comme valeurs de discrétisation. Ceci permet de prendre en compte de manière prépondérante les zones où la courbe a le plus de valeurs sans avoir à utiliser d'algorithme itératif.

Une fois créé l'histogramme de la courbe de contrôle avec $n_H > n_q$ paniers, on utilise les abscisses des n_q plus grand pics comme valeurs de quantification. La fonction d'histogramme est donnée par

$$H(n, n_H) = \sum_{k=1}^{N_T} \mathbb{1}_{\mathcal{C}(k) \in \mathcal{I}(n, n_H)}$$

avec la fonction de densité associée $\mathcal{D}(n, n_H) = s_u(n, n_H)$ et $\mathbb{1}$ la fonction indicatrice. Le nombre maximum n_q de valeurs différents de l'histogramme $H(n, n_H)$ sont $\delta(n), n = 1, .., n_q$ définies par :

$$\delta(n, n_q) = \mathcal{D}\left(\max_{k \in S(n-1)} H(k, n_H), n_H\right) \tag{6.36}$$

avec l'ensemble :

$$S(n-1) = \left\{ i \in \{1 \dots n_H\}; H(i, n_H) \in \{\delta(k, n_H)\}_{k=1 \dots n-1} \right\} \tag{6.37}$$

on obtient enfin la fonction de quantification non uniforme pondérée :

$$\mathcal{Q}_w(t, n_q) = \delta\left(\arg \min_{n \in \{1 \dots n_q\}} |\mathcal{C}(t) - s_u(n, n_q)|, n_q\right) \tag{6.38}$$

Cette quantification ne tient pas compte des extrema locaux (*cf. fig.* 6.47).

Quantification non uniforme pondérée tenant compte des extrema

On peut donner un sens musical au fait de tenir compte des extrema locaux de la courbe de contrôle (les minima et maxima locaux atteints par la courbe à discrétiser, sur des fenêtres

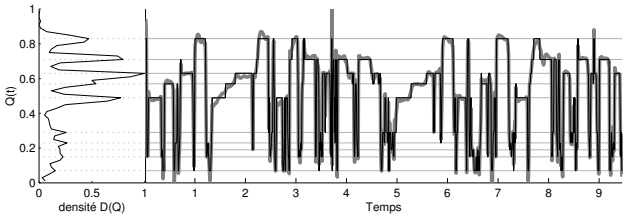


FIG. 6.47 – *Quantification non uniforme pondérée : histogramme et courbe discrétisée.*

temporelles de quelques dixièmes de secondes). En effet, les valeurs extrêmes, même si elles sont peu souvent atteintes, correspondent à une information de contrôle pertinente, puisqu'ils décrivent la plus grande variation de la courbe. Le problème de la quantification non-uniforme pondérée est qu'elle ne tient pas compte de ces extrema. Aussi, nous proposons d'y remédier avec cette nouvelle quantification que nous avons développée pour nos besoins : la **pondération non-uniforme pondérée avec pics**.

On calcule comme précédemment les $n_q - n_p$ valeurs de quantification non-uniforme pondérée $\delta(n, n_q - n_p)$, puis on calcule les n_p valeurs de quantification correspondant à une pondération entre l'extremum de la portion de courbe de contrôle et l'extremum des marques de quantification. On définit le plus petit intervalle contenant toutes les marques de quantification par :

$$\mathcal{I}_{extr} = [\min_n \delta(n, n_q - n_p); \max_n \delta(n, n_q - n_p)] = [\delta^-; \delta^+] \tag{6.39}$$

on extrait alors $2n_p$ extrema locaux : n_p maxima locaux, notés $\mathcal{P}^+(n)$ et n_p minima locaux notés $\mathcal{P}^-(n)$. On calcule ensuite leur distance à la plus proche borne de l'intervalle $\mathcal{I}_{extr}$:

$$d^{\pm}(n) = \mathcal{P}^{\pm}(n) - \delta^{\pm} \tag{6.40}$$

On définit alors les marques de quantifications de pics pondérées par :

$$\mathcal{P}^{\pm}_{\alpha}(n) = \delta^{\pm} + \alpha \left(\mathcal{P}^{\pm}(n) - \delta^{\pm}\right) \tag{6.41}$$

ainsi que leurs distances à la borne de $\mathcal{I}_{extr}$ la plus proche :

$$d^{\pm}_{\alpha}(n) = \mathcal{P}^{\pm}_{\alpha}(n) - \delta^{\pm} \tag{6.42}$$

Finalement, on classe ces marques de quantification de pics de la plus lointaine à la plus proche de l'intervalle, tout en ne gardant que celles extérieures à l'intervalle :

$$\mathcal{P}^{cl}_{\alpha} = \left\{\mathcal{P}^{\pm}_{\alpha}(k); \delta^{\pm} > 0, |\mathcal{P}^{cl}_{\alpha}(i-1) - m| > |\mathcal{P}^{cl}_{\alpha}(i) - m|\right\} \tag{6.43}$$

avec $m = \frac{\delta^- + \delta^+}{2}$ la valeur moyenne de l'intervalle $\mathcal{I}_{extr}$. L'ensemble des valeurs de quantifiation devient :

$$\overline{\Delta}(., n_q) = \{\delta(n, n_q - n_p)\}_{n \in \{1 \ldots n_q - n_p\}} \cup \left\{\mathcal{P}^{cl}_{\alpha}(i)\right\}_{i \in \{1 \ldots n_p\}} \tag{6.44}$$

La fonction de quantification non-uniforme pondérée de pics est donnée par :

$$\mathcal{Q}_{p,\alpha}(t, n_q) = \overline{\Delta}\left(\arg \min_{n \in \{1 \ldots n_q\}} |\mathcal{C}(t) - \delta(n, n_q)|, n_q\right) \tag{6.45}$$

Pour $\alpha = 0$, on a $\mathcal{Q}_{p,0}(t, n_q) = \mathcal{Q}_w(t, n_p)$: aucun pic n'est pris en compte en dehors de l'intervalle. Pour $\alpha = 1$, les extrema locaux sont directement pris en compte. Cela signifie que les valeurs proches des pics seront quantifiés aux valeurs des pics, ce qui produira une erreur de quantification moins grande (et des résultats sonores plus proches de la version temps différé). Des valeurs intermédiaires permettent de prendre en compte l'effet du pic, sans en prendre la valeur exacte. De bonnes valeurs sont données par l'intervalle $\alpha \in [0.5; 0.8]$.

Quelle quantification choisir ?

L'utilisateur doit écouter le résultat de plusieurs quantifications de la courbe de contrôle, et éventuellement les comparer au résultat de la version temps-différé : l'effet musical peut être très différent d'une quantification à une autre, et il n'y a pas d'a priori sur la manière de bien faire sonner un son traité par un effet dont la courbe de contrôle est quantifiée. La manière de choisir l'une des fonctions de quantification (avec plusieurs valeurs possibles pour α) n'est donc pas évidente. Cependant, nous pouvons donner quelques pistes que nous avons explorées.

La quantification uniforme est la plus pratique à mettre en œuvre. Elle donne la même importance à toutes les valeurs de la courbe $\mathcal{C}(t)$. A elle seule, elle permet déjà de modifier grandement le son, selon si le contrôle est quantifié ou non. Il en va de même pour la quantification uniforme avec correction de l'intervalle, à ceci près qu'il est fait une vraie différence entre les extrema de $\mathcal{C}(t)$ et les autres valeurs, du fait que la quantification possède une marque exactement sur chaque extrema. La quantification non uniforme par moyenne locale donne des différences très légères pour des effets tel que la dilatation/contraction temporelle adaptative ou la transposition adaptative : il faut vraiment tendre l'oreille pour entendre les subtilités.

La quantification non uniforme par moyenne locale et itérations en revanche se focalise sur les zones de $\mathcal{C}(t)$ souvent traversées, ce qui permet de donner plus d'importance à ces zones qu'elle n'en ont avec la quantification uniforme, sans pour autant oublier les extrema de la courbe. La quantification non uniforme pondérée réalise le même genre de quantification, à un coût de calcul moindre (pas de méthode itérative), avec des différences très faibles. La quantification non uniforme pondérée tenant compte des maxima locaux permet de donner plus d'importance aux extrema locaux, même lorsqu'ils ne sont que très rarement atteints par $\mathcal{C}(t)$.

6.4.4 Quatrième étape : filtrage passe-bas

En introduction, nous avons justifié la structure du *mapping* que nous utilisons par le fait que le contrôle de l'effet se prête bien à des variations rapides, comme c'est le cas pour les descripteurs de signal et perceptifs. Cependant, il nous faut ici nuancer notre propos : tous les effets ne sont pas prévus pour être contrôlé par un signal à fréquences audio sans que des artefacts se produisent. Certains effets, tels que la modulation d'amplitude et ses variantes ont besoin d'un paramètre de contrôle ne variant pas au-dessus de 20 *Hz*. En effet, des modulations d'amplitude par une sinusoïde de fréquence supérieure à 20 *Hz* s'entendent dans le domaine fréquentiel (*cf.* sec. 2.3.1). Si les modulations d'amplitudes ne sont pas sinusoïdales, le risque est que les variations rapides produisent des *clicks*. La panoramisation présente les mêmes risques. Des descripteurs tels que les indicateurs passent brutalement d'une valeur 0 à 1 (par exemple), ce qui implique de hautes fréquences, elles aussi nuisibles, car elles introduisent des clicks dans le son. Afin d'éviter ce désagrément (et de pouvoir choisir si oui ou non, on désire conserver les artefacts), il faut appliquer un filtrage passe-bas afin de lisser la courbe. Nous avons utilisé un filtre bi-quadratique (filtre d'ordre quatre avec deux pôles et deux zéros), autant dans les programmes sous *Matlab* que dans les *patchs* sous *Max/MSP*.

6.4.5 Cinquième étape : ajustement des bornes de variation

Finalement, une fois que l'on a modifié des descripteurs de manière à obtenir une courbe de contrôle $\mathcal{X}(t)$, cette courbe est ajustée aux bornes de variations du contrôle. Soient Δ_m le minimum et Δ_M le maximum autorisés pour le contrôle de l'effet. La courbe de contrôle est alors donnée en fonction de Δ_m et Δ_M par :

$$\mathcal{C}_{fx}(t) = \Delta_m + (\Delta_M - \Delta_m)\ \mathcal{X}(t) \tag{6.46}$$

On peut vouloir que cette courbe de contrôle ait des bornes non plus constantes dans le temps mais données par des contrôles $c_i(t)$ obtenus à partir de descripteurs. Dans ce cas, à partir de deux

courbes de contrôle normalisées dans un intervalle inclus dans $[\Delta_m; \Delta_M]$ l'intervalle de variation du contrôle, on calcule le contrôle de l'effet $\mathcal{C}_{fx}(t)$ ainsi (*cf. fig.* 6.49) :

$$\mathcal{C}_{fx}(t) = \min_{i=1,2}(c_i(t)) + \left(\max_{i=1,2} c_i(t)) - \min_{i=1,2}(c_i(t))\right) \frac{c(t) - \min c(t)}{\max c(t) - \min c(t)} \tag{6.47}$$

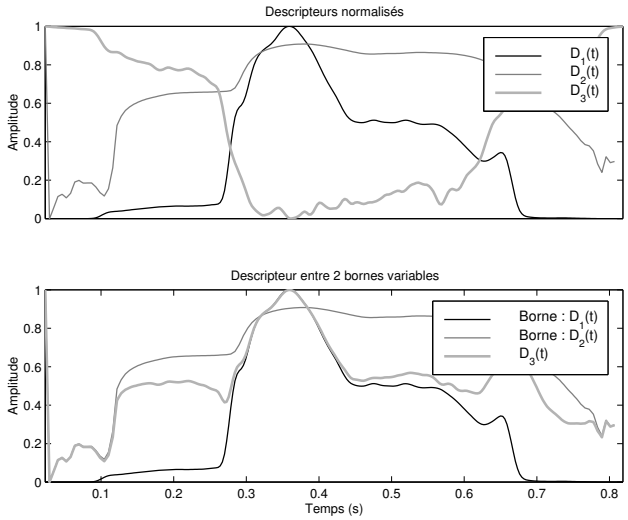


FIG. 6.48 – *Ajustement variable des bornes d'un contrôle à partir de deux descripteurs.*

Pour les effets de type trémolo adaptatif, vibrato adaptatif, on utilise un descripteur comme fréquence variable d'un oscillateur. Cet oscillateur varie entre -1 et 1, on peut donc le faire varier entre d'autres bornes, par exemple des bornes données par des contrôles (*cf. fig.* 6.49 droite).

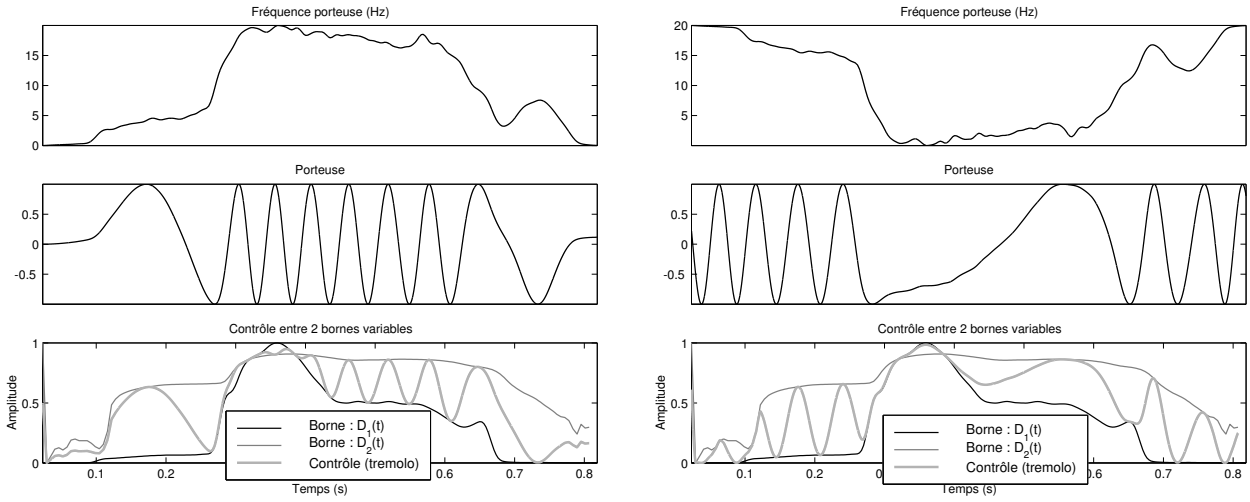


FIG. 6.49 – *Ajustement variable des bornes d'un contrôle (modulé en fréquence) à partir de deux descripteurs.*

Lorsqu'on utilise deux courbes comme bornes, on peut choisir de prendre comme borne inférieure le minimum des deux courbes, et comme borne supérieure le maximum des deux courbes. Dans ce cas, on dit que les bornes (courbes de bornes) sont ordonnées. On peut aussi choisir d'imposer que la borne inférieure soit toujours la même courbe, de même pour la borne supérieure. Dans ce cas, on prend le risque, pour une courbe sinusoïdale, d'avoir des oppositions de phase lorsque les deux courbes se croisent. Des illustrations sont données *fig.* 6.50.

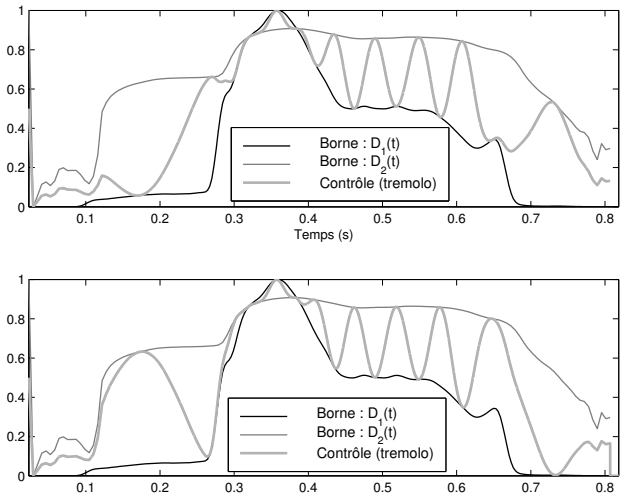


FIG. 6.50 – *Ajustement variable des bornes, pour des bornes ordonnées et non ordonnées.*

Nous avons établit les différentes étapes du premier niveau de *mapping* qui concerne le contrôle de l'effet par les descripteurs du son. Nous pouvons maintenant décrire le second niveau de *mapping*.

6.5 Second niveau de *mapping* (N2) : contrôle du premier niveau par le geste

Le second niveau de *mapping* correspond au contrôle gestuel du premier niveau. Nous présentons les points d'entrée que nous avons identifiés, ainsi que des expériences dans lesquelles nous avons mis en œuvre le contrôle gestuel.

6.5.1 Points d'entrée dans les deux niveaux du premier étage de *mapping*

Nous avons identifié plusieurs points d'entrée pour le contrôle gestuel dans les deux niveaux (N1 et N2) du premier étage (E1) de *mapping* (descripteurs du son vers contrôle de l'effet) :

- ◇ contrôle gestuel du premier niveau (combinaison de descripteurs) :
  - geste de sélection pour choisir la fonction de conformation ;
  - geste de modification pour modifier les paramètres à valeurs continues de la fonction de conformation, tels que les bornes de troncature, les exposants de la fonction puissance, la forme de la courbe logarithme ou exponentielle, etc. ;
  - geste de sélection pour modifier les paramètres à valeurs discrètes de la fonction de conformation, tel que le nombre de convolution de la fonction sinusoïde ;
  - geste de sélection pour choisir le type de combinaison : linéaire, non-linéaire (multiplication) ;
  - geste de modification pour contrôler les poids de la combinaison ;
- ◇ contrôle gestuel du deuxième étage (conformations spécifiques) :
  - quantification : geste de modification pour faire passer lentement d'un type de quantification à un autre ;
  - zoom : geste de modification pour forcer le zoom, l'accélérer ou le ralentir ; geste de sélection pour l'enclencher ;
  - ajustement aux bornes : geste de modification pour faire varier les bornes ; geste de sélection pour passer d'une préset de bornes à un autre.

De manière plus générale, il nous est apparu assez évident d'utiliser les gestes de sélection pour changer un préset (quel qu'il soit) et les gestes de modification pour modifier continûment des paramètres. Ceci dit, un grand intérêt réside dans la modification continue par interpolation linéaire pour passer d'un préset à un autre (*cf.* les GRM Tools [Favreau, 2001] qui possèdent ce

très astucieux contrôle). Ceci permet d'utiliser des gestes de modification plutôt que des gestes de sélection pour sélectionner tel ou tel préset.

6.5.2 Illustrations - Expériences

Nous avons expérimenté le contrôle gestuel d'effets adaptatifs dans trois expériences principales sous *Max/MSP* :

- la spatialisation adaptative, où le geste contrôle la largeur de l'arc, la forme de la trajectoire (*cf.* sec. 7.2.1) ;
- l'objet `vst~` qui permet d'utiliser et de contrôler tout effet au format VST ;
- l'équaliseur adaptatif (*cf.* sec. 5.6.1).

Ces expériences mettent en œuvre le *mapping* à 2 étages, avec combinaison linéaire pour le premier niveau du premier étage. Elles ont permis de formaliser l'utilisation du contrôle gestuel dans le contexte des effets audionumériques, et de trouver les points d'entrée intéressants pour le contrôle gestuel.

6.6 Interfaces graphiques pour l'utilisateur

Nous disposons maintenant à la fois de descripteurs sonores, de traitements contrôlés de manière adaptative et de moyens de contrôle. Nous allons pouvoir illustrer le contrôle en présentant les interfaces graphique utilisateur que nous avons réalisées, hors temps-réel sous *Matlab* et en temps-réel sous *Max/MSP*.

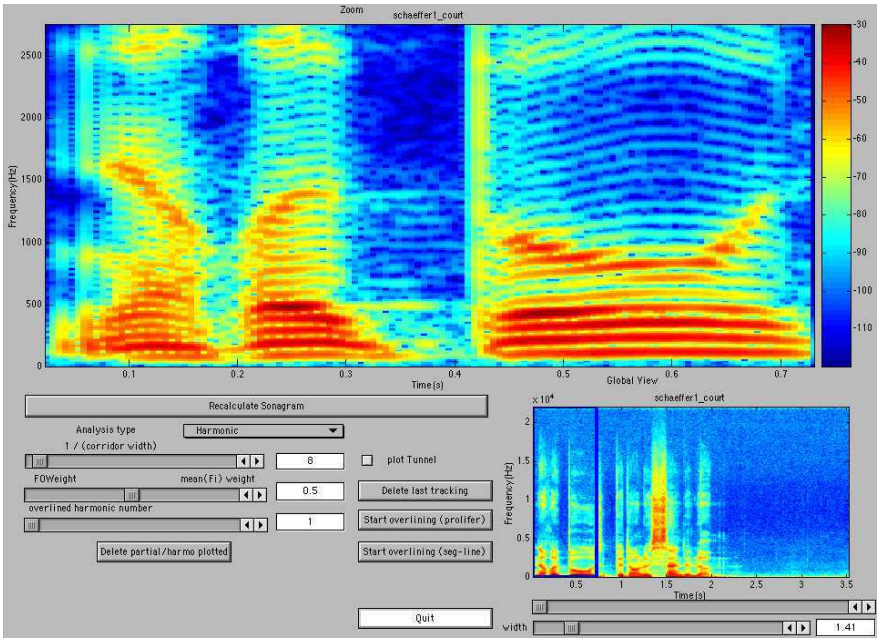


FIG. 6.51 – *Interface* Matlab *d'analyse pour l'extraction des descripteurs. Un sonagramme du son entier est indiqué en bas à droite, et une vue local (zoom) est affichée en vue du tracé manuel des partiels pour guider leur pistage lors de l'extraction.*

6.6.1 Différences entre les interfaces temps-réel et hors temps-réel

Dans une première version des effets adaptatifs hors temps-réel, l'analyse du son via ses descripteurs se faisait au fur et à mesure du traitement. C'est le meilleur choix de mise en œuvre

pour pouvoir passer à une implémentation en temps-réel, par exemple sous forme de *plug-in*. Cependant, cette mise en œuvre nous limitait en nombre de descripteurs (certains étant calculés par des programmes externes, pour toute la durée du son, d'autres n'étant pas causaux). Aussi, dans une deuxième version que nous présentons ici, nous avons fait le choix de séparer complètement l'analyse de la synthèse sous *Matlab*. Ceci implique plusieurs choses : de se limiter à des effets à post-contrôle (contrôle par le son avant l'effet), d'avoir un bien plus grand choix de descripteurs et de fonctions de *mapping*, et de ne pas pouvoir utiliser de contrôle temps-réel.

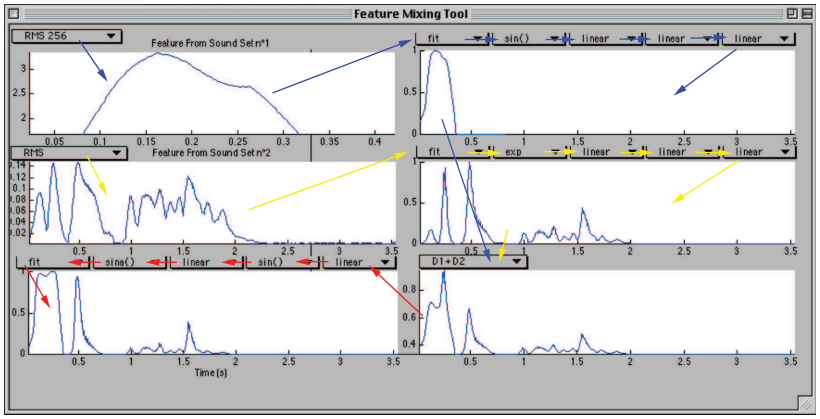


FIG. 6.52 – *Interface* Matlab *de combinaison de deux descripteurs de longueurs différentes (N1-E1) : chaque descripteur (flèches bleues et jaunes) choisi dans un menu déroulant se voit appliquer jusqu'à 5 fonctions de conformation successives (à droite) à l'aide de l'interface* fig. 6.54. *Les courbes résultantes (à droite) sont combinées (en bas), puis on applique jusqu'à cinq fonctions de conformation à la courbe obtenue (flèches rouges)).*

La mise en œuvre en temps-réel s'est faite sous *Max/MSP*. Elle permet le contrôle en temps-réel (et donc le contrôle gestuel), elle offre la possibilité de réaliser des effets adaptatifs croisés plus facilement, et des effets adaptatifs avec rétrocontrôle. On peut ainsi en jouer en direct et interagir avec le système. On dispose par contre de peu de descripteurs, mais avec toutefois le possibilité de lire les descripteurs de l'analyse sous *Matlab*, lorsqu'on traite en temps-réel un son enregistré. On peut ainsi appliquer un contrôle gestuel à la dilatation/contraction temporelle adaptative et au changement de prosodie (ceci a été développé à la conférence DAFX-03, en septembre 2003 à Londres [Arfib and Verfaille, 2003], *cf.* vidéo B.2).

6.6.2 Interface hors temps-réel (Matlab)

L'interface *Matlab* est constituée de deux programmes principaux, l'un d'analyse (*cf. fig.* 6.51) et l'autre de traitement (synthèse) avec contrôle adaptatif (*cf. fig.* 6.53 et voir vidéo B.4). L'interface pour les traitements comprend plusieurs parties permettant la combinaisons et la conformation des descripteurs (*cf. fig.* 6.54), le choix de l'effet et de ses bornes (*cf. fig.* 6.53). De plus, des interfaces spécifiques sont dédiées à la quantification (*cf. fig.* 6.56) et au zoom (*cf. fig.* 6.55). Une interface spécifique permet la combinaison de deux descripteurs dont les courbes temporelles sont de longueurs différentes (*cf. fig.* 6.52).

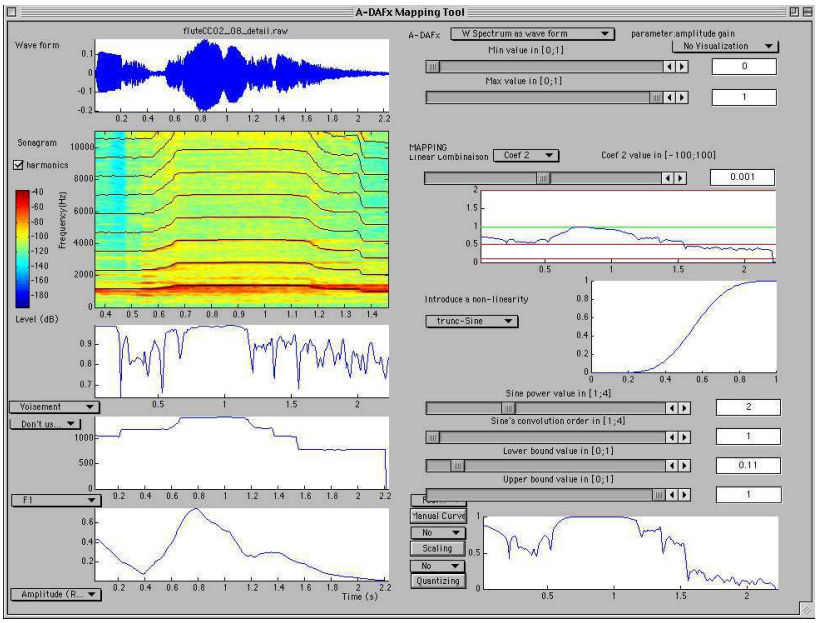


FIG. 6.53 – *Interface* Matlab *d'application de l'effet adaptatif. En haut à gauche sont indiqués la forme d'onde et le sonagramme, avec le tracé des harmoniques de l'analyse additive. En bas à gauche, trois descripteurs sont affichés (choisis dans un menu déroulant). En haut à droite, on choisit dans le menu déroulant l'effet à appliquer, et avec les curseurs les bornes des paramètres de contrôle. En bas à droite, on effectue le* mapping *pour transformer les descripteurs en contrôles (combinaison, conformation, zoom, quantification, etc.).*

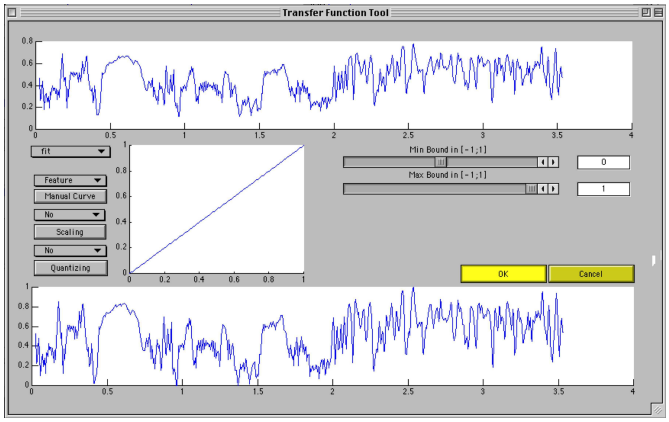


FIG. 6.54 – *Interface* Matlab *d'application d'une fonction de conformation à la courbe d'un descripteur. Le descripteur choisi par menu déroulant est affiché en haut. La fonction de conformation affichée au milieu est choisie par menu déroulant. Ses contrôles sont donnés par les curseurs. La courbe résultant est affichée en bas.*

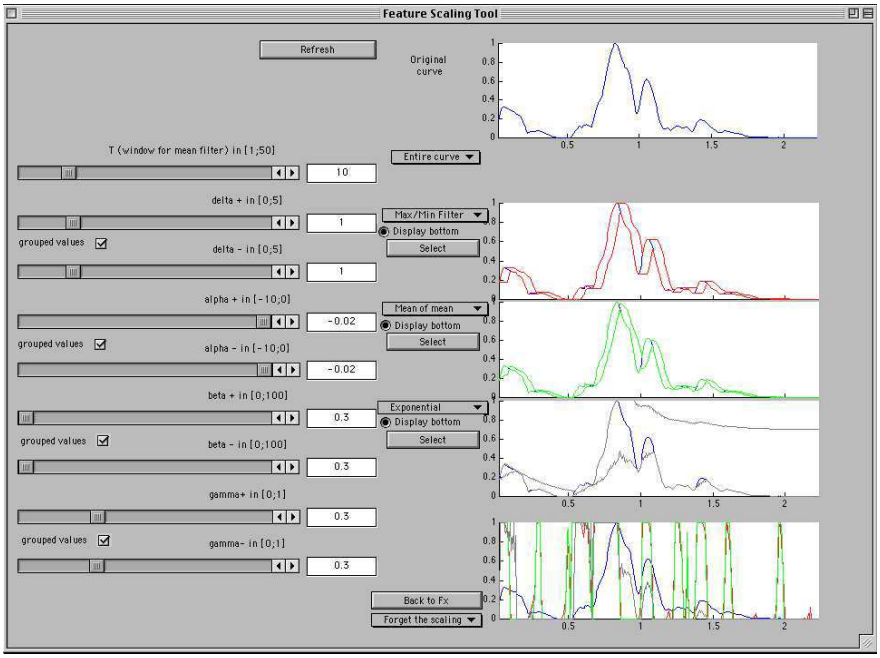


FIG. 6.55 – *Interface* Matlab *d'application d'une fonction de zoom à un descripteur. La courbe à modifier est affichée en haut. On dispose de trois graphiques pour comparer le résultat de 3 fonctions de zoom. Les bornes à partir duquel le zoom est réalisé sont superposées à la courbe d'origine, à côté de son menu déroulant. Le résultat est affiché dans le fenêtre en bas, afin de pouvoir comparer plusieurs réglages.*

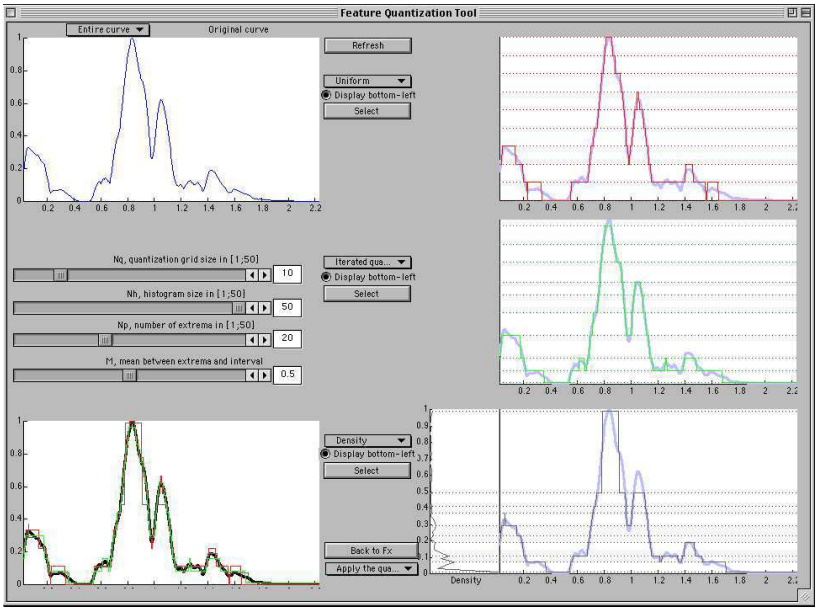


FIG. 6.56 – *Interface* Matlab *d'application d'une fonction de quantification à un descripteur. Comme pour la fonction de zoom, on dispose de trois graphiques pour pouvoir comparer le résultat de trois quantifications différentes. La courbe originale est présentée en haut à gauche, et les courbes quantifiées en bas à gauche.*

6.6.3 Interface temps-réel (Max/MSP)

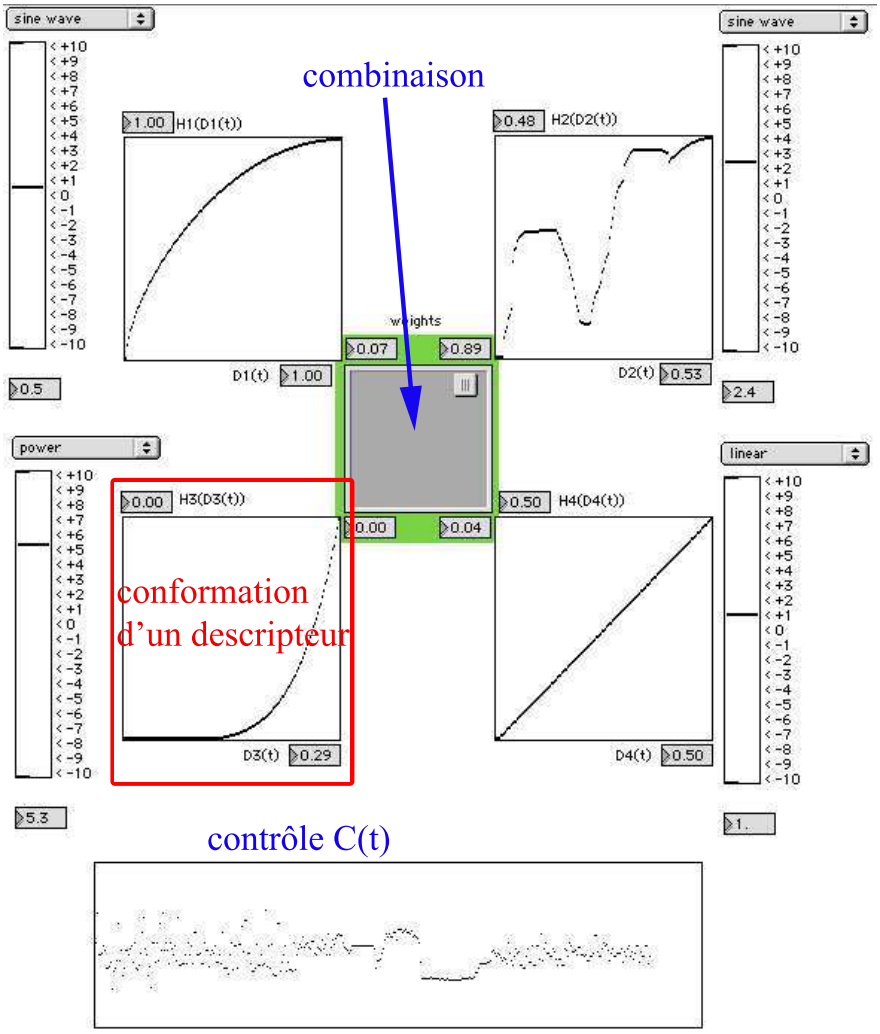


FIG. 6.57 – *Combinaison linaire de 4 descripteurs après application d'une fonction de conformation à chacun (sous* Max/MSP*). Chacune des quatre fonctions de conformation peut être affine ou donnée manuellement. Le descripteur est transformé par sa fonction de conformation. La combinaison des descripteurs après conformation se fait en déplaçant le curseur au milieu, pour pondérer chacun des 4 descripteurs. La courbe en bas est la courbe temporelle du contrôle* $C(t)$ *résultant de la combinaison.*

Pour la mise en œuvre temps-réel sous *Max/MSP*, nous avons utilisé un ensemble de descripteurs plus réduit que pour le temps différé, afin de pouvoir traiter n'importe quel son en entrée du programme. Nous préférons en effet nous placer dans l'optique du temps-réel où le signal à venir est inconnu, et nous focaliser sur le contrôle gestuel et l'interaction entre le musicien et le système. Cependant on peut utiliser un son pré-analysé et lire ses descripteurs sous forme de signaux audio à l'ouverture du *patch*, solution présentée à DAFx-03 [Arfib and Verfaille, 2003] (*cf.* vidéo B.2). L'ensemble des descripteurs utilisé est le suivant :

- le RMS (`peakamp~`) ;
- la sonie (`loudness~`, réalisé par Tristan Jehan) ;
- la brillance (`brightness~`, réalisé par Tristan Jehan) ;
- le *noisiness* (`noisiness~`) de Tristan Jehan) ;
- un détecteur de notes (`onset~`, réalisé par Tristan Jehan) ;

– un détecteur de hauteur (`fiddle~`, réalisé par Miller Puckette) ;
– l'analyse par bande filtre à Q constant (`fffb~`) ;
– le cepstre que nous avons mis en œuvre.

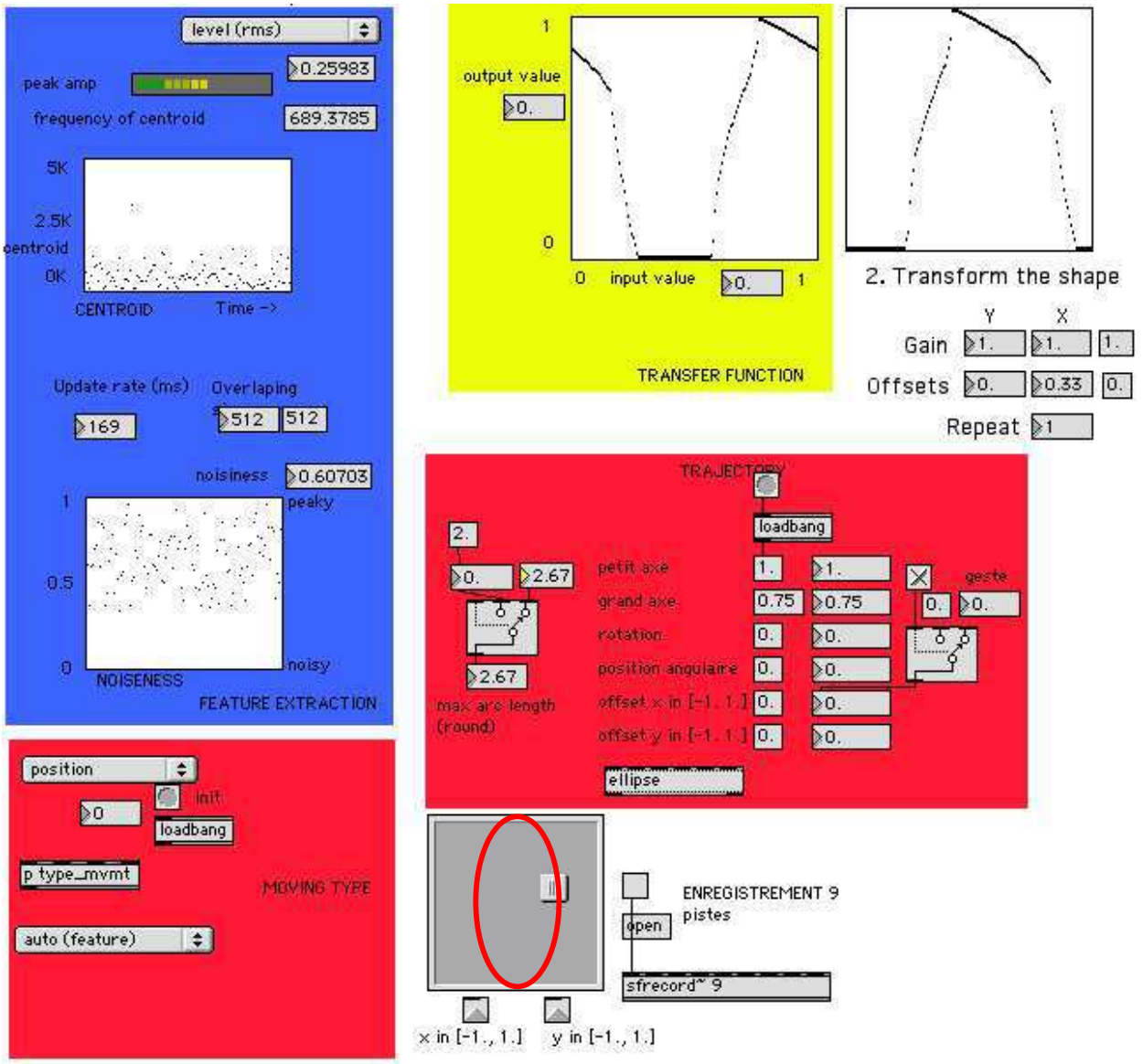


FIG. 6.58 – *Interface graphique utilisateur (*Max/MSP*) pour la spatialisation adaptative avec contrôle gestuel. La boîte bleue correspond au choix du descripteur. La boîte jaune applique la fonction de conformation, en fonction de la courbe à sa droite, donnée par l'utilisateur, et modifiée (en amplitude, par permutation circulaire, etc.) avec les valeurs de gain et d'*offset *de manière fixe ou continue (par le geste). La boîte rouge en bas sert à choisir le type de déplacement (en position, en vitesse, en accélération). La seconde boîte rouge, au milieu, permet de déformer l'ellipse et de changer sa position. Les données du contrôle gestuel agissent sur les paramètres présentés dans les boîtes rouges ou jaunes.*

Une première interface est utilisée pour combiner linéairement quatre descripteurs auxquels une seule fonction de conformation est appliquée (*cf. fig.* 6.57). Une deuxième interface permet de construire des fonctions de conformation plus complexes à partir de la combinaison de 4 fonctions de conformations affines ou données graphiquement (*cf. fig.* 6.60). Ensuite, nous présentons une partie de l'interface de la spatialisation adaptative (*cf. fig.* 6.58) et de l'équaliseur stéréophonique

adaptatif (*cf. fig.* 6.59).

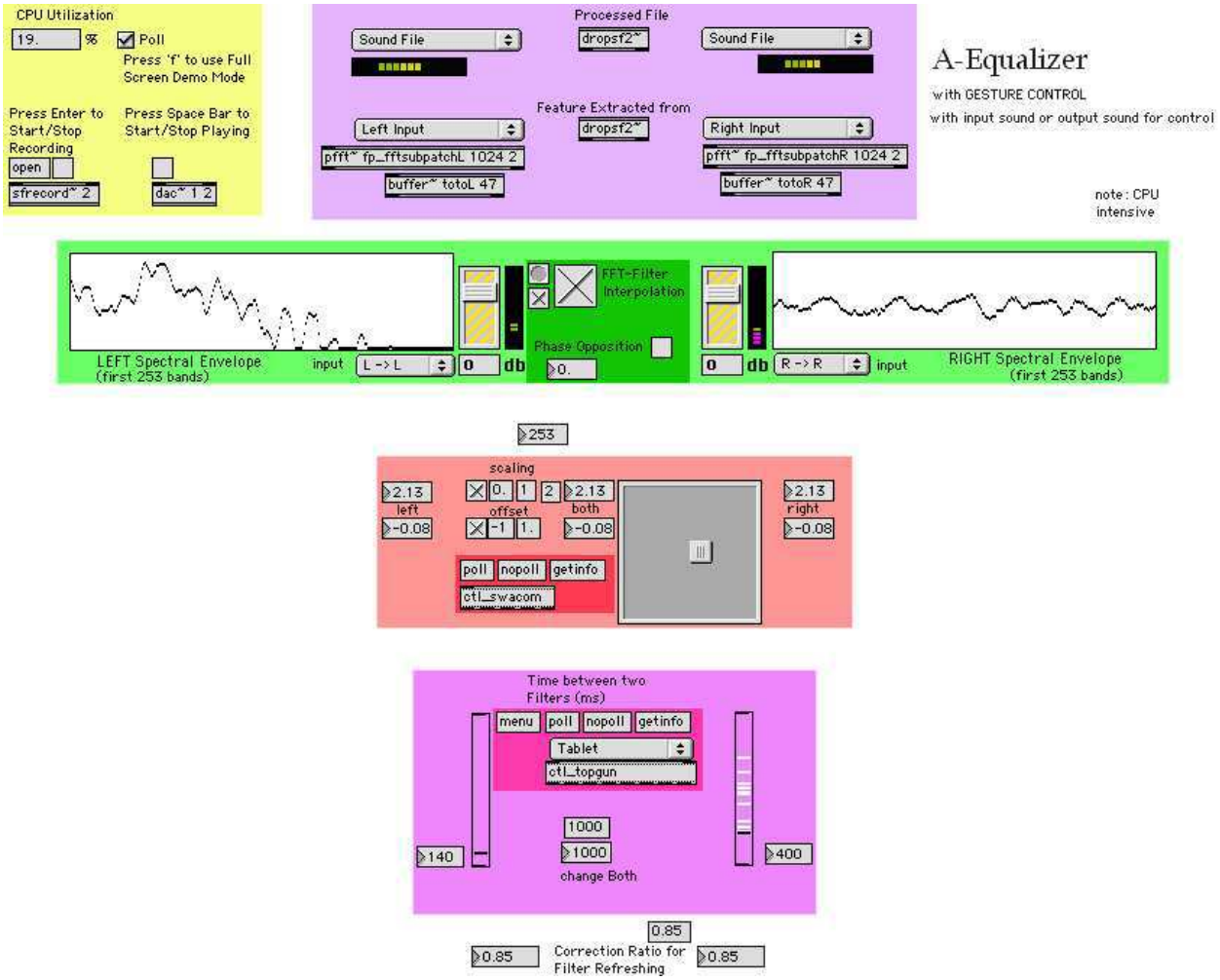


FIG. 6.59 – *Interface graphique utilisateur (*Max/MSP*) pour l'équaliseur adaptatif avec contrôle gestuel. La boîte jaune est le contrôle général de lecture et enregistrement. La boîte rose permet de choisir pour chaque canal (stéréo) le signal traité et le signal duquel les descripteurs sont extraits. La boîte verte contient le représentations TFCT de la fonction de transfert de chaque filtre. La boîte rouge sert au contrôle du changement d'échelle de la fonction de transfert (troncature, décalage vers le haut ou vers le bas et étirement), contrôlée par la tablette graphique. La boîte mauve enfin permet de contrôler (avec le joystick) la période de rafraîchissement de la fonction de transfert de chaque canal séparément.*

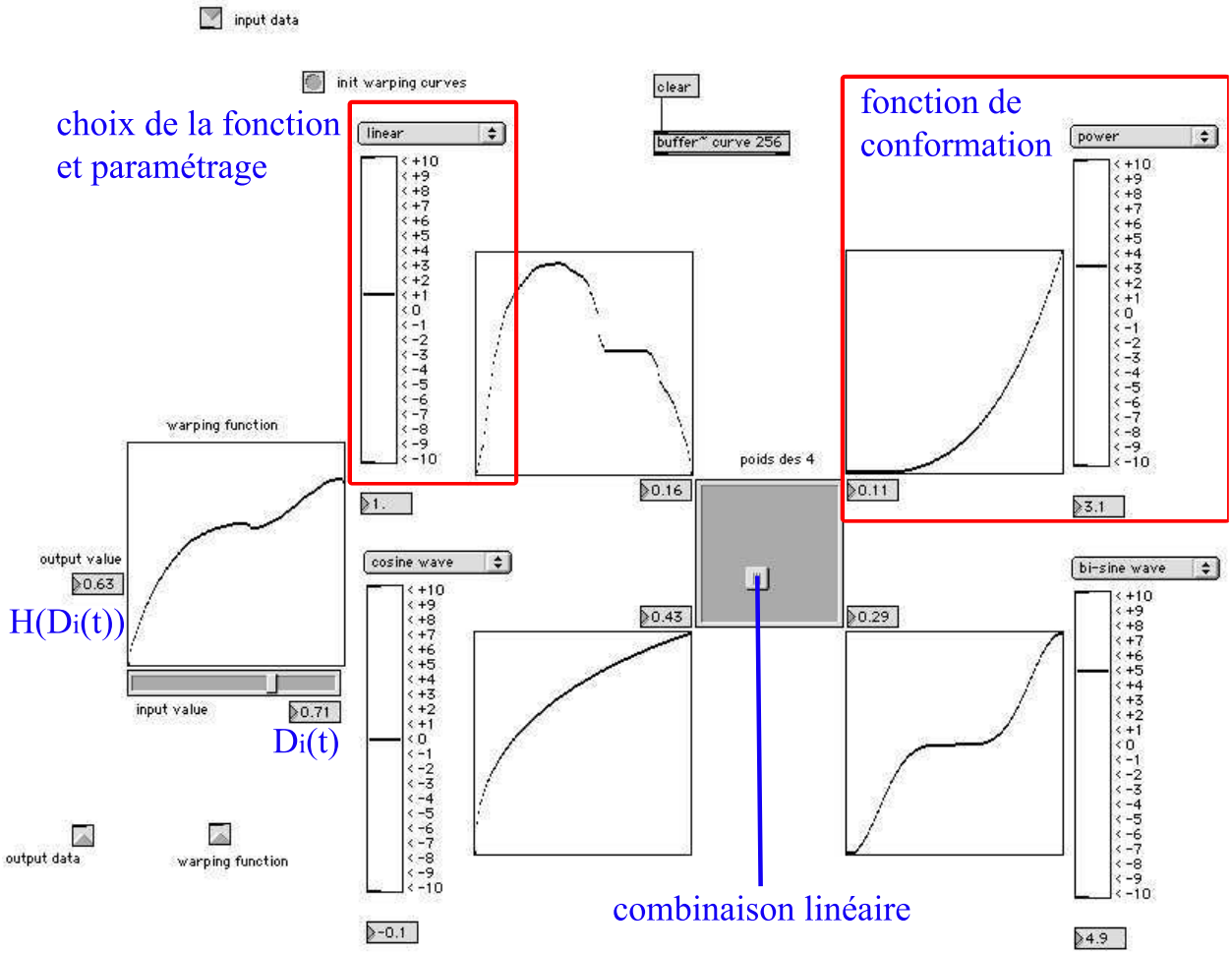


FIG. 6.60 – *Construction d'une fonction de conformation par combinaison linaire de fonctions élémentaires* Max/MSP *: chacune des quatre fonctions de conformation peut être affine ou donnée manuellement. La combinaison se fait en déplaçant le curseur au milieu, pour pondérer chacune des 4 fonctions élémentaires. La courbe de gauche est la courbe résultant de la combinaison : pour une valeur de paramètre donnée $D_i(t)$, on obtient $H(D_i(t))$.*

6.7 Conclusions

6.7.1 Intérêt du dispositif de *mapping* proposé

Nous proposons de rappeler ici les avantages qu'apporte le *mapping* que nous proposons. Tout d'abord, nous avons introduit le contrôle gestuel à un autre niveau que la simple manipulation des contrôles via des potentiomètres des effets usuels. Nous rendu possible l'utilisation des descripteurs perceptifs du son comme contrôle (via les paramètres psychoacoustiques). Le *mapping* est explicite et sa structure effectue des séparations claires entre les différentes fonctionnalités en leur assignant une place précise dans la chaîne de *mapping*. Nous avons décrit des outils de combinaison et de manipulation de courbes, dont les fonctions de conformation, le zoom pour se focaliser sur un petit espace de variation et la quantification-discrétisation pour le délai granulaire. Ce *mapping* offre la possibilité d'utiliser les descripteurs de deux sons, pour des effets croisés ou non. Pour l'utilisateur, le choix du *mapping* perceptivement intéressant est effectué suite à l'écoute de l'effet appliqué avec différentes configurations de *mapping*.

6.7.2 Prospective : apprentissage du *mapping*

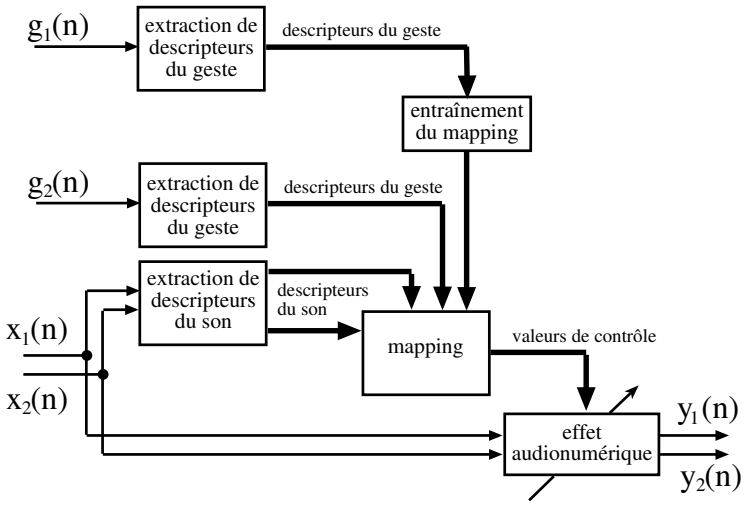


FIG. 6.61 – *Diagramme de l'effet audionumérique adaptatif croisé contrôlé gestuellement et entraîné*

Parlons maintenant de prospectives : pourquoi ne pas utiliser un schéma encore plus complet, réalisant un apprentissage des modifications de *mapping* (*cf. fig.* 6.61) effectuées par le contrôle gestuel. Ceci est encore prospectif, mais nous savons grâce à différentes études que l'on devrait pouvoir en tirer d'enrichissants résultats [Drame *et al.*, 1998; Jehan and Schoner, 2001b; Rochebois, 1997]. L'apprentissage permet, à partir d'une évolution de paramètre donnée (une courbe donnée par l'utilisateur), de calculer le *mapping* correspondant pour que, à partir des descripteurs extraits du geste et des sons, on puisse produire le même effet, sans que l'utilisateur n'ait à contrôler gestuellement l'effet. Ceci fait appel à des modèles statistiques (ACP) pour les *mappings* explicites, combinatoires (réseaux de neurones) pour les *mappings* implicites.

Chapitre 7

Spécificités, applications musicales et intérêts des effets adaptatifs

Je suis contre l'utilisation unilatérale d'une fonction : il faut en trouver d'autres. Et je ne suis pas d'accord de dire qu'un instrument a une fonction unique, qu'il doit être utilisé comme ceci ou comme cela. Lorsque Monteverdi a utilisé pour la première fois le trémolo, c'était une révolution !
Luigi Nono [Albèra, 1997b]

Sommaire

7.1 Spécificités de la mise en œuvre effectuée

7.1.1 Extraction de partiels guidée

Dans notre interface utilisateur permettant d'effectuer une extraction de descripteurs d'un son hors temps-réel (sous *Matlab*, [Mathworks, 2003]), nous avons mis en place un guidage de l'analyse additive par l'utilisateur. Ceci nous a paru important parce que les méthodes simples de poursuite de partiels [McAulay and Quatieri, 1986; Maher and Beauchamp, 1994] fonctionnent mal pour des sons réverbérés (on observe des répliques du son décalées et superposées). En général, dans le cadre de l'analyse-synthèse additive, on utilise tant que possible des sons relativement secs, sans réverbération, ou sinon on met en place des méthodes de poursuite très sophistiqués. Cela demanderait trop de temps pour en faire la revue et tester les qualités de chacune, et ce n'est pas notre propos ici. Nous avons préféré développer une analyse guidée par l'utilisateur.

Le guidage se fait comme suit : l'utilisateur dispose dans une fenêtre du sonagramme du son, avec une vue globale miniature, et une vue agrandie et parcellaire (un zoom). Ce zoom est visualisé dans la vue globale, et il peut déplacer la zone temporelle visualisée à l'aide d'un curseur. Le nombre d'harmoniques à pister est indiqué par l'utilisateur. Avec la souris, il dessine une courbe en segments de droites afin de "surligner" un partiel. Ensuite, il indique son numéro au programme. L'analyse peut alors commencer.

L'analyse se fait harmonique par harmonique : pour chaque harmonique, on dispose d'un guide, et on recherche le maximum local (dans une zone limitée, que nous appelons tunnel) sur deux trames temporelles successives. Le tunnel est défini par l'intervalle $[\hat{f}_i - \frac{\hat{f}_i}{i\,\alpha}; \hat{f}_i + \frac{\hat{f}_i}{i\,\alpha}]$. Plus le paramètre α est grand, plus le tunnel est étroit. La fréquence du partiel est alors calculée à partir du déroulement de phase, de manière classique. Le guide est remis à jour pour chaque harmonique, selon un calcul de pondération entre le guide original donné par l'utilisateur et la fréquence fondamentale calculée à partir des harmoniques déjà pistées. Si l'on note le guide $\hat{f}_1$ donné par l'utilisateur et f_{i-1} la $(i-1)^{\text{e}}$ harmonique pistée, la suivante sera calculée selon :

$$\hat{f}_i(t) = (1-\beta) i \hat{f}_1(t) + \beta \sum_{k=1}^{i-1} \frac{f_k(t)}{k} \frac{\rho_k(t)}{\sum_{l=1}^{i-1} \rho_l(t)} \tag{7.1}$$

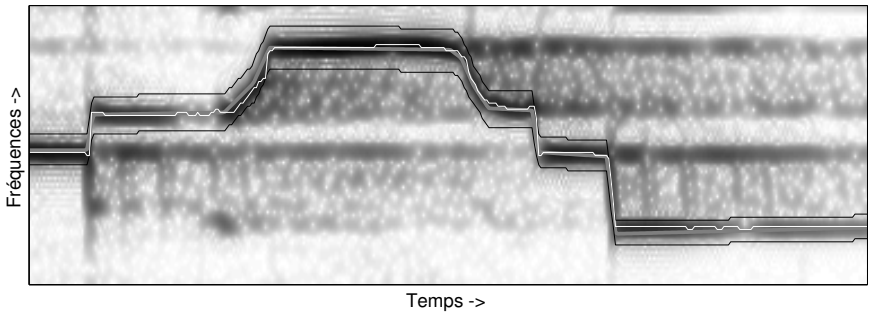


FIG. 7.1 – *Extraction de partiels guidée par l'utilisateur : définition du guide (trait gris large), puis recherche du partiel (trait blanc) dans un tunnel (traits noirs).*

Cette méthode permet de forcer l'analyse à suivre les partiels, sans s'égarer lorsque le son est réverbéré. De plus, elle permet de prendre en compte des sons légèrement inharmoniques, voire de définir séparément, un par un, des partiels pour un son inharmonique.

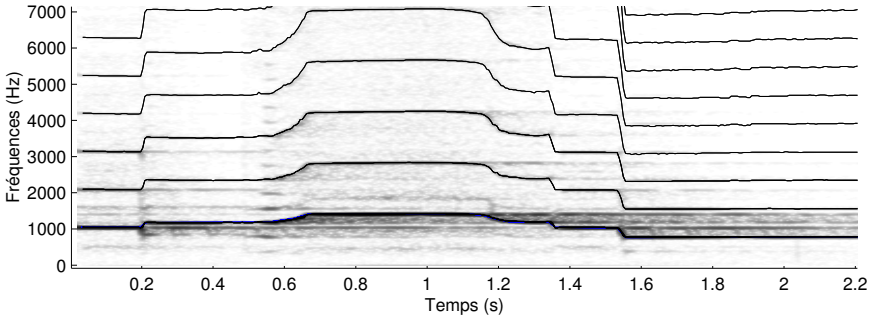


FIG. 7.2 – *Sonagramme et ses partiels pistées.*

7.1.2 Mise en œuvre bloc par bloc, avec pas et taille de fenêtres variables et/ou constants

Pour les effets temps-réel traditionnels, le son est traité sur des fenêtres temporelles de longueur et de pas constant [Arfib, 1998a]. Pour les effets adaptatifs, au contraire, le pas comme la longueur de la fenêtre (du grain) peuvent varier, ce qui peut poser des problèmes pour la mise en œuvre en temps-réel. Remarquons que le traitement lui-même est la plupart du temps appliqué fenêtre par fenêtre (traitements spectraux et temps-fréquence), parfois par pseudo-vecteur (traitements temporels), mais les données sont transmises à l'algorithme par pseudo-vecteur. Cela signifie qu'il y a deux niveaux de calcul : celui du traitement lui-même (bloc, fenêtre, ou grain) et celui de la procédure (pseudo-vecteur).

La mise en œuvre bloc par bloc a été utilisée en temps différé (*Matlab*) pour les effets à l'aide de la technique du vocodeur de phase : la transposition (et l'harmonisation le vibrato, le martien), les effets sur l'enveloppe spectrale (équalisation, changement de couleur ou de voyelle, décalage ou étirement de l'enveloppe, conformation spectrale, trémolo spectral, modulation en anneau spectrale, compresseur spectral, etc.), les effets sur le spectre (robotisation, chuchotement, etc.), les effets sur la durée (dilatation/contraction temporelle).

Taille de fenêtres et pas constants $L_A^w, L_S^w = cte$, $R_A, R_S = cte$

Les effets non adaptatifs fonctionnant bloc par bloc utilisent des tailles de fenêtres d'analyse et de synthèse constantes, ainsi que des pas constants. C'est le cas le plus simple, puisque si la fenêtre de synthèse est bien choisie, le son est normalisé, à une constante multiplicative près. Le vocodeur de phase et l'analyse-synthèse additive traditionnels utilisent cette configuration. Dans l'exemple donné *fig.* 7.3, la fenêtre d'analyse et de synthèse est une fenêtre de Hanning. Lors de la resynthèse, le résultat serait identique si on utilisait une fenêtre triangulaire.

Taille de fenêtres et pas de synthèse constant, et pas d'analyse variable $L_A^w, L_S^w = cte$, $R_A = f(t), R_S = cte$

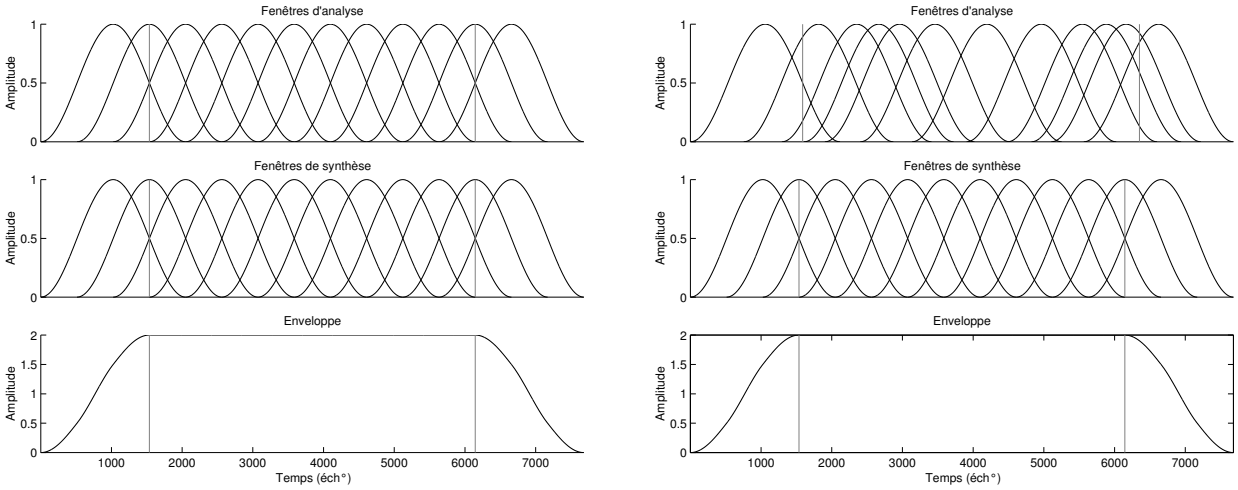


FIG. 7.3 – *Mise en œuvre bloc par bloc avec $L_A^w = cte$, $L_S^w = cte$, $R_S = cte$.*
A gauche : $R_A = cte$, le son est normalisé en amplitude.
A droite : $R_A \neq cte$, il n'est pas nécessaire de renormaliser, mais le pas de synthèse doit être suffisamment petit pour que le recouvrement à l'analyse correct.

Il n'y a aucun problème de normalisation lorsque le pas d'analyse est variable et le pas de synthèse est constant (*cf. fig.* 7.3 droite). C'est l'une des mise en œuvre possible de la dilatation/contraction temporelle adaptative. Nous l'avons utilisée pour cette qualité : le temps de calcul du son de synthèse s'en trouve diminué, puisqu'il ne faut pas calculer l'enveloppe ni effectuer la post-correction. Un

reproche que l'on pourrait faire à cette mise en œuvre dans le cas précis de la dilatation/contraction temporelle, c'est que pour des pas d'analyse grands (ie. un fort taux de contraction) et supérieurs à la taille de la fenêtre d'analyse, une partie de l'information du signal d'analyse soit perdue et ne figure pas dans le son de synthèse. Pour éviter ceci, il faut choisir un pas de synthèse très petit, de manière à ce que le pas d'analyse permette un recouvrement correct du son d'analyse.

Taille de fenêtre d'analyse et pas constants, taille de fenêtre de synthèse variable $L_A^w, L_S^w = f(t)$, $R_A, R_S = cte$

Pour la mise en œuvre adaptative de la transposition par vocodeur de phase, avec la méthode du Cepstre pour la conservation de l'enveloppe et le ré-échantilonnage pour la transposition, (*sec.* 5.4.2), la taille de la fenêtre de synthèse L_S^w varie du fait du ré-échantillonnage à taux variable. Le pas de synthèse R_S doit être choisi pour que le recouvrement soit correct, c'est-à-dire :

$$R_S \leq \frac{\min(L_S^w)}{K} \quad \text{avec} \quad K \in \mathbb{N} \quad \text{et} \quad K \geq 2 \tag{7.2}$$

Sans modification du traitement (*cf. fig.* 7.4), le son de synthèse n'est pas normalisé en énergie. Il faut donc calculer l'enveloppe d'amplitude afin de pouvoir effectuer une correction d'amplitude échantillon par échantillon.

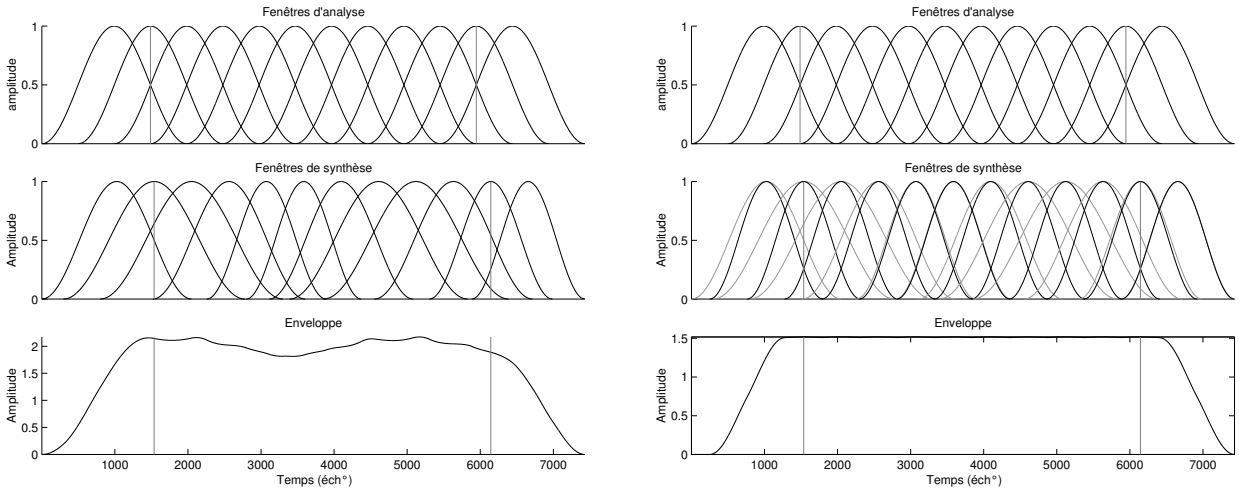


FIG. 7.4 – *Mise en œuvre bloc par bloc avec $L_A^w = cte$, $R_A = cte$, $R_S = cte$.*
A gauche : $L_S^w \neq cte$, le son de synthèse n'est pas renormalisé, une post-correction est nécessaire.
A droite : $L_S^w = \min L_S^w(t)$, l'utilisation d'une fenêtre de synthèse constante corrige la normalisation du son de synthèse et la post-correction n'est plus nécessaire.

Une solution que nous avons proposé consiste à utiliser une fenêtre de synthèse de la taille de la plus petite fenêtre de synthèse. Du fait que la fenêtre est de taille constante, il n'y a plus à effectuer de post-correction d'amplitude (*cf. fig.* 7.4).

Taille de fenêtres et pas d'analyse constant, et pas de synthèse variable $L_A^w, L_S^w = cte$, $R_A = cte, R_S = f(t)$

Dans le cas où les tailles de fenêtres et le pas d'analyse sont constantes et le pas de synthèse variable, le signal doit être renormalisé. Il faut une fois de plus calculer l'enveloppe d'amplitude et appliquer une post-correction (*cf. fig.* 7.5).
Une solution que nous proposons dans ce cas consiste à ne plus utiliser de fenêtre de Hanning à la synthèse, mais un fondu-enchaîné par segments de droites : la fenêtre est triangulaire asymétrique, ce qui permet que deux fenêtres successives apportent des contributions complémentaires. Le son de synthèse est alors normalisé automatiquement.

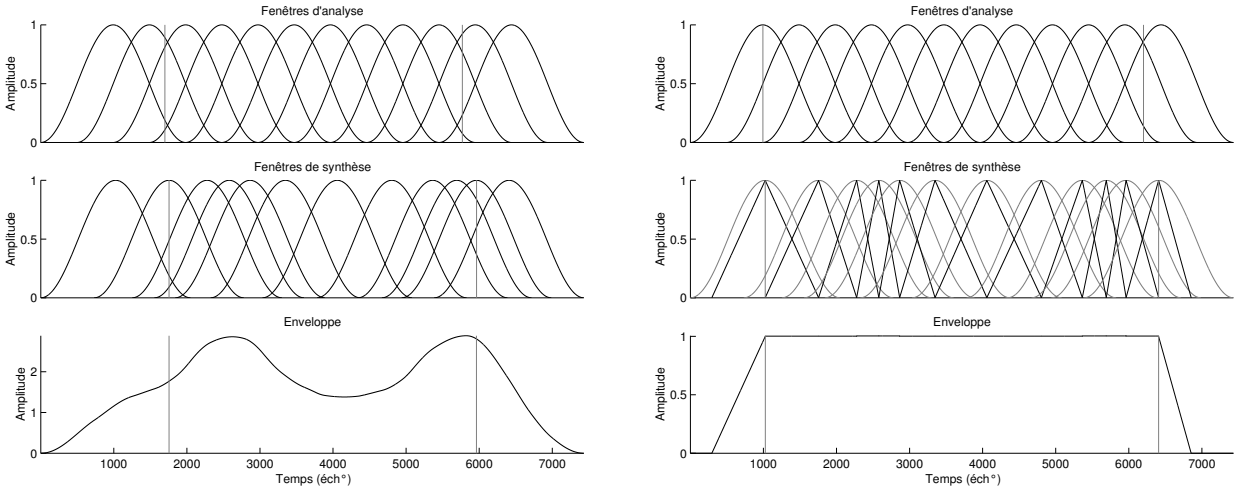


FIG. 7.5 – *Mise en œuvre bloc par bloc avec $L_A^w = cte$, $R_A = cte$, $R_S \neq cte$.*
A gauche : $L_S^w = cte$, il faut renormaliser le signal, et le pas d'analyse doit être suffisamment petit pour que le recouvrement à la synthèse soit correct.
A droite : $L_S^w \neq cte$, l'utilisation de fenêtres triangulaires asymétriques permet une normalisation.

Taille de fenêtres et pas variables $L_A^w, L_S^w, R_A, R_S = f(t)$

Lorsque à la fois les tailles de fenêtres et les pas peuvent varier, on utilise une technique développée précédemment, à savoir l'utilisation d'une fenêtre de synthèse de taille constante correspondant à la taille de la plus petite fenêtre. On s'assure de plus que le recouvrement est suffisant, en choisissant bien le plus grand pas d'analyse.

Un cas particulier est celui de la robotisation adaptative : dans ce cas, le pas et la taille de fenêtre varient, mais le problème de recouvrement et de normalisation ne se pose pas, puisque les pas peuvent volontairement dépasser la taille de fenêtre, les valeurs intermédiaires entre deux fenêtres étant remplacées par des 0.

7.1.3 Mise en œuvre pseudo-vectorielle

La mise en œuvre pseudo-vectorielle doit être modifiée pour permettre d'appliquer des traitements avec pas et taille de fenêtres variables et/ou constants.

Double boucle de traitement pour les sons longs

Afin de pouvoir traiter de longs fichiers son en s'affranchissant de la limite en taille de mémoire vive, nous avons mise en œuvre un traitement pseudo-vectoriel, où le fichier son est lu par vecteurs (de l'ordre de 1 seconde de son, soit $44,1\ kHz$). Le pseudo-vecteur est un grand bloc, dont la taille est inférieure à celle du son dans son intégralité, et supérieure à la taille de bloc. Le traitement du vecteur se fait par une boucle travaillant bloc par bloc. Il y a donc deux niveaux de boucles de traitement : le niveau global des vecteurs, et le niveau local des blocs. Pratiquement tous les effets mis en œuvre hors temps-réel utilisent la double boucle de traitement (par vecteur puis par bloc). Seuls les effets hors temps-réel dans le domaine temporel (à savoir les effets adaptatifs sur le niveau, la panoramisation adaptative, la distorsion, le compresseur) utilisent le traitement vecteur par vecteur directement, sans bloc, du fait qu'ils permettent aussi une implémentation échantillon par échantillon.

Concernant le traitement à deux boucles imbriquées, lorsque l'on passe d'un vecteur au suivant, il faut prendre une précaution élémentaire si l'on veut s'assurer de traiter tous les échantillons. On utilise un bloc de recouvrement entre deux vecteurs successifs. La taille de ce bloc de recouvrement

varie en fonction de la taille de fenêtre de synthèse. Si la fenêtre de synthèse est de taille fixe, le bloc de recouvrement est de la taille de la fenêtre de synthèse, quel que soit le pas de synthèse. Si par contre la taille de cette fenêtre de synthèse varie, il faut que le bloc de recouvrement soit de la taille de la plus grande fenêtre, et ce quel que soit le pas de synthèse.

Récapitulatif

Nous donnons dans le tableau 7.1 un récapitulatif des situations de mises en œuvre bloc par bloc lorsque les pas et les tailles de fenêtres (à l'analyse et à la synthèse) peuvent varier. Nous y précisons s'il faut appliquer une post-correction, une solution intermédiaire éventuelle, et la taille du bloc de recouvrement.

L_A^w	L_S^w	R_A	R_S	**normalisation**	**vectorisation**
constant	constant	constant	constant	OK	$B = L_S^w$
variable	constant	constant	constant	OK	$B = L_S^w$
constant	constant	variable	constant	OK	$B = L_S^w$
constant	variable	constant	constant	nécessaire ou utiliser $\min(L_S^w)$	$B = \max(L_S^w)$ $B = \max(L_S^w)$
constant	constant	constant	variable	nécessaire ou utiliser w_{triang}	$B = L_S^w$
constant	variable	constant	variable	nécessaire ou utiliser w_{triang}	$B = \max(L_S^w)$

TAB. 7.1 – *Mise en œuvre par bloc : nécessité ou non d'une post-correction de normalisation, taille du recouvrement de bloc.*

7.2 Applications musicales

7.2.1 Spatialisation adaptative au CICM, Paris VIII

Introduction, principe

On cherche des manières de spatialiser un son monophonique sur un système multi haut-parleurs (dans notre cas, le système est octophonique). Pour ce faire, on dispose déjà d'un système logiciel de spatialisation, ou plus précisément de positionnement grâce à l'intensité, comme décrit précédemment (Ambisonic), permettant de calculer les signaux sur les 8 haut-parleurs afin de simuler la position demandée.

Les contraintes que nous nous sommes fixées sont les suivantes : on désire que le son puisse se déplacer : soit seul, sans action extérieure, soit en fonction de gestes ; de plus, le système doit être temps-réel. Le système complet répondant à ces contraintes utilise un contrôle adaptatif de la spatialisation (voir le principe des effets adaptatifs [Verfaille, 2003]), ainsi que des capteurs à ultrasons, utilisés comme transducteurs gestuels, dans un dispositif développé par Todor Todoroff [Todoroff, 2003] pour une étude menée par Anne Sédès [Sédès, 2003] avec les danseuses Katia Légeret [Légeret, 2003] et Laurence Marthouret [Marthouret, 2003].

Le logiciel est développé sous l'environnement temps-réel *Max/MSP*, et le contrôle gestuel est effectué par un interprète (une danseuse dans les expériences que nous avons menées). Dans ce dispositif précis, le son est généré par la danseuse ; il pourrait toutefois provenir d'enregistrement, ou d'une pièce produite en direct.

Afin de permettre à l'auditeur et à l'interprète de mieux comprendre ce qui se passe spatialement, une représentation 3D du son et de la position de l'interprète est donnée à l'écran à l'aide du logiciel graphique Jitter. Ainsi, nous utilisons les outils de visualisation, de spatialisation et de description de trajectoires de l'espace sonore, présentés précédemment, dans le cadre d'un projet artistique.

Mise en œuvre

La chaîne de traitement de la spatialisation adaptative est la suivante : le son est analysé et des paramètres descriptifs $D_i(t)$ en sont extraits. Ensuite, ces paramètres sont combinés en une courbe $p(t)$. On applique ensuite à cette courbe de combinaison une fonction de conformation non linéaire H. Cette courbe après la fonction de conformation $H(p(t))$ correspond soit au déplacement sur une trajectoire, soit à la déformation de la trajectoire, soit aux deux.

Le déplacement sur une trajectoire s'obtient en donnant la position, la vitesse, ou l'accélération de la source sonore : elle parcourt d'une manière non linéaire une trajectoire prédéfinie. Cette trajectoire peut être donnée analytiquement, ou graphiquement par l'utilisateur, à l'aide de l'interface présentée dans cet article. Dans le cas de la déformation de la trajectoire, on utilise le paramètre gestuel pour modifier l'un des paramètres de la courbe définie analytiquement. Dans le cas où le geste modifie à la fois le déplacement sur une trajectoire et la déformation de cette trajectoire, il vaut mieux utiliser plusieurs paramètres, un pour chaque type de modification.

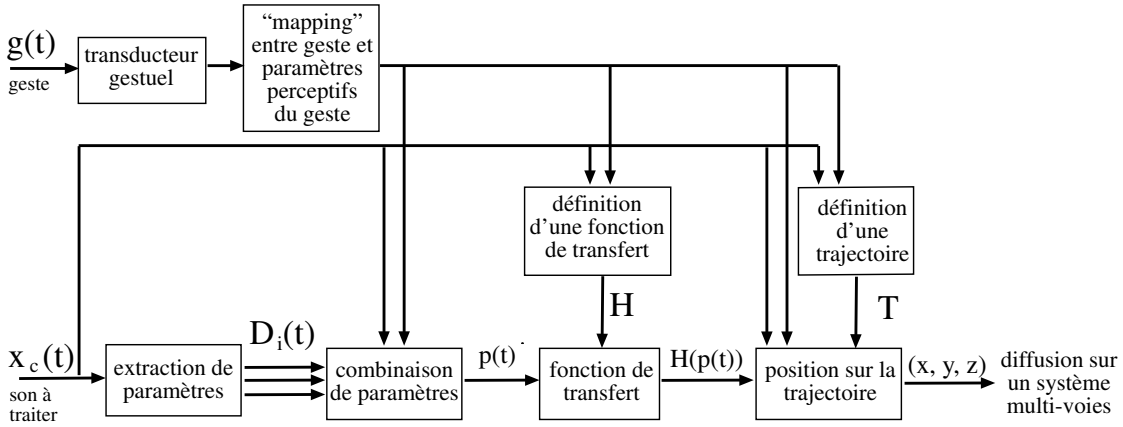


FIG. 7.6 – *Diagramme de la spatialisation adaptative contrôlée gestuellement.*

Le geste peut alors agir sur différentes étapes du traitement : pour la combinaison des paramètres, pour déformer, translater la fonction de conformation, pour choisir le type de déplacement (position, vitesse, accélération) sur la trajectoire, ou enfin pour modifier cette dernière.

La visualisation à l'écran de la position de l'interprète détectée par les capteurs ultrasons et de la position du son permet de mieux comprendre les implications des gestes sur la spatialisation.

Quelques configurations

Nous avons identifié quelques configurations principales, dont les résultats sonores sont perceptivement intéressants. Nous les présentons par la suite : déplacement automatique, modification de longueur d'arc, position donnée par le geste (ou contrôle direct), modification de fonction de conformation, modification de trajectoire, modification d'ouverture spatiale.

- **Déplacement automatique** : un paramètre du son donne la position du son x (ou sa vitesse v, ou son accélération a) sur une trajectoire donnée T, avec une combinaison de paramètres et une fonction de conformation donnée. Par exemple, pour une trajectoire circulaire, lorsque la vitesse est donnée par le RMS (mesure de l'énergie du signal) et la fonction de conformation par l'une des deux courbes *fig.* 7.7, alors le son se déplacera soit pendant les silences (fonction de conformation de gauche), soit pendant les sons de forts niveaux (fonction de conformation de droite).
- **Modification de longueur d'arc** : étant donnée une trajectoire curviligne $(x, y) = (x(t), y(t))$ (fixe ou variable) la source est placée par exemple en fonction du CGS (centre de gravité spectral ou centroïde, descripteur de l'ouverture du spectre fréquentiel, et donc du timbre),

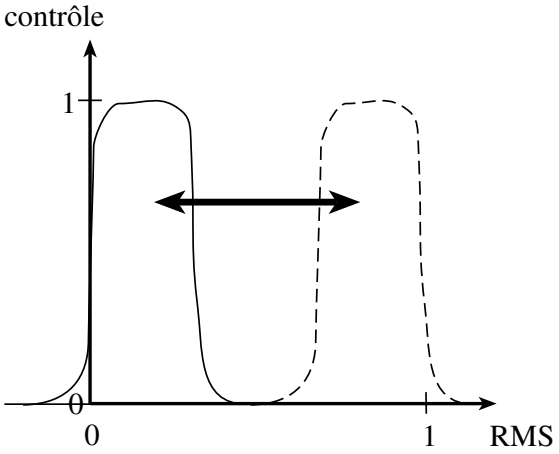


FIG. 7.7 – *Exemple de fonction de conformation pour la spatialisation adaptative : déplacement de la source uniquement pendant les silences (gauche) ou pendant les sons forts (droite) ; le paramètre extrait du son est alors le RMS.*

sur un arc de la trajectoire $[t^-; t^+]$ de longueur variable, donnée par un geste ou un paramètre à évolution lente. Si l'arc possible est petit, le son reste local ou en tout cas localisé près d'un point, qui semblera fixe ; si l'arc fait un demi-cercle ou un cercle complet, le son semble multi-localisé, plusieurs sources différentes apparaissent car les mouvements sont trop rapides pour être perçus comme tels, et trop différents de mouvements sinusoïdaux pour donner l'effet modulation en anneau attendu. En fait, le son est segmenté en plusieurs flux auditifs de propriétés distinctes mais variant avec cohérence : à la fois en localisation et en énergie si le RMS guide le déplacement ; à la fois en localisation et en enveloppe spectrale si c'est le CGS qui guide le déplacement du son. Le fait que deux dimensions (localisation et énergie ou centroïde) soit cohérentes aident grandement à dissocier ces flux auditifs. Ceci corrobore des résultats connus sur l'analyse de scènes auditives et la ségrégation de flux [Bregman, 1990].

- **Contrôle direct** : à partir de la position de la danseuse (captée par deux transducteurs à ultrasons, par exemple), on positionne le son selon des règles de symétrie centrale ou axiale, de translation, d'homothétie, à partir des coordonnées en cartésien (x, y) ou polaire (ρ, ϕ). Ainsi, le geste contrôle directement la position du son.
- **Fonction de conformation** : le geste modifie la fonction de conformation : par décalage en x avec une fenêtre rectangulaire, *cf. fig.* 7.7 ; par décalage en y (lorsqu'on augmente le décalage en y, on augmente la vitesse moyenne de déplacement) ; ou par un *morphing* plus complexe.
- **Ouverture spatiale** : le son est séparé en fréquences selon 8 bandes (à l'aide de l'objet `fffb~`), chaque bande étant positionnée à un emplacement fixe ou mobile. Dans le cas de positions fixes, le son s'ouvre spatialement lorsque le spectre s'enrichit ; le résultat sonore ne correspond pas à une ségrégation de flux, bien que l'ouverture spatiale soit corrélée à l'ouverture spectrale. Dans le cas de positions variables, et donc de déplacements des différentes bandes de fréquences, le son est plus difficile à décrire et à comprendre perceptivement. Une vie interne au son se met en place, et s'externalise par la spatialisation.
- **trajectoire** : soit une ellipse comme exemple de base. Son équation est :

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1 \tag{7.3}$$

Supposons que le son se déplace en position ou en vitesse sur l'ellipse. Le geste va déformer l'ellipse (la trajectoire) de plusieurs manières : par décalage selon l'un ou plusieurs des axes, et par changement de la valeur de l'un ou des deux axes (de longueurs a et b), de façon à pouvoir passer du cercle au segment en passant par l'ellipse, de manière continue.

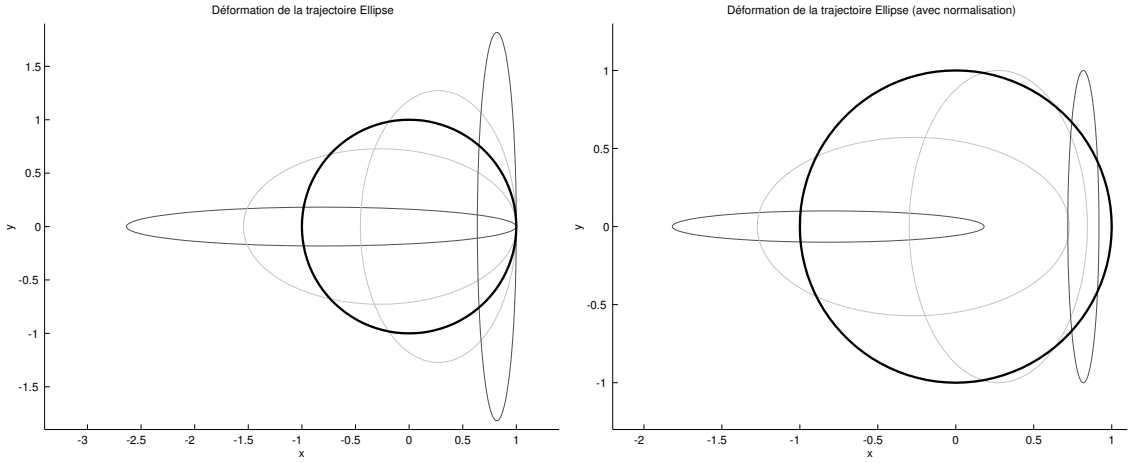


FIG. 7.8 – *Exemples de déformation de trajectoire (ellipse) par le geste : les axes de l'ellipse changent de taille, et permettent de passer du déplacement sur un cercle au déplacement sur un segment en passant par l'ellipse, le tout avec une ellipse dont les axes sont normalisés (*fig. *gauche) ou non (*fig. *droite). De plus, l'ellipse est translatée de gauche à droite au fur et à mesure que ses axes sont modifiés. Le cercle au trait épais correspond au cercle sur lequel les enceintes sont positionnées ; il se confond avec la trajectoire circulaire sans translation.*

Visualisation

Le système de spatialisation adaptative a été combiné avec le système de représentation 3D et de navigation dans l'espace sonore. La position donnée pour la danseuse correspond en fait à la position donnée par la dernière interruption du faisceau vertical et la dernière interruption du faisceau horizontal, qui peut être totalement différente de la position réelle de la danseuse dès lors qu'elle sort de l'un des deux faisceaux. Notons que le point de vue utilisé ici ne correspond pas au point d'écoute, qui est placé au centre du dispositif. Ici, la déambulation n'est pas rendu possible, mais le même système de représentation avec *Jitter* et la librairie graphique *OpenGL* est utilisé.

Ce retour visuel (ou *feedback*) déjà évoqué dans [Verfaille, 2003; Couturier, 2003] permet dans un premier temps de calibrer le système, mais surtout dans un second temps d'avoir la possibilité de contrôler plus finement les implications des interruptions de faisceaux ultrasons, et de choisir quand et comment on agit sur la synthèse du son ainsi que sur sa spatialisation, de manière conjointe ou séparée. En effet, l'image est présentée en continu à la danseuse via l'écran.

Implications sur le son et le geste

Les possibilités offertes par la spatialisation adaptative contrôlée gestuellement sont immenses. Nous n'avons ici exploré qu'une petite partie des possibilités offertes par ce système de double contrôle par le son et par le geste. La configuration d'ellipse avec déformation et translation a été mise en place au fur et à mesure du développement du système, en suivant les demandes précises de l'interprète. La dimension X était utilisée pour générer le son, et la dimension verticale Z pour la spatialisation. Le son était spatialisé de manière adaptative, se déplaçant de lui-même en position sur l'ellipse. La forme et la translation de l'ellipse quant à elles étaient données par le geste. Cette configuration a ceci d'intéressant que les positions possibles du son suivent l'évolution verticale de l'interprète ; le champ d'action sur la spatialisation peut être vu comme un cylindre central.

Lorsqu'on inverse les capteurs, c'est-à-dire lorsque le son est généré selon la dimension Z et la spatialisation contrôlée selon la dimension X, les positions possibles du son suivent l'évolution horizontale de l'interprète. Ainsi, lorsqu'il est face au capteur, à droite de la salle (à droite de la *fig.* 7.9 gauche), le son se déplace devant lui, de gauche à droite. Au centre, le son se place partout autour de lui, tandis qu'à gauche, il évolue sur un segment placé derrière lui. L'évolution spatiale

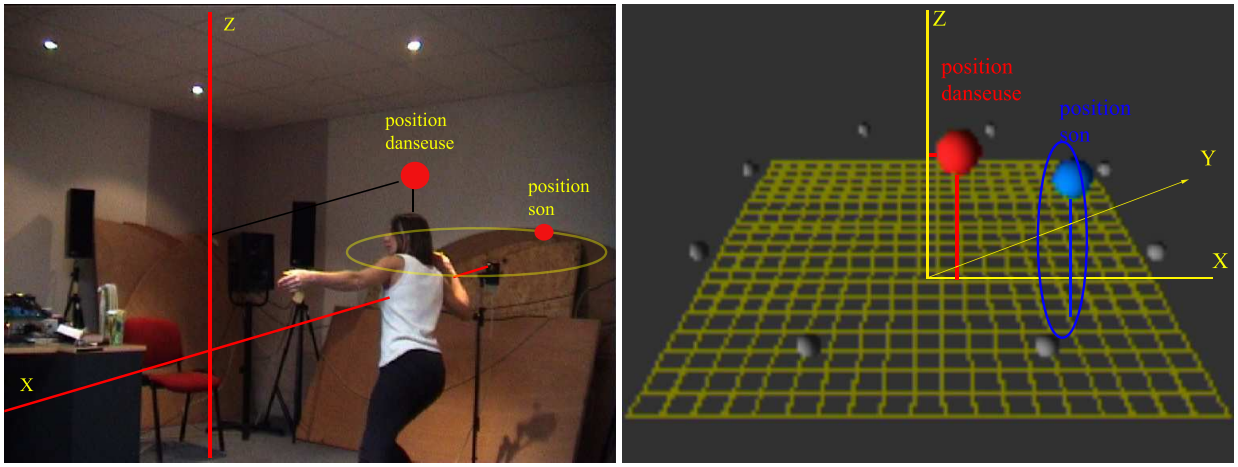


FIG. 7.9 – *Spatialisation adaptative : représentation 3D de la scène. A gauche, installation en cours de jeu : la danseuse (ici Cécile Vallet) dispose de l'espace entre les haut-parleurs pour sa chorégraphie, improvisée et adaptée au son généré par le système. A droite, représentation 3D de la position du son, de l'interprète et des 8 haut-parleurs à l'aide du système de déambulation. Notons que les angles de vue des deux figures sont différents : l'angle de vue de la figure de gauche correspond à la direction de l'axe Y sur la figure de droite. La position du son est donnée par la moyenne sphère à droite (en bleu sur la vidéo). La position de la danseuse, ou du moins la position donnée par les dernières interruptions des faisceaux vertical et horizontal, est donnée par la grande sphère au milieu (en rouge sur la vidéo).*

est donc plus corrélée à la position de l'interprète sur la ligne du capteur.

La configuration où la dimension verticale donne la longueur de l'arc d'ellipse sur lequel le son peut se déplacer permet quant à elle de passer d'un son localisé en une position choisie par l'interprète lorsqu'il entre dans le faisceau vertical, à hauteur d'épaule, à un son multi-localisé lorsque l'interruption du faisceau vertical se fait entre les épaules et le sol. Des mouvements de jambes ou de pieds influent sur la localisation précise ou floue du son.

Ces trois exemples montrent la richesse de jeu que l'on peut attendre d'un tel système.

D'autre part, le retour visuel présente un grand intérêt, quelle que soit la configuration. Le retour visuel se compose à la fois de la visualisation du son ou de la dernière interruption des faisceaux par l'interprète, ainsi que des diodes positionnées sur les capteurs ultrasons qui permettent à l'interprète de savoir s'il est en champ ou hors champ. Ces indicateurs visuels permettent de développer des stratégies d'action sur le son, qui impliquent des stratégies sur le geste : ainsi contourner le faisceau vertical lorsqu'il génère le son permet de figer le son en tant qu'événement sonore, tout en le déplaçant spatialement. De plus, rester au centre implique dans cette configuration un son placé partout autour, enveloppant. Une expressivité basée sur la verticalité se limitera alors à la production de son, dans une configuration de spatialisation figée. Le regard, déjà utilisé par l'interprète pour conserver une prise sur le réel et maîtriser à souhait les déplacements de son corps et de ses membres dans l'espace, trouve ici un point d'ancrage supplémentaire.

Prospectives

On imagine aisément d'autres applications et utilisations de ces outils dans le cadre de la danse. En effet, une multitude de trajectoires différentes peuvent être proposées. De même, l'interaction du geste avec la position du son se fait selon un seul capteur : l'utilisation de deux voire trois capteurs permettra des raffinements dans l'interaction geste-son. Une autre piste consiste à considérer qu'une scène sonore est pré-existant, et que l'interprète, au lieu de déformer la trajectoire d'un son monophonique, agit de manière plus fine sur plusieurs sons (comme dans le cas de l'ouverture spatiale, où le spectre du son est séparé en 8 bandes de fréquences), via des gestes soit sur des

capteurs différents, soit sur le même. Une autre idée, tout aussi intéressante, serait d'utiliser le système de déambulation dans une scène sonore présenté au tout début de l'article en tant que tel, et non plus seulement en tant que documentation de protocoles expérimentaux ou en tant que retour visuel. De cette manière, on remplacerait l'action de l'interprète sur la position du son par une déambulation de l'interprète dans une scène sonore, scène pouvant être générée par l'interprète lui-même ou lui pré-exister.

Vidéo

Une démonstration est jointe à ce document sous format vidéo (*cf.* p. 283), dans laquelle le dispositif utilisé (*cf. fig.* 7.9) correspond à celui où le son est généré selon la dimension Z et la spatialisation contrôlée selon la dimension X. La trajectoire du son se fait en vitesse sur une ellipse (*cf. fig.* 7.8) avec translation corrélée avec le mouvement en X de la danseuse. La représentation graphique, dans le coin en bas à droite de la vidéo, correspond à la modélisation de la salle et de ses 8 haut-parleurs, de la position du son (en bleu), et de la position donnée par les deux dernières interruptions de faisceau (en X et Z) par la danseuse (en rouge).

Remarque : cette partie a été présentée dans l'ouvrage "Espaces Sonores, Actes de Recherche" ; elle est intégrée dans le dernier chapitre, intitulé "*Visualisation de l'Espace Sonore, vers la Notion de Transduction : une Approche Intéractive Temps-Réel*" [Sédès *et al.*, 2003].

7.2.2 Composition de pièces de musique électroacoustique

L'utilisation des effets adaptatifs a été pour moi a la fois une grande source d'inspiration pendant mes trois année d'électroacoustique au Conservatoire National de Région de Marseille, durant la thèse. En effet, j'ai utilisé cette démarche de contrôle adaptatif en filigrane de la pièce de concours de fin de première année. Quelques autres réalisations en cours utilisent pleinement ces traitements. Ils permettent à partir d'un seul matériau sonore de produire une multitude de sons différents, à la fois en terme de texture, d'expressivité, de forme, et se prête tout à fait à ce type de composition. De nombreuses discussions avec des étudiants, des enseignants et compositeurs m'ont aidé à donner une dimension musicale à mon approche, ceci permettant d'éviter de rester dans la théorie. Cependant, je ne pense avoir exploité qu'une infirme partie des potentialité de tels effets.

7.2.3 Contrôle adaptatif du multi-effet Digitech Valve-Fx

Le multi-effet **Digitech Valve-Fx** est dédié aux traitements sonores analogiques et numériques de sons de guitare. Il comprend un étage analogique de pré-amplification avec 3 lampes, et un DSP pour les traitements numériques. Depuis plusieurs années, je l'utilisais à la fois de manière traditionnelle pour jouer du jazz ou du rock, et à la fois de manière expérimentale pour produire du matériau sonore utilisé dans des pièces électroacoustiques. Je contrôlais volontairement des paramètres qui n'était pas prévus à cet effet, via des contrôleurs MIDI tels qu'une pédale d'expression et des interrupteurs pour changer de paramètre.

Tout naturellement, j'ai expérimenté son contrôle adaptatif, en analysant le son sous *Max/MSP* [Cycling'74, 2003] et en envoyant les valeurs de contrôles sous le protocole MIDI. Les résultats ont souvent été à la fois stimulants par les potentialités révélées, et parfois désastreux concernant la qualité sonore des sons produits. En effet, les algorithmes utilisés n'ont pas été prévus pour ce type de contrôle. Cependant, toutes les idées étaient là, il ne restait plus qu'à creuser pour proposer des contrôles plus poussés et des algorithmes améliorés : tout ceci a été investigué durant le projet de recherche, afin de fournir des réponses valides.

7.2.4 “Flûte-Salade n°7”, une pièce pour flûte traversière, effets audionumériques adaptatifs et bande

J'ai composé une pièce pour flûte traversière, effets audionumériques adaptatifs et bande en septembre 2003, *cf.* Piste n°61-CD2 ♫♪. Elle a été créée le 11 octobre 2003 à Lorgues, dans le cadre du concert des “Explorateurs associés”, puis rejouée le 18 octobre 2003 au Café Julien à Marseille. Les effets adaptatifs utilisés sont l'équaliseur adaptatif, dans sa forme d'instrument “Noisonic” (pour la bande), la panoramisation adaptative, le trémolo adaptatif, l'écho adaptatif et le filtrage adaptatif pour le dispositif traitant les sons de flûte. L'ensemble a été réalisé en temps-réel sous *Max/MSP*, et contrôlé par le *rack* d'effets **Digitech Valve-Fx** et le pédalier de contrôle associé **Control One**. Cette création a été l'occasion pour moi de présenter en publique une réalisation musicale achevée, en tant que pièce de musique et non plus uniquement en tant que démonstration des possibilités de ces effets.

7.3 Quelques réflexions sur les effets adaptatifs

Nous proposons maintenant quelques réflexions sur les effets audionumériques adaptatifs, sous forme d'une conclusion partielle.

Remarquons tout d'abord que la démarche de proposer un contrôle adaptatif des effets est une source de création de nouveaux effets. Ainsi, nous avons proposé l'effet de martianisation, le trémolo spectral qui combine un trémolo à un effet de *phasing*, la modulation en annneau spectrale qui réalise la combinaison de la modulation en anneau et du *phasing*. De même, nous avons proposé des traitements jouant sur la dimension spatiale avec la panoramisation adaptative. Les changements de *groove*, de prosodie et d'expressivité via la dilatation/contraction temporelle adaptative sont aussi des nouvelles possibilités offertes par ces effets. Notre tour des effets n'a pas été exhaustif, certes, mais il a révélé de grandes potentialités.

Une question que l'on peut se poser est le choix des exemples sonores. En effet, pour chaque son, certains traitements sont plus appropriés, et réciproquement. Il en résulte qu'il faut du temps pour apprendre à connaître ses effets et le *mapping* de contrôle. Il n'y a pas réellement à notre connaissance de “recette magique” dans ce domaine, dès que le but est la création musicale. On peut comparer la nécessité d'apprentissage de ces effets à celle des instruments traditionnels ou à celle des effets traditionnels.

Pour résumer les résultats présentés dans le chapitre sur les effets adaptatifs, on peut dire que :

- ils sont une généralisation des effets existants ;
- les liens des effets adaptatifs avec les paramètres perceptifs modifiés sont beaucoup plus serrés que ceux des effets traditionnels ;
- ils apportent des améliorations aux algorithmes et aussi au rendu des effets existants :
- ils permettent un contrôle gestuel de plus haut niveau (sur le *mapping*) : équaliseur adaptatif, compresseur spectral adaptatif, spatialisation adaptative ;
- ils permettent une grande cohérence entre le son et l'évolution de l'effet qui lui est appliqué ;
- leur étude implique une réflexion sur la classification des effets et transformations audionumériques en général, et sur les liens effet-paramètre perceptif modifié ;
- ils offrent un cadre pour penser des solutions à des problèmes complexes, tels que la dilatation/contraction temporelle respectant des critères perceptifs (préservation du timbre, des attaques, du vibrato, etc.).

De nombreuses possibilités ont été présentées, mais le champ d'investigations restant est bien plus large encore, et promet d'autres découvertes dans les années à venir.

Troisième partie

Conclusions - annexes

Conclusion

Comme l'écrit Bergson à propos du possible, « Il y a plus, et non pas moins, dans la possibilité de chacun des états successifs que dans leur réalité ». La mise en mouvement de ces germes de transformation inscrits dans les actions sonores, permet de trouver une logique de continuité du mouvement intérieur de l'œuvre qui paraisse évidente et non arbitraire. De telle sorte que les actions sonores semblent s'induire, l'une l'autre.
Christophe Bruno, Philippe Leroux [Bruno and Leroux, 1999]

L'oeuvre d'art qui ne fait pas surgir davantage de problèmes qu'elle (ne) se propose (d'en) résoudre (..) sera toujours insuffisante.
Brian Ferneyhough [Bons, 1997]

Au cours de cette étude sur les effets audionumériques adaptatifs, nous avons mené plusieurs réflexions simultanées, portant sur les effets audionumériques, sur leur mise en œuvre, sur les modèles utilisés, sur leur contrôle à la fois automatique et gestuel. Cette étude aboutit notamment à la représentation faite par le diagramme *fig.* 10. La perception du son et la cognition de l'effet ont été notre fil d'Ariane et ont permi de mieux appréhender les nouvelles transformations sonores proposées. En effet, ces transformations agissent sur la perception du son en modifiant une ou plusieurs de ses dimensions.

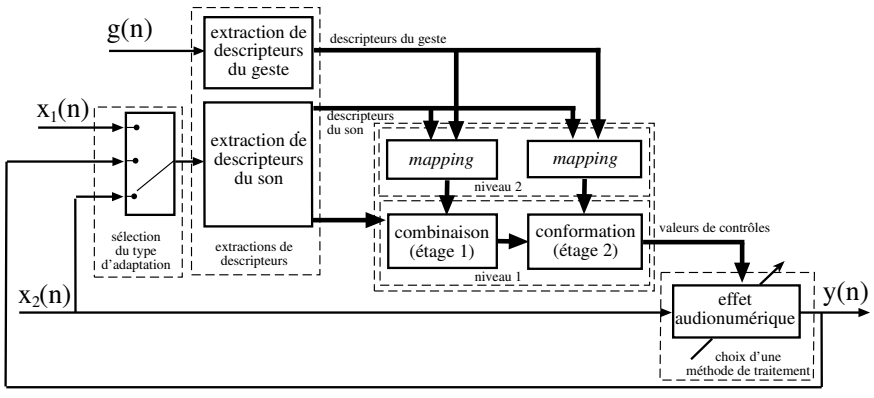


FIG. 10 – *Diagramme complet d'un effet audionumérique adaptatif.*

Les effets adaptatifs permettent un contrôle cohérent avec le son, puisque les courbes de contrôle proviennent de descripteurs du son. Ils permettent notamment des effets ou synthèses croisées, et réintroduisent le contrôle gestuel au niveau du *mapping* entre descripteurs et contrôles de l'effet, permettant un contrôle de plus haut niveau. Pour les exemples sonores obtenus en temps différé, les descripteurs les plus souvent utilisés sont les suiveurs d'enveloppe et de hauteur, le centroïde, l'indice de voisement, l'enveloppe spectrale ; le contrôle est quant à lui très simple, souvent composé d'une seule fonction de transfert non linéaire. Par contre, pour les exemples sonores en temps réel et pour notre utilisation personnelle du système en temps-différé, l'ensemble du *mapping* proposé est mis en œuvre, afin d'offrir plus de précision dans la manipulation, ceci permettant plus de finesse et de subtilité de jeu.

Le contrôle des effets audionumériques adaptatifs se fait à l'aide d'un contrôle à fréquences audio par les descripteurs sonores, et à fréquences sub-audio par le geste. Ceci permet d'apporter de nombreuses améliorations aux effets traditionnels. Tout d'abord, le geste est utilisé à un niveau supérieur au niveau de contrôle des effets usuels : le contrôle n'est plus simple manipulation des contrôles via des potentiomètres et interrupteurs (effets usuels), mais contrôle des configurations de *mapping* par des gestes de modification ou de sélection. Ensuite, il est possible d'utiliser les paramètres perceptifs et psychoacoustiques comme moyen de contrôle, et ainsi d'apporter une grande cohérence entre le son et le contrôle de l'effet par le synchronisme et la concordance d'évolution du contrôle et du son. Une séparation claire a été proposée entre les différentes fonctionnalités du *mapping*, en deux niveaux, l'un pour le contrôle par le son et l'autre pour le contrôle par le geste. De même, le contrôle par le son a été décrit par deux étages, l'un de combinaison des descripteurs, l'autre de conformation du contrôle aux spécificités de l'effet et de son contrôle.

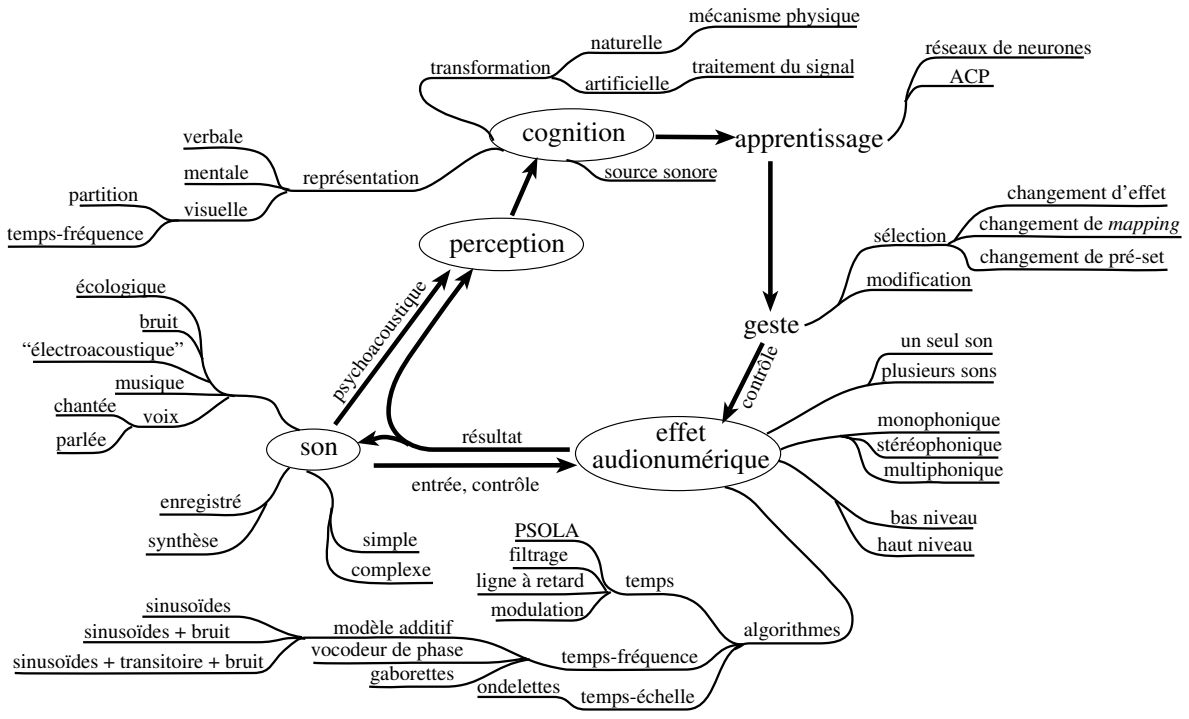


FIG. 11 – *Relations entre son, traitement sonore et perception du son.*

Nous avons développé les programmes correspondants en grande partie sous *Matlab*, mais aussi en temps-réel sous *Max/MSP*. Des outils de manipulation de courbes par des fonctions de transfert ont été réalisés, permettant de les déformer en temps et en amplitude, de les combiner, d'y appliquer des fonctions évolutives (zoom, hystérésis, quantification). Finalement, le choix de *mapping* se fait suite à l'écoute de l'effet appliqué selon différentes configurations de *mapping*, en fonction de celui rendant au mieux l'effet attendu. C'est par le contrôle que la composition commence et que chacun peut exprimer sa sensibilité personnelle. On peut en effet obtenir une multitude de résultats sonores différents pour un même son à traiter et pour un même effet.

Si l'on reprend la figure *fig.* 1 de l'introduction de ce document, on peut en donner une représentation plus complète (diagramme heuristique, *cf.* [Buzan and Buzan, 2003]) par la figure *fig.* 11 d'après tout ce qui a été vu au cours de l'étude.

Concernant la perception des sons traités, nous avons vérifié plusieurs faits perceptifs avérés par des expériences réalisées au cours du développement des effets adaptatifs. Par exemple, de grandes variations non sinusoïdales de hauteur provoquent des changements de perception du timbre (martianisation, *cf.* sec. 5.7.4), qui ne sont pas forcément perçues comme des modulations en anneau puisque le contrôle n'est pas sinusoïdal. Ensuite de grandes variations de panoramisation affectent la sensation de flux auditifs, quitte à les ségréguer, lorsque les déplacements selon la dimension perceptive d'espace sont corrélés à des variations selon une autre dimension perceptive (hauteur, intensité, timbre), *cf.* la panoramisation adaptative (sec. 5.5.2), la spatialisation adaptative (sec. 5.5.4). De même, de grandes variations de gain modifient le rythme perçu du son (sec. 5.2.1).

Des variations de la dilatation/contraction temporelle, même minimes, permettent de modifier l'expressivité (*cf.* sec. 5.3), tout en permettant de conserver la durée globale du son, voire d'imposer des temps de synchronisation : c'est alors le rythme qui est localement modifié. Des variations de hauteur de la voix parlée combinées à des variations de contraction/dilatation temporelle permettent de modifier la prosodie, et ainsi le sens donné à une phrase. Les variations de l'enveloppe spectrale, du spectre d'amplitude ou de l'ensemble du spectre offrent eux aussi des outils de composition du son très utiles en musique électroacoustique.

Pour conclure, nous dirons qu'à partir d'une idée nouvelle (le contrôle adaptatif des effets audionumériques), nous avons proposé une formalisation et une généralisation des effets adaptatifs. Ce cadre de référence vaut pour tous les effets. De plus, il inclut le geste tout en reportant une partie du contrôle sur le son : nous avons proposé une structure de *mapping* entre le geste et le son d'une part et le contrôle de l'effet d'autre part. Nous avons indiqué des points d'entrée pour le contrôle et des descripteurs permettant le contrôle adaptatif. Les résultats sonores montrent comment l'expressivité et la musicalité peuvent être modifiées avec sens : pour travailler musicalement le son, il faut tenir compte de la perception. Nous dirons pour terminer qu'on ne peut utiliser un effet sans tenir compte du contrôle, qu'il soit automatique, adaptatif ou gestuel : une vision globale du contrôle telle que celle nous proposons s'imposait. La transition des effets audionumériques[1] vers le contrôle gestuel s'opère[2] !

[1] Conférences **DAFx** : *Digital Audio Effects*
[2] Conférences **ConGAS** : *Control of Gestural Audio Systems*

Annexe A

Descriptif des exemples sonores

Le son habite partout. Mais les sons, je veux dire les mélodies
qui parlent la langue supérieure du royaume de l'esprit,
ne reposent que dans le sein de l'Homme.

Hauffmann, cité par Pierre Schaeffer,
"Solfège des Objets Sonores" [Schaeffer, 1966]

Sommaire

Nous présentons ici les différents exemples sonores joints au mémoire. Les premiers sont les sons de "référence", avant traitement. Certains de ces sons de référence ont été utilisés uniquement pour les illustrations des descripteurs (*cf.* chap. 4). Les deux CD qui accompagnent le document contiennent les sons de référence dans les premières plages, puis les sons traités illustrant des effets adaptatifs.

A.1 Sons de référence

Piste n°1 ♫♪ (p. 188) : boucle de son de guitare *funk* [Zero G, 2000b].
Piste n°2 ♫♪ (p. 13) : son de cloche, extrait du "Poème électronique" d'Edgar Varèse [Varèse, 1958].

Piste n°3 ♫♪ (p. 12, 122, 170, 178 et 189) : introduction de “Like Someone in Love” de Sylvain Beuf (improvisation au saxophone) [Beuf, 2002].
Piste n°4 ♫♪ (p. 189) : son synthétique de trompette réalisés par John Chowning en synthèse par modulation de fréquence dans les années 1970.
Piste n°5 ♫♪ (p. 270) : extrait de “... entendue sous le sable” [CNRS, 2002], du CD *Chant des Dunes* accompagnant le *Journal du CNRS* [CNRS éditions, 2002].
Piste n°6 ♫♪ (p. 197 et 285) : Earth, Wind and Fire en concert [Earth Wind & Fire, 2001].
Piste n°7 ♫♪ (p. 184 et 282) : extrait de “Mallorca” de Chick Corea et Steve Kujala [Corea and Kujala, 1985].
Piste n°8 ♫♪ (p. 170, 169 et 191) : extrait de “Mallorca” de Chick Corea et Steve Kujala [Corea and Kujala, 1985].
Piste n°9 ♫♪ (p. 193) : glissements de doigts sur les cordes d’une basse électrique [Zero G, 2000a].
Piste n°10 ♫♪ (p. 187 et 206) : “Lalula”, phrase dans un langage imaginaire [Dutilleux, 1992].
Piste n°11 ♫♪ (p. 170) : extrait court de “Lalula” [Dutilleux, 1992].
Piste n°12 ♫♪ (p. 118, 184 et 210) : voix parlée (‘love”).
Piste n°13 ♫♪ (p. 202) : “bavazakamasarpataparda”. Extrait de la chanson “buvons un coup, ma serpette est perdue”, selon les modalités d’un jeu où toutes les voyelles sont remplacées par une seule (un [a] dans l’exemple).
Piste n°14 ♫♪ (p. 190) : extrait de musique électro, Aphex Twin [Aphex Twin, 2001].
Piste n°15 ♫♪ (p. 170, 206 et 207) : extrait du “*Solfège des Objets Sonores*” [Schaeffer and Reibel, 1998] CD1 Piste 1.
Piste n°16 ♫♪ (p. 10, 15, 117, 167, 193 et 235) : extrait de la Piste n°15 ♫♪.
Piste n°17 ♫♪ (p. 282 et 91) : extrait du “*Solfège des Objets Sonores*” [Schaeffer and Reibel, 1998], CD 1 Piste 3.
Piste n°18 ♫♪ (p. 122, 128 et 146) : extrait de la Piste n°17 ♫♪.
Piste n°19 ♫♪ (p. 203) : introduction de “Tom’s Diner” de Suzanne Vega [Vega, 1993].
Piste n°20 ♫♪ (p. 120, 133 et 91) : introduction de “Spain” de Chick Corea [Camilo and Tomatito, 2000].

A.2 Exemples accompagnant les effets adaptatifs sur la dynamique

A.2.1 Changement de niveau adaptatif

Piste n°21-CD1 ♫♪ (p. 162) : changement de niveau adaptatif appliqué à la Piste n°3 ♫♪.
Contrôle : le niveau est fonction du RMS et du centroïde.
Remarque : les notes sont plus courtes, plus “sèches”.
Piste n°22-CD1 ♫♪ (p. 162) : changement de niveau adaptatif appliqué à la Piste n°5 ♫♪.
Contrôle : on utilise le RMS (qui varie très peu dans le son original) normalisé avec $\mathcal{N}_1$ et modifié par la fonction $\mathcal{H}_{sin}$. Le gain varie alors dans $[0;1]$ en renforçant la proximité du 0 ou du 1, ce qui creuse davantage la courbe de contrôle.
Remarque : pour ce son de type “trame temporelle” (ou “nappe”), l’utilisation du changement de niveau adaptatif permet de rajouter de grandes variations de niveau.
Piste n°23-CD1 ♫♪ (p. 162) : changement de niveau adaptatif appliqué à la Piste n°5 ♫♪.
Contrôle : on utilise maintenant la courbe du RMS inversée en amplitude, avec le même *mapping*.
Remarque : les observations sont identiques, et cette fois-ci, c’est un peu le “son complémentaire” qui a été réalisé, puisque d’amplitude maximale lorsque l’autre est d’amplitude minimale, et réciproquement.
Piste n°24-CD1 ♫♪ (p. 162) : changement de niveau adaptatif appliqué à la Piste n°10 ♫♪.
Contrôle : on utilise l’indice de voisement auquel on applique la fonction de conformation $1 - \mathcal{H}_{sin}$.
Remarque : on obtient une sorte de sélecteur de consonnes.

A.2.2 Trémolo adaptatif

Piste n°25-CD1 ♫♪ (p. 163) : trémolo adaptatif appliqué à la Piste n°2 ♫♪.
Contrôle : la courbe de contrôle a une échelle linéaire.
Remarque : Le trémolo est assez “mou”.
Piste n°26-CD1 ♫♪ (p. 163) : trémolo adaptatif appliqué à la Piste n°2 ♫♪.
Contrôle : la courbe de contrôle a une échelle logarithmique.
Remarque : Le trémolo est plus sec que précédemment.

A.2.3 Pseudo-trémolo adaptatif

Piste n°27-CD1 ♫♪ (p. 163) : pseudo-trémolo adaptatif appliqué à la Piste n°2 ♫♪.
Contrôle : la forme est donnée par une table d'amplitude donnée par l'utilisateur.
Remarque : L'effet de double rebond ne s'entend que pour les modulations lentes, à la fin de l'exemple.
Piste n°28-CD1 ♫♪ (p. 163) : pseudo-trémolo adaptatif appliqué à la Piste n°2 ♫♪.
Contrôle : table d'amplitude contenant une forme triangulaire.
Remarque : on obtient un générateur d'enveloppes triangulaires de longueurs variables.
Piste n°29-CD1 ♫♪ (p. 163) : pseudo-trémolo adaptatif appliqué à la Piste n°2 ♫♪.
Contrôle : table d'amplitude contenant une forme de type pallier.
Remarque : on obtient un générateur d'enveloppes “pallier” de longueurs variables.

A.2.4 Trémolo spectral adaptatif

Piste n°30-CD1 ♫♪ (p. 167) : trémolo spectral adaptatif appliqué à la Piste n°2 ♫♪.
Contrôle : la fréquence du trémolo spectral ($f_{trem} = 2 + 4[1 - \mathcal{N}_1(20\log_{20}(\mathcal{E}))] \in [2;6]$) est fonction de l'enveloppe spectrale , avec une fenêtre de 1024 points et un pas de 512 échantillons.
Remarque : un effet de *phasing* apparaît, en plus du trémolo perceptible principalement dans les basses fréquences.
Piste n°31-CD1 ♫♪ (p. 167 et 168) : trémolo spectral adaptatif de profondeur 100 dB, appliqué à la Piste n°2 ♫♪.
Contrôle : $f_{trem} = 2[1 - \mathcal{N}_1(20\log_{20}(\mathcal{E}))] + 4 \in [4;6]$ avec comme profondeur $d_{trem} = 100\ dB$.
Remarque : L'effet de *phasing* coexiste avec le trémolo dans les graves.
Piste n°32-CD1 ♫♪ (p. 167) : trémolo spectral adaptatif de profondeur 50 dB, appliqué à la Piste n°2 ♫♪.
Contrôle : identique avec comme profondeur $d_{trem} = 50\ dB$.
Remarque : L'effet de *phasing* coexiste avec le trémolo dans les graves.
Piste n°33-CD1 ♫♪ (p. 167) : trémolo spectral adaptatif de profondeur 10 dB, appliqué à la Piste n°2 ♫♪.
Contrôle : identique avec comme profondeur $d_{trem} = 10\ dB$.
Remarque : L'effet de *phasing* est léger mais prépondérant, l'effet de trémolo ayant quasiment disparu (la profondeur est trop faible).

A.3 Exemples accompagnant les effets adaptatifs sur la durée

A.3.1 Dilatation/contraction temporelle non linéaire

Piste n°34-CD1 ♫♪ (p. 170) : dilatation/contraction temporelle non linéaire appliquée Piste n°8 ♫♪.
Contrôle : le ratio de dilatation/contraction est fonction de F_0.
Remarque : cela correspond à une contraction de la note la plus aiguë et une dilatation des autres notes.

Piste n°35-CD1 ♫♪ (p. 170) : dilatation/contraction temporelle non linéaire appliquée Piste n°8 ♫♪.
Contrôle : le ratio de dilatation/contraction est fonction de $-F_0$ (la transformation inverse).
Remarque : cela correspond à une dilatation de la note la plus aiguë et une contraction des autres notes.
Piste n°36-CD1 ♫♪ (p. 170) : dilatation/contraction temporelle non linéaire appliquée Piste n°3 ♫♪.
Contrôle : le ratio de dilatation/contraction est fonction de l'énergie.
Remarque : cette improvisation, originellement joué "dans le temps", devient hors temps avec ce traitement.
Piste n°37-CD1 ♫♪ (p. 170) : dilatation/contraction temporelle non linéaire appliquée Piste n°11 ♫♪.
Contrôle : le ratio de dilatation/contraction est contrôlé par l'indice de voisement.
Remarque : l'intelligibilité du contenu est altérée.
Piste n°38-CD1 ♫♪ (p. 170) : dilatation/contraction temporelle non linéaire appliquée Piste n°11 ♫♪.
Contrôle : le ratio de dilatation/contraction est contrôlé par l'indice de voisement avec une conformation non linéaire de la courbe de contrôle.
Remarque : l'intelligibilité du contenu est altérée.
Piste n°39-CD1 ♫♪ (p. 170) : dilatation/contraction temporelle non linéaire appliquée Piste n°16 ♫♪.
Contrôle : le ratio de dilatation/contraction est fonction de l'indice de voisement.
Remarque : l'expressivité est modifiée. Le locuteur est pressé, sans pause de respiration entre les mots.
Piste n°40-CD1 ♫♪ (p. 171) : dilatation/contraction temporelle non linéaire appliquée Piste n°16 ♫♪.
Contrôle : le ratio de dilatation/contraction est fonction du RMS et de l'indice de voisement.
Remarque : l'expressivité est modifiée. Le locuteur ralenti par moments, sans respiration ; il semble réfléchir à ce qu'il dit.
Piste n°41-CD1 ♫♪ (p. 171) : dilatation/contraction temporelle non linéaire appliquée Piste n°16 ♫♪.
Contrôle : le ratio de dilatation/contraction est fonction du RMS et de l'indice de voisement.
Remarque : l'expressivité est modifiée. Le locuteur ralenti aux pauses respiratoires et au milieu de certains mots ; il semble réfléchir entre chaque mot.
Piste n°42-CD1 ♫♪ (p. 174) : dilatation/contraction temporelle non linéaire appliquée Piste n°16 ♫♪.
Contrôle : le ratio de dilatation/contraction est positif ou négatif.
Remarque : illustration de l'effet "attracteur" lorsque le ratio de dilatation/contraction temporelle peut être négatif et que l'index de synthèse reste confiné dans une zone réduite.

A.3.2 Dilatation/contraction non linéaire avec préservation de la durée globale

Piste n°43-CD1 ♫♪ (p. 178) : dilatation/contraction non linéaire avec préservation de la durée globale, Piste n°16 ♫♪.
Contrôle : le ratio de dilatation/contraction est fonction de l'indice de voisement, et la synchronisation se fait par multiplication du ratio.
Remarque : changement d'expressivité de la voix de Pierre Schaeffer, utile pour réinterpréter une voix créer un chorus naturel.
Piste n°44-CD1 ♫♪ (p. 178) : dilatation/contraction non linéaire avec préservation de la durée globale, Piste n°18 ♫♪.
Contrôle : le ratio de dilatation/contraction est fonction de l'indice de voisement et du RMS, et la synchronisation se fait en élevant le ratio à une puissance.
Remarque : changement d'expressivité de la voix de Pierre Schaeffer, utile pour réinterpréter une voix créer un chorus naturel.
Piste n°45-CD1 ♫♪ (p. 178) : dilatation/contraction non linéaire avec préservation de la durée globale, Piste n°3 ♫♪.
Contrôle : le ratio de dilatation/contraction est fonction du CGS, et la synchronisation se fait selon une loi puissance.
Remarque : le thème joué au saxophone par Sylvain Beuf est rendu hors temps ; des arythmies

apparaissent.
Piste n°46-CD1 ♫♪ (p. 178) : piste stéréophonique avec le signal original à gauche (Piste n°3 ♫♪) et le signal Piste n°45-CD1 ♫♪ à droite.
Piste n°47-CD1 ♫♪ (p. 178) : dilatation/contraction non linéaire avec préservation de la durée globale de la Piste n°3 ♫♪.
Contrôle : le ratio de dilatation/contraction est fonction du CGS avec un *mapping* différent.
Remarque : le thème joué au saxophone par Sylvain Beuf est rendu hors temps par Des arythmies apparaissent.
Piste n°48-CD1 ♫♪ (p. 178) : piste stéréophonique avec le signal original à gauche (Piste n°3 ♫♪) et le signal Piste n°47-CD1 ♫♪ à droite.

A.3.3 Dilatation/contraction non linéaire avec préservation de la durée globale et synchronisation

Piste n°49-CD1 ♫♪ (p. 181) : dilatation/contraction non linéaire avec synchronisation toutes les deux secondes, appliquée à la Piste n°18 ♫♪.
Contrôle : le ratio de dilatation/contraction est fonction du CGS et du RMS.
Remarque : le changement d'expressivité est flagrant. Les deux voix sont néanmoins synchrones par moments, notamment au début et à la fin.
Piste n°50-CD1 ♫♪ (p. 181) : canal de gauche, son original Piste n°18 ♫♪ ; canal de droite, Piste n°49-CD1 ♫♪. .
Remarque : on obtient un effet (ici exagéré) de chorus plus naturel que par les méthodes de modulation de ligne à retard usuelles, du fait que les décalages temporels sont audibles (un peu trop dans ces exemples), mais surtout n'ont pas l'aspect systématique habituel.
Piste n°51-CD1 ♫♪ (p. 181) : dilatation/contraction non linéaire avec synchronisation toutes les deux secondes, appliquée à la Piste n°18 ♫♪.
Contrôle : le ratio de dilatation/contraction est fonction de l'indice de voisement et du CGS, avec des choix de conformations différents, mais aussi un type de synchronisation entre chaque couple de points de synchronisation différent.
Remarque : le changement d'expressivité est marqué. Les deux voix sont synchrones aux instants imposés par l'utilisateur.
Piste n°52-CD1 ♫♪ (p. 181) : canal de gauche : son original Piste n°3 ♫♪ ; canal de droite, Piste n°51-CD1 ♫♪.
Remarque : les deux voix ont un déroulement temporel différent, même si elles se rejoignent de temps en temps (et à la fin).
Piste n°53-CD1 ♫♪ (p. 181) : dilatation/contraction non linéaire avec synchronisation toutes les deux secondes, appliquée à la Piste n°3 ♫♪.
Contrôle : le ratio de dilatation/contraction est fonction du CGS et de la fréquence fondamentale. La synchronisation se fait selon une loi puissance.
Remarque : l'expressivité diffère du son original.
Piste n°54-CD1 ♫♪ (p. 181) : canal de gauche : son original Piste n°3 ♫♪ ; canal de droite : Piste n°53-CD1 ♫♪.
Remarque : l'effet de chorus est réel, il semble que deux musiciens jouent en simultanés la même ligne mélodique, avec des expressivités différentes.
Piste n°55-CD1 ♫♪ (p. 181) : dilatation/contraction non linéaire avec synchronisation toutes les deux secondes appliquée à la Piste n°3 ♫♪.
Contrôle : le ratio de dilatation/contraction est fonction du RMS, avec et des choix de fonctions de conformation différents de l'exemple précédent.
Remarque : l'expressivité est différente.
Piste n°56-CD1 ♫♪ (p. 181) : canal de gauche : son original Piste n°3 ♫♪ ; canal de droite : Piste n°55-CD1 ♫♪.

Remarque : l'effet de chorus est réel, il semble que deux musiciens jouent en simultanés la même ligne mélodique, avec des expressivités différentes.

A.4 Exemples accompagnant les effets adaptatifs sur la hauteur

A.4.1 Transposition adaptative

Piste n°57-CD1 ♫♪ (p. 183) : transposition adaptative de la voix parlée Piste n°10 ♫♪.
Contrôle : le facteur de transposition est fonction de la fréquence fondamentale.
Remarque : l'intonation est modifiée.
Piste n°58-CD1 ♫♪ (p. 183) : transposition adaptative de la voix parlée Piste n°10 ♫♪.
Contrôle : le facteur de transposition est fonction de la fréquence fondamentale et du RMS.
Remarque : l'intonation est modifiée.
Piste n°59-CD1 ♫♪ (p. 183) : transposition adaptative d'une phrase musicale Piste n°3 ♫♪.
Contrôle : le facteur de transposition est fonction de la fréquence fondamentale, de manière à l'inverser.
Remarque : l'effet est assez indigeste dès lors qu'on ne prend pas en compte des données de segmentation afin de transposer de manière constante sur chaque note.

A.4.2 Harmoniseur adaptatif

Piste n°60-CD1 ♫♪ (p. 184) : harmonisation adaptative de la Piste n°12 ♫♪.
Contrôle : le son passe de manière continue de l'accord mineur à l'accord majeur, en faisant glisser la tierce. Le facteur de transposition de la tierce est fonction du RMS.
Remarque : cet effet permet d'appliquer des glissement d'accord, ce qui induit des sensations de flottement (l'accord n'est ni tout le temps majeur, ni tout le temps mineur).
Piste n°61-CD1 ♫♪ (p. 184) : harmonisation adaptative de la Piste n°12 ♫♪.
Contrôle : on fait passer la première transposition de l'octave inférieure à la tierce majeure supérieure, et la seconde transposition de l'octave supérieure à la quinte juste supérieure.
Remarque : ce type de contrôle est moins lisible pour l'auditeur, mais intéressant à l'écoute.
Piste n°62-CD1 ♫♪ (p. 184) : harmonisation non adaptative de la Piste n°7 ♫♪.
Contrôle : le son passe de manière continue de l'accord de septième mineure à l'accord de septième majeure.
Remarque : cet effet permet d'appliquer des glissement d'accord, ce qui induit des sensations de flottement (l'accord n'est ni tout le temps septième majeure, ni tout le temps septième mineure).
Piste n°63-CD1 ♫♪ (p. 184) : harmonisation et transposition adaptatives de la Piste n°7 ♫♪.
Contrôle : on applique une harmonisation passant de l'accord mineur sixième à l'accord de majeur septième en fonction du RMS, en ajoutant une transposition adaptative de tout le son allant jusqu'à moins un ton, conduite par le taux de basses énergies.
Remarque : ce type de contrôle rend le son bien moins "stable" harmoniquement que précédemment.
Piste n°64-CD1 ♫♪ (p. 184) : harmonisation et transposition adaptatives de la Piste n°7 ♫♪.
Contrôle : applique une harmonisation en accord de majeur septième, en ajoutant une transposition adaptative allant jusqu'à moins un ton, conduite par le RMS.
Remarque : ce type de contrôle rend le son bien moins "stable" harmoniquement que précédemment.

A.5 Exemples accompagnant les effets adaptatifs sur la spatialisation

A.5.1 Echo granulaire adaptatif

Piste n°65-CD1 ♫♪ (p. 187) : écho granulaire adaptatif appliqué à la Piste n°1 ♫♪.
Contrôle : seul le gain de réinjection est contrôlé par une courbe dérivée du RMS ($\tau = 0.3\ s$),

Remarque : on applique l'écho de manière prépondérante sur les attaques. C'est aussi un exemple de *morphing* de timbre au cours du temps.

Piste n°66-CD1 ♫♪ (p. 187) : écho granulaire adaptatif appliqué à la Piste n°1 ♫♪.
Contrôle : seul le gain de réinjection est contrôlé par une courbe dérivée du RMS ($\tau = 0.3\ s$),
Remarque : on applique l'écho de manière prépondérante sur les parties harmoniques. C'est aussi un exemple de *morphing* de timbre au cours du temps.

Piste n°67-CD1 ♫♪ (p. 187) : écho granulaire adaptatif appliqué à la Piste n°10 ♫♪.
Contrôle : le gain est de 0.4 et le temps de délai varie entre 0.2 et 1 seconde, en utilisant comme courbe de contrôle la fréquence fondamentale.
Remarque : on obtient un son de synthèse granulaire dont la hauteur monte, étiré dans le temps au fur et à mesure de ses répétitions. Ceci est plus proche de la synthèse granulaire que de l'écho traditionnel.

Piste n°68-CD1 ♫♪ (p. 187) : écho granulaire adaptatif appliqué à la Piste n°10 ♫♪.
Contrôle : avec un temps de délai à 0.2 seconde et un gain de réinjection variable, en fonction de la balance de voisement :
Remarque : on obtient une voix dont seules les consonnes sont répétées. C'est aussi un exemple de *morphing* de timbre au cours du temps.

A.5.2 Panoramisation adaptative

Piste n°69-CD1 ♫♪ (p. 189) : panoramisation adaptative de la Piste n°4 ♫♪.
Contrôle : l'angle de panoramisation du son est fonction du CGS.
Remarque : ceci illustre la perception de plusieurs flux sonores à partir d'un son monophonique lorsque le son se déplace à des positions préférentielles. La première trompette synthétique semble déjà "à cheval" sur les deux canaux, et lorsque la seconde trompette entre en jeu, elle semble elle aussi se déplacer, par symétrie avec la première.

Piste n°70-CD1 ♫♪ (p. 189) : panoramisation adaptative de la Piste n°4 ♫♪.
Contrôle : l'angle de panoramisation du son est fonction de l'indice de voisement et du RMS.
Remarque : ceci illustre les problèmes dus à des variations trop rapides de l'angle $\theta(t)$. On entend aux débuts et fins de chaque note un effet de type *noise gate*, où le bruit de fond est supprimé brusquement. Ceci semble dû au mouvement rapide du son à cet instant où il passe d'un son harmonique de niveau fort à un son bruité de niveau faible.

Piste n°71-CD1 ♫♪ (p. 189) : panoramisation adaptative de la Piste n°3 ♫♪.
Contrôle : l'angle de panoramisation du son est fonction du contenu en hautes fréquences (HFC).
Remarque : ceci illustre la façon dont cet effet peut produire des déplacements assez lents pour une source seule.

A.5.3 Panoramisation spectrale adaptative

Piste n°72-CD1 ♫♪ (p. 190) : panoramisation spectrale obtenue à partir de la piste monophonique Piste n°14 ♫♪.
Contrôle : l'angle de panoramisation spectrale de chaque panier de fréquence est fonction d'une interpolation de formes d'ondes.
Remarque : Cet effet permet d'ajouter plusieurs plans sonores à un seul son. En effet, les composantes spectrales sont traités de manières différentes, et l'on entend le son comme provenant de plusieurs sources sonores se déplaçant selon des trajets différents. Le son n'est pas placé quelque part entre la gauche et la droite, mais c'est une modification plus subtile et plus complexe à saisir pour l'auditeur qui s'opère.

A.6 Exemples accompagnant les effets adaptatifs sur le timbre

A.6.1 Décalage adaptatif de l'enveloppe spectrale

Piste n°73-CD1 ♫♪ (p. 193) : décalage adaptatif de $\pm 300\ Hz$ de l'enveloppe spectrale appliqué à la Piste n°9 ♫♪.
Contrôle : le décalage de l'enveloppe spectrale est fonction du RMS.
Remarque : certaines parties du son traité ont plus de présence qu'auparavant, voire "sonnent" très différemment, de fait que des fréquences très faibles sont grandement amplifiées. On apporte une plus grande variabilité au timbre du son obtenu par le traitement.
Piste n°74-CD1 ♫♪ (p. 193) : décalage adaptatif de $\pm 300\ Hz$ de l'enveloppe spectrale appliqué à la Piste n°9 ♫♪.
Contrôle : le décalage de l'enveloppe spectrale est fonction d'une combinaison entre RMS et centroïde.
Remarque : on apporte une fois encore une plus grande variabilité au timbre du son obtenu par le traitement.

A.6.2 Dilatation/contraction adaptative de l'enveloppe spectrale

Piste n°75-CD1 ♫♪ (p. 193) : dilatation/contraction adaptative de l'enveloppe spectrale de la Piste n°9 ♫♪.
Contrôle : le facteur de dilatation/contraction de l'enveloppe spectrale est fonction du RMS.
Remarque : cet effet augmente la variabilité spectrale.
Piste n°76-CD1 ♫♪ (p. 193) : dilatation/contraction adaptative de l'enveloppe spectrale de la Piste n°9 ♫♪.
Contrôle : le facteur de dilatation/contraction de l'enveloppe spectrale est fonction du taux de basses énergies.
Remarque : cet effet augmente la variabilité spectrale.
Piste n°77-CD1 ♫♪ (p. 193) : dilatation/contraction adaptative de l'enveloppe spectrale de la Piste n°16 ♫♪.
Contrôle : le facteur de dilatation/contraction de l'enveloppe spectrale est fonction du CGS.
Remarque : l'effet fonctionne bien pour des sons riches en harmoniques ou partiels.
Piste n°78-CD1 ♫♪ (p. 193) : dilatation/contraction adaptative de l'enveloppe spectrale de la Piste n°8 ♫♪.
Contrôle : le facteur de dilatation/contraction de l'enveloppe spectrale est fonction du RMS, du voisement et du CGS.
Remarque : l'effet ressemble par certains moment à du *phasing*.

A.6.3 Conformation adaptative de l'enveloppe spectrale (NTR)

Piste n°79-CD1 ♫♪ (p. 194) : conformation adaptative de l'enveloppe spectrale de la Piste n°9 ♫♪.
Contrôle : la courbe de conformation adaptative de l'enveloppe spectrale est fonction de l'intégrale de l'enveloppe spectrale.
Remarque : les composantes amplifiées par la conformation de l'enveloppe peuvent être bruités, et inversement les composantes diminuées peuvent être harmoniques, si bien que la sonie du signal change considérablement.
Piste n°80-CD1 ♫♪ (p. 194) : conformation adaptative de l'enveloppe spectrale de la Piste n°16 ♫♪.
Contrôle : la courbe de conformation adaptative de l'enveloppe spectrale est fonction de l'intégrale de l'enveloppe spectrale.
Remarque : même remarque sur le changement de l'intensité perçue (sonie). Les formants de la voix parlée sont moins reconnaissables que pour un décalage ou une dilatation/contraction de l'enveloppe.

Piste n°81-CD1 ♫♪ (p. 194) : conformation adaptative de l'enveloppe spectrale de la Piste n°8 ♫♪.
Contrôle : la courbe de conformation adaptative de l'enveloppe spectrale est fonction de l'intégrale de l'enveloppe spectrale.
Remarque : la composante bruitée prend facilement le pas sur la composante harmonique.

A.6.4 Equaliseur adaptatif

Piste n°82-CD1 ♫♪ (p. 197) : équalisation adaptative stéréophonique de la Piste n°6 ♫♪.
Contrôle : le méta-descripteur de contrôle de l'équaliseur adaptatif utilisé est la forme d'onde.
Remarque : on ressent l'effet de filtrage évolutif comme une "modulation de spectre".

A.6.5 Instrument "Noisonic"

Piste n°83-CD1 ♫♪ (p. 198) : exemples de sons obtenus avec l'instrument Noisonic, ou synthèse croisée adaptatif entre un bruit blanc et un son riche, Piste n°6 ♫♪, utilisé comme contrôle de l'équaliseur.
Contrôle : le méta-descripteur de contrôle de l'équaliseur adaptatif utilisé est la forme d'onde. Le taux de rafraîchissement varie entre 20 ms (milieu du morceau) et 1 seconde, et l'interpolation d'une forme d'onde à la suivante est incomplète.
Remarque : le fait que l'interpolation de la forme de contrôle soit incomplète implique un contrôle discontinu, ce qui permet des changements brusques de sonorité (bruits filtrés impulsifs).

A.6.6 Filtre en peigne

Piste n°84-CD1 ♫♪ (p. 199) : filtrage en peigne (cascade puis parallèle) adaptatif de la Piste n°6 ♫♪.
Contrôle : le délai et les gains du filtre en peigne adaptatif est contrôlé par le RMS et le CGS.
Remarque : ceci illustre l'effet "moustique".

A.6.7 Changement de voyelle

Piste n°85-CD1 ♫♪ (p. 202) : changement de voyelle adaptatif.
Contrôle : les [a] de la Piste n°13 ♫♪ sont remplacés par la suite de voyelles {[o],[u],[é],[a],[o],[u],[a]}. Le filtrage est évolutif pour essayer de respecter la co-articulation.
Remarque : il reste des artefacts sur les consonnes et certaines co-articulations.
Piste n°86-CD1 ♫♪ (p. 202) : changement de voyelle adaptatif.
Contrôle : les [a] de la Piste n°13 ♫♪ sont remplacés par la suite de voyelles {[i],[o],[a],[e],[u],[é],[i]}. Le filtrage est évolutif pour essayer de respecter la co-articulation.
Remarque : il reste des artefacts sur les consonnes et certaines co-articulations.

A.6.8 Chuchotement

Piste n°87-CD1 ♫♪ (p. 203) : chuchotement adaptatif appliqué à la Piste n°19 ♫♪.
Contrôle : le chuchotement adaptatif est contrôlé par la fréquence fondamentale avec une taille de grain varie de 64 à 4096 échantillons.
Remarque : la voix semble chuchotée au début de chaque phrase, puis parlée en fin de phrase.

A.6.9 Vibrato adaptatif

Piste n°88-CD1 ♫♪ (p. 204) : vibrato adaptatif appliqué à la Piste n°8 ♫♪.
Contrôle : la profondeur du vibrato adaptatif dépend de la fréquence fondamentale.
Remarque : seule la note la plus aiguë a un vibrato perceptible.

A.6.10 Dual detune

Piste n°89-CD1 ♫♪ (p. 205) : double transposition, vers le haut d'un facteur $\rho(t)$ et vers le bas d'un facteur $\frac{1}{\rho(t)}$, appliqué à la Piste n°8 ♫♪.
Contrôle : l'amplitude de la transposition est inférieure au demi-ton, avec un ambitus de transposition de ± 1 ton, et contrôlé par le RMS.
Remarque : cet effet de *detune* permet d'ajouter de l'inharmonicité à un son.

A.6.11 Conformation spectrale adaptative

Piste n°90-CD1 ♫♪ (p. 205) : conformation spectrale adaptative de la Piste n°8 ♫♪.
Contrôle : la courbe de conformation spectrale adaptative est fonction de la somme cumulative de l'enveloppe spectrale, sans préservation de l'enveloppe spectrale.
Remarque : le son peut devenir méconnaissable, du fait de la non-préservation de l'enveloppes spectrale.
Piste n°91-CD1 ♫♪ (p. 205) : conformation spectrale adaptative de la Piste n°8 ♫♪.
Contrôle : la courbe de conformation spectrale adaptative est fonction de la somme cumulative de l'enveloppe spectrale, avec conservation de l'enveloppe.
Remarque : le son résultant est différent, et pas vraiment plus reconnaissable sur le plan harmonique, mais par contre plus proche sur le plan formantique.

A.7 Exemples accompagnant les effets adaptatifs portants sur plusieurs paramètres perceptifs

CD2

A.7.1 Robotisation adaptative

Piste n°21-CD2 ♫♪ (p. 206) : modification de l'intonation et du timbre par robotisation adaptative de la Piste n°15 ♫♪.
Contrôle : le pas de la robotisation est contrôlé par l'indice de voisement, entre 80 et 100 Hz.
Remarque : ceci permet d'effectuer des changements de timbre et aussi d'intonation, donc de prosodie.
Piste n°22-CD2 ♫♪ (p. 206) : modification de l'intonation et du timbre de la Piste n°15 ♫♪ par robotisation adaptative et harmonisation (accord Majeur).
Contrôle : le pas de la robotisation est contrôlé par le RMS.
Remarque : ceci permet d'effectuer des changements de timbre et aussi d'intonation, donc de prosodie, tout en ajoutant un aspect musical par l'harmonisation.
Piste n°23-CD2 ♫♪ (p. 206) : conservation de l'intonation et modification du timbre par robotisation adaptative de la Piste n°10 ♫♪.
Contrôle : le pas de la robotisation adaptative est contrôlé par F_0.
Remarque : ceci permet de ne modifier que le timbre de la voix, sans affecter la hauteur.
Piste n°24-CD2 ♫♪ (p. 206) : modification de l'intonation et du timbre de la Piste n°10 ♫♪ par robotisation adaptative.
Contrôle : le pas de la robotisation adaptative est contrôlé par l'indice de voisement.
Remarque : ceci permet de modifier à la fois le timbre de la voix et son intonation.

A.7.2 Ré-échantillonnage adaptatif

Piste n°25-CD2 ♫♪ (p. 207) : ré-échantillonnage adaptatif appliqué à la Piste n°16 ♫♪.
Contrôle : le facteur de ré-échantillonnage est fonction du RMS avec un *mapping* non linéaire, variant entre 0.5 et 1.4.
Remarque : cet effet permet à la fois une modification de la hauteur et du timbre, et peut s'entendre comme une "cartoonification" du son.

Piste n°26-CD2 ♫♪ (p. 207) : ré-échantillonnage adaptatif appliqué à la Piste n°16 ♫♪.
Contrôle : le facteur de ré-échantillonnage est fonction du RMS avec un *mapping* non linéaire différent, variant entre 0.25 et 2.
Remarque : cet effet permet à la fois une modification de la hauteur et du timbre, et peut s'entendre comme une "cartoonification" du son.
Piste n°27-CD2 ♫♪ (p. 207) : ré-échantillonnage adaptatif appliqué à la Piste n°16 ♫♪.
Contrôle : le facteur de ré-échantillonnage est fonction du RMS avec un *mapping* non linéaire encore différent, variant entre 0.25 et 4.
Remarque : cet effet permet à la fois une modification de la hauteur et du timbre, et peut s'entendre comme une "cartoonification" du son. Les facteurs extrêmes 0.25 et 4 est souvent atteint du fait du choix du *mapping*, ce qui explique que le son soit très différent de l'original (très grave et lent, ou très aigu et rapide).

A.7.3 Brassage adaptatif

Piste n°28-CD2 ♫♪ (p. 207) : brassage adaptatif de la voix de Pierre Schaeffer Piste n°16 ♫♪.
Contrôle : la position variable de lecture (avec respect de la hauteur et calcul des phases par vocodeur de phase) est fonction du RMS.
Remarque : le son est réinterprété, et le choix du ratio global de dilatation/contraction permet de choisir la vitesse de déplacement dans le son initial (ici élevée). Ceci permet notamment de créer des sons qui semblent répétitifs (boucles), mais varient d'une répétition à l'autre.
Piste n°29-CD2 ♫♪ (p. 207) : brassage adaptatif du son de flûte Piste n°7 ♫♪.
Contrôle : la position variable de lecture (avec respect de la hauteur et calcul des phases par vocodeur de phase) est fonction du RMS.
Remarque : le son est réinterprété, et le ratio global de dilatation/contraction est élevé. Ceci permet notamment de créer des sons qui semblent répétitifs (boucles), mais varient d'une répétition à l'autre. La voix devient incompréhensible, mais le timbre est toujours reconnaissable.
Piste n°30-CD2 ♫♪ (p. 207) : brassage adaptatif du glissé de cordes de basse Piste n°9 ♫♪.
Contrôle : la position variable de lecture (avec respect de la hauteur et calcul des phases par vocodeur de phase) est fonction du RMS.
Remarque : le son est réinterprété, et le choix du ratio global de dilatation/contraction permet de choisir la vitesse de déplacement dans le son initial (ici élevée). Ceci permet notamment de créer des sons qui semblent répétitifs (boucles), mais varient d'une répétition à l'autre.

A.7.4 Martianisation

Piste n°31-CD2 ♫♪ (p. 208) : martianisation de la voix parlée Piste n°15 ♫♪.
Contrôle : modulation de fréquence adaptative (transposition) allant jusqu'à ± 2 octaves.
Remarque : pour des variations de hauteur trop grandes ou trop rapides de hauteur, on peut perdre complètement le sens du message original, et obtenir un son de "martien".

A.7.5 Modulation en anneau adaptative

Piste n°32-CD2 ♫♪ (p. 210) : modulation en anneau adaptative de la Piste n°12 ♫♪.
Contrôle : la fréquence de modulation est fonction de la demi-fréquence fondamentale.
Remarque : la hauteur baisse d'une octave et le son reste harmonique.
Piste n°33-CD2 ♫♪ (p. 210) : modulation en anneau adaptative de la Piste n°12 ♫♪.
Contrôle : la fréquence de modulation est fonction du quart de la fréquence fondamentale.
Remarque : La hauteur baisse de deux octaves et le son reste harmonique.
Piste n°34-CD2 ♫♪ (p. 210) : modulation en anneau adaptative de la Piste n°12 ♫♪.
Contrôle : la fréquence de modulation est fonction de la fréquence fondamentale.

Remarque : la hauteur est conservée, et l'enveloppe spectrale (donc le timbre) légèrement modifiée, bien que le son reste harmonique.
Piste n°35-CD2 ♫♪ (p. 210) : modulation en anneau adaptative de la Piste n°12 ♫♪.
Contrôle : la fréquence de modulation est fonction du double de la fréquence fondamentale.
Remarque : la hauteur est conservée, et le timbre plus modifié encore que précédemment.

A.7.6 Modulation en anneau sans et avec conservation de l'enveloppe

Piste n°36-CD2 ♫♪ (p. 211) : modulation en anneau adaptative de la Piste n°12 ♫♪.
Contrôle : la fréquence de modulation est fonction de 1.8 F_0, sans conservation des formants.
Remarque : le timbre est grandement modifié du fait que le son n'est plus harmonique.
Piste n°37-CD2 ♫♪ (p. 211) : modulation en anneau adaptative de la Piste n°12 ♫♪.
Contrôle : la fréquence de modulation est fonction de 1.8 F_0, avec préservation des formants.
Remarque : le timbre est bien plus reconnaissable puisque les formants sont conservés. Il reste la rugosité et la légère inharmonicité, dues à la modulation en anneau non contrôlée par un multiple ou un sous-multiple de F_0.
Piste n°38-CD2 ♫♪ (p. 211) : modulation en anneau adaptative de la Piste n°16 ♫♪.
Contrôle : la fréquence de modulation est fonction de 1.8 F_0, sans conservation des formants.
Remarque : le timbre est beaucoup modifié par rapport à la Piste n°15 ♫♪ du fait que le son n'est plus harmonique.
Piste n°39-CD2 ♫♪ (p. 211) : modulation en anneau adaptative de la Piste n°16 ♫♪.
Contrôle : la fréquence de modulation est fonction de 1.8 F_0, avec préservation des formants (quéfrence de 538).
Remarque : le timbre est plus reconnaissable que la Piste n°38-CD2 ♫♪ puisque les formants sont conservés. Il reste la rugosité et la légère inharmonicité, dues à la modulation en anneau.
Piste n°40-CD2 ♫♪ (p. 211) : modulation en anneau adaptative de la Piste n°16 ♫♪.
Contrôle : la fréquence de modulation est fonction de 1.8 F_0, avec préservation des formants (quéfrence de 1077).
Remarque : le timbre est encore plus reconnaissable que la Piste n°39-CD2 ♫♪ puisque les formants sont mieux conservés par l'application d'une enveloppe plus précise (quéfrence de coupure plus élevée). Il reste toujours la rugosité.

A.7.7 Modulation en anneau spectrale et adaptative

Piste n°41-CD2 ♫♪ (p. 211) : modulation en anneau spectrale adaptative de la Piste n°12 ♫♪.
Contrôle : la fréquence de modulation varie entre 64 et 80 Hz, et est donnée pour chaque panier de fréquence par l'enveloppe spectrale.
Remarque : l'effet ressemble à la modulation en anneau adaptative, mais permet d'obtenir un peu plus de rugosité, et des parcours fréquentiels différents pour chaque harmonique.

A.7.8 Panoramisation-octaviation adaptative

Piste n°42-CD2 ♫♪ (p. 211) : panoramisation-octaviation adaptative sur la voix parlée Piste n°16 ♫♪.
Contrôle : le son est à la fois panoramisé adaptativement, et mixé entre sa version transposée à l'octave inférieure et sa version originale.
Remarque : cet effet joue sur l'ambiguïté de position et d'octave.
Piste n°43-CD2 ♫♪ (p. 211) : panoramisation-octaviation adaptative sur la voix parlée Piste n°19 ♫♪.
Contrôle : le son est à la fois panoramisé adaptativement, et mixé entre sa version transposée à l'octave supérieure et sa version originale.
Remarque : cet effet joue sur l'ambiguïté de position et d'octave.

A.8 Exemples accompagnant les stratégies de *mapping*

A.8.1 Quantification

Piste n°44-CD2 ♫♪ (p. 188) : écho granulaire adaptatif obtenu hors temps-réel appliqué à la Piste n°11 ♫♪.
Contrôle : le gain est constant et le délai variable, à l'échantillon près.
Remarque : cet exemple sert de référence pour la recherche d'une bonne quantification de la courbe de contrôle.
Piste n°45-CD2 ♫♪ (p. 188) : écho granulaire adaptatif temps réel obtenu par une quantification uniforme de la courbe de contrôle, appliqué à la Piste n°11 ♫♪.
Contrôle : la courbe de délai est quantifié avec $n_q = 3$.
Remarque : le nombre de valeurs de quantification est très insuffisant ; l'utilisation de seulement trois lignes à retard s'entend comme tel, et le son obtenu diffère trop du son de référence.
Piste n°46-CD2 ♫♪ (p. 188) : écho granulaire adaptatif temps réel obtenu par une quantification uniforme de la courbe de contrôle, appliqué à la Piste n°11 ♫♪.
Contrôle : la courbe de délai est quantifiée avec $n_q = 10$.
Remarque : le nombre de valeurs de quantification est insuffisant, même si on commence à reconnaître le son de référence.
Piste n°47-CD2 ♫♪ (p. 188) : écho granulaire adaptatif temps réel obtenu par une quantification uniforme de la courbe de contrôle, appliqué à la Piste n°11 ♫♪.
Contrôle : la courbe de délai est quantifiée avec $n_q = 30$.
Remarque : le son est assez ressemblant au son de référence, mais il manque encore de précision.
Piste n°48-CD2 ♫♪ (p. 188) : écho granulaire adaptatif temps réel obtenu par une quantification uniforme de la courbe de contrôle, appliqué à la Piste n°11 ♫♪.
Contrôle : la courbe de délai est quantifiée avec $n_q = 60$.
Remarque : le nombre de valeurs de quantification est suffisant : le son est quasiment identique au son de référence.

A.8.2 Zoom

Piste n°49-CD2 ♫♪ (p. 235) : écho granulaire adaptatif, appliqué à la Piste n°16 ♫♪.
Contrôle : le contrôle est donné par une combinaison de descripteurs, sans application de fonction loupe. *Remarque* : cet exemple sert de référence pour présenter l'effet de la fonction de zoom sur un contrôle de l'écho granulaire adaptatif.
Piste n°50-CD2 ♫♪ (p. 235) : écho granulaire adaptatif, appliqué à la Piste n°16 ♫♪.
Contrôle : la courbe de contrôle se voit appliquer une fonction loupe $\mathcal{Z}_2(t)$.
Remarque : certaines portions du son se sont vues attribuer des gains et délais différents, et sont alors plus mises en valeur, par exemple à la fin du son.
Piste n°51-CD2 ♫♪ (p. 235) : transposition adaptative de la voix parlée Piste n°16 ♫♪ (à la limite de la robotisation).
Contrôle : le contrôle est donné par une combinaison de descripteurs, sans application de fonction loupe. *Remarque* : cet exemple sert de référence pour présenter l'effet de la fonction de zoom sur un contrôle de la transposition adaptative.
Piste n°52-CD2 ♫♪ (p. 235) : transposition adaptative de la voix parlée Piste n°16 ♫♪.
Contrôle : la courbe de contrôle se voit appliquer une loupe $\mathcal{Z}_2(t)$ pour modifier la courbe de contrôle.
Remarque : cela permet de donner plus d'ambitus, notamment sur la fin du son où le paramètre de contrôle varie peu, avant application de la loupe.
Piste n°53-CD2 ♫♪ (p. 235) : transposition adaptative de la voix parlée Piste n°16 ♫♪.
Contrôle : avec utilisation de la loupe $\mathcal{Z}_4(t)$ pour modifier la courbe de contrôle.
Remarque : cela permet de donner plus d'ambitus, notamment sur la fin du son où le paramètre de contrôle varie peu, avant application de la loupe.

Piste n°54-CD2 ♫♪ (p. 235) : dilatation/contraction temporelle adaptative de la Piste n°16 ♫♪.
Contrôle : le contrôle est donné par une combinaison de descripteurs, sans application de fonction loupe. *Remarque* : cet exemple sert de référence pour présenter l'effet de la fonction de zoom sur un contrôle de la dilatation/contraction adaptative.
Piste n°55-CD2 ♫♪ (p. 235) : dilatation/contraction temporelle adaptative.
Contrôle : avec utilisation de la loupe $\mathcal{Z}_1(t)$ pour modifier la courbe de contrôle.
Remarque : le son est plus dilaté et contracté aux passages où le contrôle variait peu avant application de la loupe, notamment au milieu où le son est moins ralenti et sur la fin où il est beaucoup plus ralenti.

A.9 Exemples complémentaires utilisés lors de la soutenance

Certains sons utilisés lors de la soutenance sont présentés ici, du fait de leur intérêt quant à leur aspect illustratif des effets adaptatifs.
Piste n°56-CD2 ♫♪ : bruit des freins d'un bus (utilisé pour l'illustration *fig.* 1.5).
Piste n°57-CD2 ♫♪ (p. 91) : effet adaptatif de synthèse croisée.
Contrôle : l'enveloppe spectrale est extraite de la Piste n°17 ♫♪ et la source de la Piste n°20 ♫♪.
Remarque : la synchronisation entre les deux sons (au niveau rythmique) explique la réussite de l'effet musical.
Piste n°58-CD2 ♫♪ : effet adaptatif croisé de changement d'amplitude adaptative (canal de gauche).
Contrôle : le son de la Piste n°5 ♫♪ est modulé en amplitude par l'amplitude du son de la Piste n°7 ♫♪ (canal de droite).
Remarque : les deux sons ayant la même courbe d'amplitude, le son du "chant des Dunes" vient renforcer la présence de la flûte, en lui ajoutant un double dans les basses fréquences.
Piste n°59-CD2 ♫♪ : effet croisé de conformation spectrale adaptative.
Contrôle : on applique l'effet à la Piste n°7 ♫♪ en fonction de l'amplitude du son de la Piste n°17 ♫♪.
Remarque : le son de flûte est transformé pendant les moments de parole, et non-transformé pendant les silences.
Piste n°60-CD2 ♫♪ : Piste n°59-CD2 ♫♪ au canal de gauche et Piste n°17 ♫♪ au canal de droite.
Remarque : le fait d'entendre les deux sons en même temps permet de mieux saisir l'action de la voix de Pierre Schaeffer sur le son de flûte.
Piste n°61-CD2 ♫♪ (p. 262) : création du 18 octobre 2003 au Café Julien à Marseille. Cette pièce est écrite pour flûte traversière, effets audionumériques et bande. Les effets adaptatifs utilisés sont l'équaliseur adaptatif, dans sa forme d'instrument "Noisonic" (pour la bande), la panoramisation adaptative, le trémolo adaptatif, l'écho adaptatif et le filtrage adaptatif pour le dispositif traitant les sons de flûte. L'ensemble a été réalisé en temps-réel sous *Max/MSP*, et contrôlé par le *rack* d'effets **Digitech Valve-Fx** et le pédalier de contrôle associé **Control One**.

Annexe B

Descriptif des vidéos

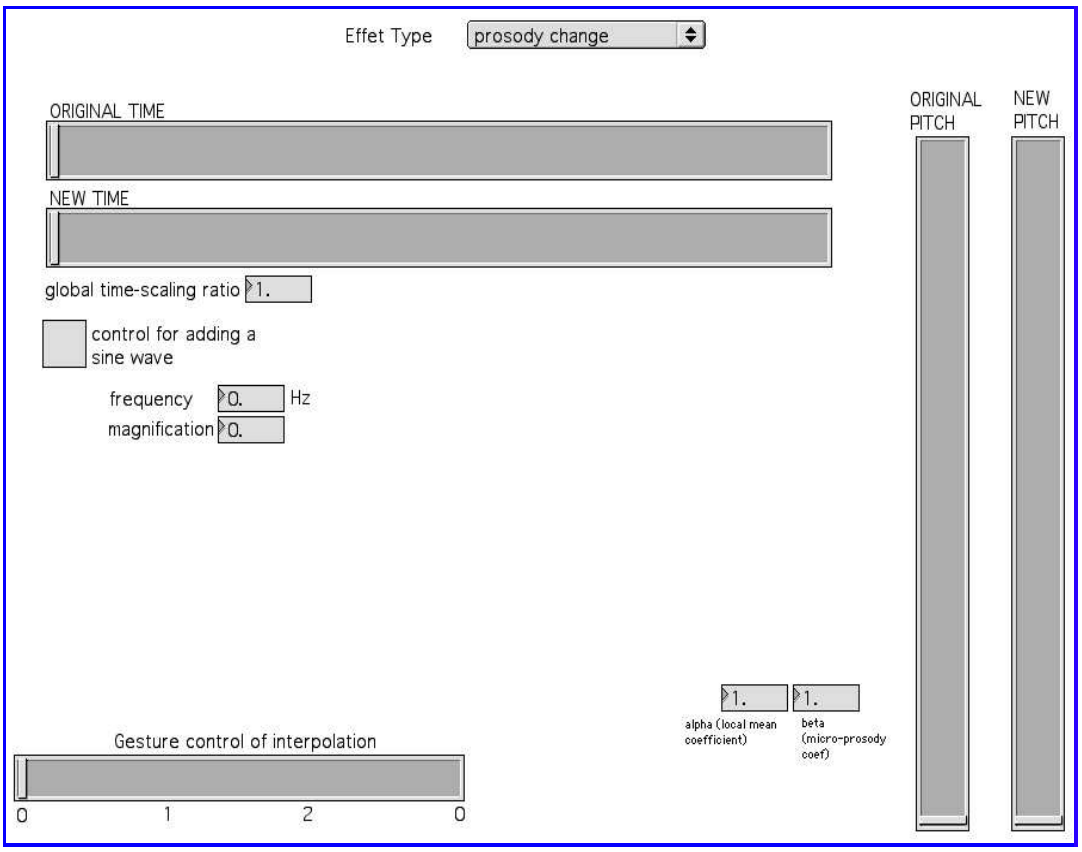
Nous présentons ici les quatre vidéos en les explicitant. Chaque vidéo peut être vue directement dans le document PDF avec **Acrobat Reader** (version 6), en cliquant directement sur l'image.

B.1 Spatialisation adaptative, janvier 2002

Dans cette vidéo, le dispositif octophonique de spatialisation adaptative est contrôlé par la danseuse, via les capteurs à ultrasons. La danseuse (Cécile Vallet) dispose de l'espace entre les haut-parleurs pour sa chorégraphie, improvisée et adaptée au son généré par le système. L'animation 3D en bas à droite représente le dispositif, avec la position de la danseuse et celle du son. Notons que les angles de vue des deux figures sont différents : l'angle de vue de la figure de gauche correspond à la direction de l'axe Y sur la figure de droite. La position du son est donnée par la moyenne sphère à droite (en bleu). La position de la danseuse, ou du moins la position donnée par les dernières

interruptions des faisceaux vertical et horizontal, est donnée par la grande sphère au milieu (en rouge). La configuration utilisée pour la vidéo est celle où la dimension verticale donne la longueur de l'arc d'ellipse sur lequel le son peut se déplacer. Elle permet de passer d'un son localisé en une position choisie par l'interprète lorsqu'il entre dans le faisceau vertical, à hauteur d'épaule, à un son multi-localisé lorsque l'interruption du faisceau vertical se fait entre les épaules et le sol. Des mouvements de jambes ou de pieds influent sur la localisation précise ou floue du son.
Pour retourner au texte, se rendre p. 261.

B.2 Conférence DAFx-03 à Londres, septembre 2003



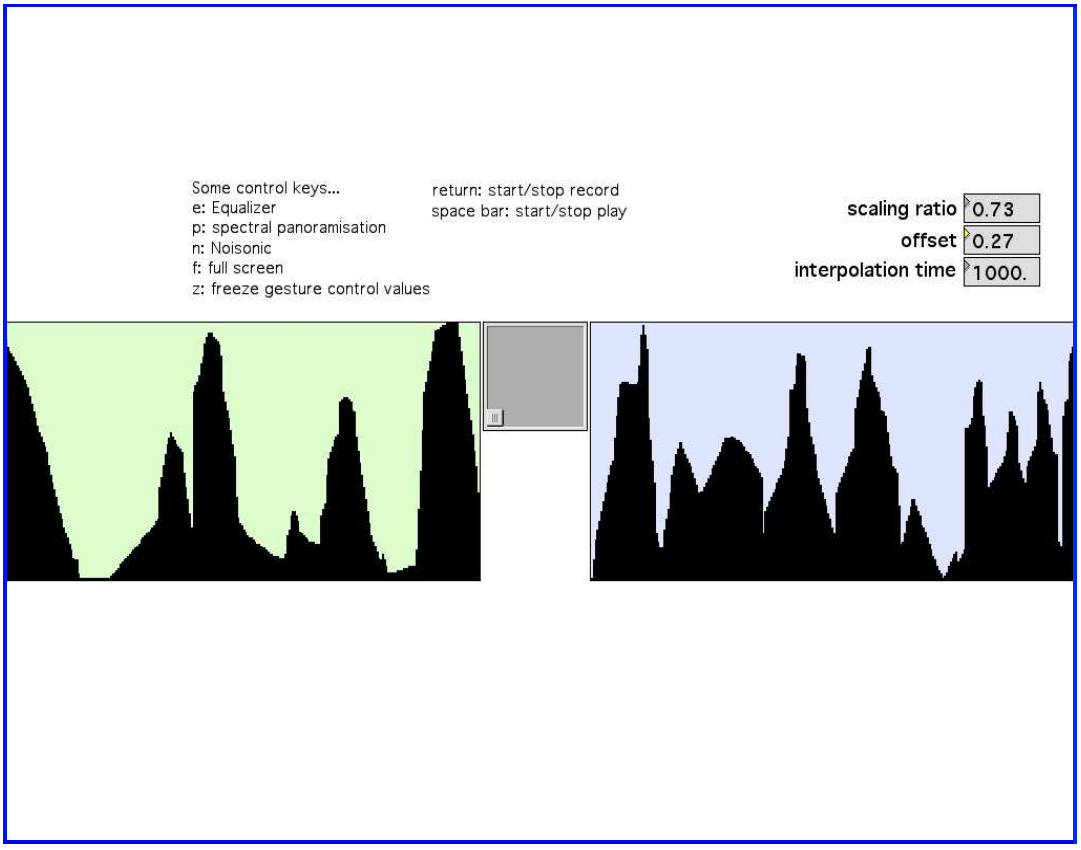
Dans cette vidéo présentée lors de la conférence DAFx-03 à Londres, on montre le contrôle gestuel et adaptatif de la transposition et de la dilatation/contraction temporelle en quatre étapes :

- dans la première partie, le contrôle gestuel porte directement sur le facteur de transposition (gauche/droite) et sur le ratio de contraction/dilatation (haut/bas), de manière exploratoire.
- dans la deuxième partie, le contrôle gestuel porte sur le ratio de dilatation/contraction temporelle, puis sur le facteur de transposition, mais cette fois-ci synchronisé avec le déroulement du son. On voit bien qu'il faut une bonne connaissance du son et des mouvements rapides si l'on veut suivre finement l'évolution du son.
- dans la troisième partie, le contrôle gestuel porte sur le *mapping* de la transposition adaptative, de façon à permettre des changements de prosodie. Ainsi, la prosodie est tout d'abord complètement aplatie (*joystick* au centre, $\alpha = \beta = 0$), puis amoindrie (*joystick* légèrement à droite (*joystick* un peu à droite et au fond, $\alpha, \beta \in [0; 0.5]$), puis exagérée (*joystick* à droite et au fond, $\alpha, \beta > 1$). La prosodie est ensuite inversée (*joystick* à gauche et devant, $\alpha, \beta < 0$).
- dans la quatrième partie, le contrôle gestuel porte sur le *mapping* de la dilatation/contraction adaptative, de manière à permettre des changements d'expressivité et de déroulement tem-

porel. Deux courbes de ratio de dilatation/contraction temporelle permettant une dilatation/contraction avec préservation de la durée sont pré-calculées. Le mouvement de torsion du *joystick* permet de calculer l'indice d'interpolation entre ces courbes. Le mouvement d'avant en arrière permet de modifier le ratio global de dilatation/contraction temporel.
- dans la cinquième et dernière partie, le contrôle gestuel porte sur l'amplitude d'une fonction ajoutée au ratio de dilatation/contraction temporelle. Cette fonction est sinusoïdale, d'amplitude dépendant du mouvement gauche/droite et de fréquence dépendant du mouvement avant/arrière. Ceci permet de créer un effet de hachage, de granulation de la voix.

Pour retourner au texte, se rendre p. 246.

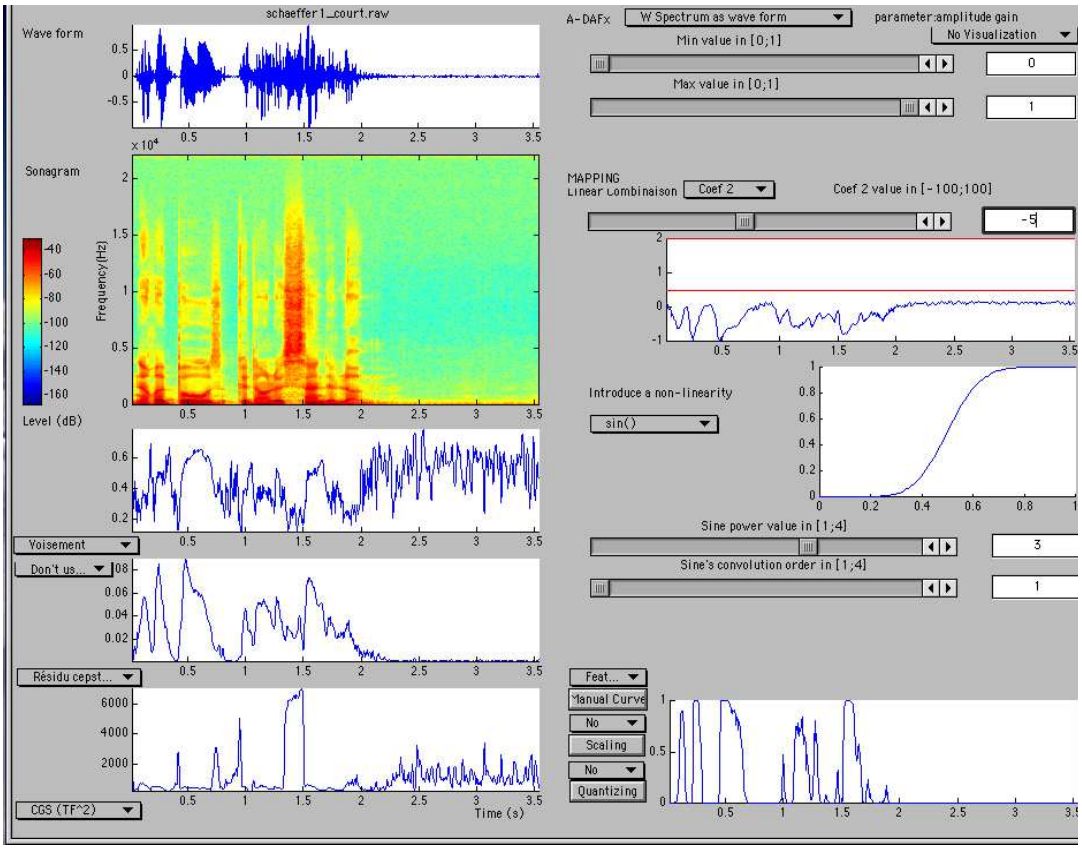
B.3 Equaliseur adaptatif



Dans cette vidéo, l'équaliseur adaptatif est présenté selon plusieurs configurations :
- la première configuration est l'équaliseur adaptatif stéréophonique contrôlé par le *joystick*. Selon l'axe gauche-droite, on contrôle l'amplitude de la fonction de transfert du filtre (formes de gauche et de droite sur la vidéo). Selon l'axe de torsion, on contrôle le taux de rafraîchissement de la forme de contrôle, calculée par interpolation entre deux formes d'ondes de grains extraits de la Piste n°6 ♫♪. Le bouton de "tir" du *joystick* permet de choisir si l'axe de torsion est actif ou inactif.
- la deuxième configuration est l'équaliseur adaptatif stéréophonique avec deux fonctions de transfert complémentaires (au sens de la puissance). Ainsi, l'effet devient une panoramisation spectrale adaptative à puissance constante, ou chaque panier de fréquence est spatialisé différemment de ses voisins.
- la troisième configuration est l'équaliseur adaptatif appliqué à un bruit blanc. Il s'agit d'un effet adaptatif croisé, que l'on peut encore considéré comme un instrument de synthèse de

bruits (corrélés avec un autre son, utilisé pour le contrôle), appelé ici “Noisonic”.
Pour retourner au texte, se rendre p. 199.

B.4 Interface en temps-différé (*Matlab*)



Dans cette vidéo, on fait la démonstration de l'interface *Matlab* de contrôle des effets audio-numériques adaptatifs en temps-différé, en plusieurs étapes :

- on présente d'abord le menu permettant d'accéder aux fonctionnalités ;
- on montre ensuite les différentes parties de l'interface :
  - la représentation du signal, avec la forme d'onde, le sonagramme et les harmoniques ;
  - les descripteurs extraits du son et accessibles via un menu déroulant ;
- on applique enfin différents *mappings* :
  - on choisit l'effet et les valeurs des bornes de ses contrôles (ainsi que du type d'harmonisation dans l'exemple vidéo) ;
  - on applique la première étape de *mapping* (combinaison) : la combinaison linéaire des descripteurs, puis l'application d'une fonction de conformation à cette combinaison linéaire (troncature puis fonction sinusoïdale) ;
  - on applique la deuxième étape du *mapping* (conformation spécifique) : l'application d'une fonction de zoom, puis d'une fonction de quantification.

Pour retourner au texte, se rendre p. 243.

Annexe C

Nouveau Chapitre de la Thèse : conduite du projet de recherche

Sommaire

Depuis plusieurs années, le Ministère de l'Education Nationale et plusieurs Ecoles Doctorales d'Ile-de-France ont décidé de s'attacher à rendre les thèses plus professionnalisantes, en vue de permettre aux doctorants ne pouvant pas ou ne voulant pas continuer dans le domaine de la recherche de mieux valoriser leurs travaux vis-à-vis d'un employeur extérieur au monde de la recherche publique française. Il a été mis en place une expérience pilote, qui consiste à demander à quelques étudiants en thèse d'inclure dans leur mémoire un chapitre présentant leurs travaux en tant que projet de recherche. Cette expérience en est maintenant à sa troisième année. L'Association Bernard Gregory (ABG) est maître d'œuvre : 74 doctorants on participé à la deuxième année et 112 participent à la troisième année. Par cette expérience, j'espère contribuer à la réflexion des autres doctorants, que j'ai représentés pendant une année auprès du Conseil de mon Ecole Doctorale, ainsi qu'à leurs encadrants.

C.1 Cadre général et enjeux de ma thèse

C.1.1 Pourquoi ce sujet ?

Après une mûre réflexion durant l'année 1999 où je travaillais comme consultant en informatique, je me suis décidé à me réorienter vers un domaine qui me passionnait déjà mais ne correspondait pas à mon secteur d'activité : la recherche, plus particulièrement en informatique musicale. Aussi ai-je postulé au DEA Acoustique, Traitement du Signal et Informatique Appliqués à la Musique (ATIAM), mis en place en 1993 par les Université de Paris VI et Aix-Marseille II[1]. Dès le début, j'ai fait connaître ma motivation pour entreprendre une thèse. J'ai effectué mon stage de DEA à Sup'Télécom Paris, tout en cherchant activement un financement de thèse. Finalement, j'ai eu le

[1] ainsi que l'Ecole Normale Supérieure d'Ulm, l'IRCAM, Sup'Télécom Paris, Polytechnique Grenoble et l'Université du Mans.

choix entre plusieurs bourses, dont la bourse docteur-ingénieur (BDI) cofinancée par le CNRS et la Région Provence-Alpes-Côte-d'Azur.

Mon choix s'est porté à la fois sur le laboratoire (CNRS-LMA), le sujet, le directeur de thèse (Daniel Arfib), le financement et la situation géographique. Il me semblait important que les conditions scientifiques, relationnelles et financières soit suffisantes. Mon directeur de thèse et moi avons défini le domaine et le sujet précis de la thèse. Nous avons souhaité que l'originalité de la recherche soit porteuse en résultats scientifiques et techniques et qu'elle offre des débouchés intéressants pour les années à venir. Mon sujet de thèse au Laboratoire de Mécanique et d'Acoustique (LMA) répondait à tous ces critères.

C.1.2 Présentation des effets audionumériques adaptatifs

Ma thèse de doctorat porte sur les effets audionumériques adaptatifs. Dans le cadre des outils numériques de traitement du son, les effets audionumériques sont très couramment utilisés dans les studios par les professionnels ainsi que dans les installations dites «*Home Studio* »par les particuliers. Leur place dans la production musicale est très importante, tout au long de la chaîne de traitement sonore : de l'acquisition jusqu'à la finition (*mastering*). Leur utilisation est grandissante en musique électroacoustique et contemporaine ainsi que dans les courants populaires tel que la «Techno ». Ces effets offrent à l'utilisateur des contrôles simples. Ils se présentent sous forme de boutons ou de potentiomètres (linéaires ou circulaires). Ils peuvent être réels (manipulés à l'aide de consoles ou de périphériques utilisant la norme MIDI[2]) ou virtuels (visualisés à l'écran et manipulés via une souris ou une tablette graphique).

Ce que l'on propose avec les effets adaptatifs, c'est que le contrôle des effets soit automatisé par le son, à l'aide de descripteurs issus du son (des paramètres qui explicitent le son, la perception que l'on en a et leurs variations au cours du temps). Le son pilote directement l'effet qui lui est appliqué. Pour y parvenir, il faut dans un premier temps extraire des descripteurs du son, puis dans un second temps modifier les contrôles de l'effet en fonction des descripteurs. Le contrôle gestuel influence la manière dont on fait correspondre les descripteurs du son aux contrôles de l'effet (on appelle ces fonctions de correspondance le *mapping*). Cette approche plus générale des effets audionumériques permet de décrire de nouveaux effets, de nouveaux types de contrôle, de nouvelles voies d'utilisation pour la composition.

C.1.3 Enjeux

Cette étude comporte deux enjeux : mener une **recherche pluridisciplinaire** et **réaliser des programmes informatiques** de traitements qui soient utilisables pour la composition. Concernant la pluridisciplinarité, le projet s'inscrit dans le contexte «Arts, Science et Technologie ». La partie scientifique fait appel à des connaissances en traitement du signal musical (pour l'analyse et la transformation du son), en psychoacoustique, en systèmes de description de contenu sonore (MPEG[3], SDIF[4] pour l'extraction des paramètres descriptifs du son), en automatique, en contrôle gestuel (à l'aide de transducteurs), en cognition musicale. Rassembler toutes ces connaissances a permis de poser le cadre de mon travail. La réalisation d'effets et de traitements sonores à paramètres variables dans le temps nécessite des mises en œuvres spécifiques, via la réalisation de programmes informatiques. Deux environnements ont été utilisés : *Matlab* pour le temps différé, *Max/MSP* pour le temps-réel et le contrôle gestuel. Concernant la composante musicale de ma thèse, j'ai suivi les

[2]MIDI : *Musical Instrument Digital Interface*, interface numérique pour instrument de musique.
[3]MPEG = *Moving Picture Experts Group*, groupe international d'experts pour le développement de normes de codage audio et vidéo.
[4]SDIF = *Spectral Data Information File*, format de fichier d'informations spectrales.

cours de Composition en Electroacoustique et les cours d'Histoire de la Musique Contemporaine, au Conservatoire National de Région de Marseille (CNR).

Nous attendions plusieurs types de résultats :

- des résultats pratiques, via la mise au point de programmes de traitement adaptatifs ;
- des résultats scientifiques de recherche en amont, via la mise en place du cadre général incluant les effets traditionnels et le contrôle gestuel (systèmes de synthèse sonore) ;
- des résultats musicaux et des résultats techniques : la réalisation de transformations d'aspect moins systématique grâce au contrôle adaptatif (de nouveaux résultats ou effets perceptifs), la création de transformations sonores inconnues, la réalisation de transformations plus réalistes que certaines existantes (ex : le ralenti sélectif d'un son), l'utilisation des programmes en situation musicale (utilisation en concert, composition de pièces).

Depuis quatre ans, quelques logiciels incluant des effets adaptatifs sont commercialisés[5]. Lors de présentations dans des conférences internationales, nous avons pu remarquer qu'aucun autre centre de recherche ni aucune entreprise du secteur audionumérique n'utilise notre approche systématique d'investigation. Il s'agit d'une recherche originale et innovante du fait de la largeur du spectre de traitements investigués. Elle se place dans la continuité de la créativité et de l'inventivité dans les traitements sonores, amorcée par l'utilisation de l'électronique au début du XX^{e} siècle.

C.1.4 Projet de recherche dans son contexte

Mon sujet se place comme un pont allant des effets audionumériques vers le contrôle gestuel. Depuis sa création, l'équipe «Acoustique Perceptive et Informatique Musicale[6] »du CNRS-LMA est tournée vers les effets et les méthodes de synthèse numérique du signal musical. Durant les années quatre-vingt dix, des méthodes de haute technologie[7] ont été utilisées à la fois pour la compréhension du signal et pour la composition. Daniel Arfib et Jean-Claude Risset[8] ont participé au développement scientifique de ces méthodes et à leur utilisation artistique, par la composition d'œuvres musicales utilisant ces systèmes. En 1998, le cycle de conférences européennes *DAFx* [9] a été initié par Daniel Arfib et d'autres scientifiques européens. Depuis, l'équipe APIM s'est tournée vers le contrôle gestuel de modèles de synthèse, en commençant par le projet «le Geste Créatif en Informatique Musicale[10] ». Elle participe au projet CONGAS[11] qui a démarré en mars 2003. Ces deux thèmes, effets audionumériques et contrôle gestuel de systèmes de synthèse, continuent à être développés simultanément.

C.1.5 Moyens

Des moyens techniques et financiers ont été mis à ma disposition tout au long du projet :

- techniques : un ordinateur *Apple Macintosh* G4 équipé d'une carte son professionnelle (*Digi001* de marque *DigiDesign*), un studio d'écoute en quadriphonie, des logiciels commerciaux de traitement du signal sonore (*Matlab* de *MathWorks* et *Max/MSP* de *Cycling'74*), une bibliothèque de programmes en *Matlab* développée dans le cadre du livre «*DAFx : Digital Audio Effects* ».
- financiers : financement de mes interventions à deux conférences internationales DAFx (Limerick en 2001, Hambourg en 2002), de deux missions dans le cadre du projet COST-G6

[5] Il s'agit notamment de l'effet *Contrast* des *GRM Tools*, du logiciel *Fraktal Delay* de *Native Instruments*, de quelques pré-réglages du multi-effets logiciel *Sfx Machine RT* de *Sound Guys*.

[6] APIM = Acoustique Perceptive et Informatique Musicale.

[7] telles que le vocodeur de phase, les ondelettes, PSOLA.

[8] J.-C. Risset : chercheur émérite et médaille d'or du CNRS en 1999, compositeur (nombreux prix dont Ars Electronica 1987, Concours International de Bourges 1980 et 1998, grand prix national de la musique 1990) et interprète.

[9] DAFx : *Digital Audio Effects*, conférences internationale sur les effets audionumériques (1998-2001), projet européen COST-G6.

[10] «le Geste Créatif en Informatique Musicale » : projet de recherche financé par le Conseil Général des Bouches du Rhône.

[11] CONGAS : *Control of Gestural Audio Systems*, contrôle de systèmes audionumériques gestuels.

(*cf. tab.* C.2).
Sur le plan scientifique et humain, Daniel Arfib est particulièrement disponible lorsque j'ai besoin d'aide ou de conseils.

C.2 Déroulement, gestion et coût de mon projet

C.2.1 Préparation, financements et cadrage du projet

Bien que s'intégrant dans une discipline des sciences appliquées (Sciences Pour l'Ingénieur), ma recherche s'est déroulé très en amont. Une profonde réflexion a été menée de manière à généraliser le plus possible la méthode en envisageant tous les effets possibles. Ceci n'a pas empêché la réalisation d'outils de traitements sonores adaptatifs, mais en a conditionné la forme. Les programmes réalisés l'ont été avec des outils universitaires, dans une optique de validation.

J'ai fait appel à des financements extérieurs (*cf. tab.* C.2) : la Société Française d'Acoustique (SFA), le projet COST-G6, la Maison des Sciences de l'Homme et le CICM de l'Université Paris VIII, Supélec.

Au cours du projet, plusieurs partenaires internationaux ont été contactés, pour des travaux et publications communes :

- le Music Technology Group (MTG) à Barcelone, pour un article au *Journal of New Music Research* (avec Xavier Serra, Xavier Amatriain) ;
- l'Université des FAF de Hambourg pour un article dans le *IEEE Transactions on Audio and Speech Processing* (avec Udo Zoelzer).

C.2.2 Conduite du projet

Le programme de recherche de départ n'était ni trop vague, de manière à cerner le domaine de l'étude, ni trop précis, de manière à garder une liberté d'orientation et d'investigation, que j'ai mise à profit pour proposer de nouvelles idées. Dès le début du projet, nous avons fixé une réunion mensuelle avec Daniel Arfib et Patrick Boussard (société GENESIS S.A., Aix-en-Provence), à laquelle nous n'avons pas toujours réussi à nous tenir. A ma demande, un point d'avancement hebdomadaire a été réalisé avec Daniel Arfib, de manière à avoir l'avis averti d'une personne en retrait par rapport à mon travail quotidien. Ceci m'a permis d'éviter bien des erreurs. La qualité de son écoute et sa disponibilité m'ont permis de le rencontrer aux moments décisifs de mon travail, aussi bien pour le projet de recherche que pour les projets d'après doctorat.

C.2.3 Problèmes rencontrés et solutions apportées

Par chance, je n'ai pas rencontré de problème, excepté ceux de tout thésard, à savoir la dispersion et les baisses de motivation :

- la dispersion, due à une grande masse de travail et à la conduite de plusieurs projets à la fois, se résout très bien avec des méthodes classiques : listes et séquençage des tâches à effectuer, points réguliers (hebdomadaires, mensuels et annuels) ;
- les baisses de motivation se gèrent par une certaine opiniâtreté ainsi que par des discussions amicales dans le cadre du laboratoire visant à prendre du recul (évaluation du chemin parcouru en regard du travail restant à effectuer).

Personnel	Temps (mois)	Salaire Mensuel brut (€)	Coût (€)
Doctorant	36	1874	68000 50% CNRS 50% région PACA
Directeur de thèse	6	3500	21000
Technicien	2	3000	6000
Autres chercheurs	1	3500	3500
Autres doctorants	2	0	0
Total			98000

TAB. C.1 – *Coût en ressources humaines.*

Dates	Mission/Formation	Intitulé	Lieu	Montant	Financement
04/01	mission court-terme	COST-G6 STSM	Barcelone	1000 €	COST-G6
05/01	Communication orale	JJCAM'01	IRCAM, Paris	60 €	SFA
06/01	mission court-terme	COST-G6 STSM	Madrid	842 €	COST-G6
06/01	Formation	Matlab	Marseille	434 €	LMA
12/01	Congrès	DAFx-01	Limerick	828 €	50% SFA
					50% LMA
01/02	Colloque	Journées GMM	INSA Toulouse	64 €	INSA Tlse
02/02	Réunion – séminaire		Supélec, Paris	82 €	Supélec
04/02	Journées d'Etude	ACI Espaces Sonores	CICM , Paris	325 €	CICM
05/02	Congrès	JIM'02	Marseille	15 €	LMA
09/02	Congrès	DAFx-02	Hambourg	582 €	LMA
10/02	Séminaire	Les Doctoriales	Carry-le-Rouet	137 €	LMA
01/03	Collaboration	ACI Espaces Sonores	CICM , Paris	220 €	CICM
09/	Formation	Secourisme du Travail	Marseille	228 €	LMA
02/03	Formation	Technical Scientific Communication	Marseille	230 €	Ecole Doctorale ED 353

TAB. C.2 – *Dépenses associées au projet.*

C.2.4 Evaluation du coût consolidé de mon projet

Ma participation à ce projet constitue la principale ressource humaine. J'ai bénéficié ponctuellement de l'aide d'autres personnes (*cf. tab.* C.1).

Le budget total de ce travail est d'environ 110 k€ dont 88% sont imputables à la masse salariale. Les 12% restant concernent les frais de missions, les frais généraux (consommables) et le matériel (studios sons). Ce projet n'a pas nécessité d'investissement lourd, les moyens techniques utilisés étant tous disponibles dans l'équipe.

C.3 Compétences, savoir faire, qualités professionnelles et personnelles acquises et/ou mises en œuvre

Cette thèse constitue pour moi une expérience professionnelle majeure. Elle me permet d'avoir aujourd'hui des compétences et connaissances à la fois scientifiques, techniques et méthodologiques.

Financement	interne	Externe	global
	74 k€(67%)	37 k€(33%)	111 k€

TAB. C.3 – *Evaluation du coût total consolidé.*

C.3.1 Compétences techniques et scientifiques

Par ce projet de recherche, j'ai développé une expertise scientifique en traitement du signal sonore, en informatique musicale, en contrôle et en contrôle gestuel, en psychoacoustique. Concernant les compétences techniques, cette étude s'est réalisée dans un studio son (quadriphonique au LMA et octophonique au CICM) avec utilisation des logiciels de Musique Assistée par Ordinateur (*Max/MSP, Pro Tools, Logic Audio, Digital Performer, Cubase, Peak*). Il a donc été nécessaire d'apprendre la configuration matérielle et logicielle des ordinateurs *Apple Macintosh* utilisés. Parallèlement à ma thèse, j'ai suivi les cours de composition électroacoustique au CNR de Marseille pendant 3 ans, afin de tenir compte dans mon approche des besoins et démarches des compositeurs. J'ai développé d'autres compétences techniques personnelles, liées au domaine :

- la programmation (*Matlab, Code Warrior*) ;
- la réalisation de présentations (*PowerPoint*) ;
- le montage vidéo et son (*Final Cut Pro, iMovie*) ;
- la réalisation de sites Web (*DreamWeaver, Flash*) ;
- la prise de parole en public (théâtre d'improvisation) ;
- une connaissance et un suivi des courants de musique contemporaine ;
- une pratique de la musique en tant que musicien, technicien (son et lumières) lors de concerts ;
- l'édition de textes scientifiques (LaTeX).

C.3.2 Compétences méthodologique en conduite de projet

Cette période m'a permis de conforter des compétences utiles dans tout travail de chercheur et d'ingénieur : analyse et modélisation d'un problème, évaluation des difficultés, proposition de méthodes pour les résoudre, mise en place d'un planning. J'ai acquis des compétences méthodologiques en conduite de projet :

- gestion du temps : évaluation de la durée, respect de dates limites ;
- gestion de tâches menées en parallèle : définition des étapes, analyse et contrôle de leur réalisation ;
- réalisation de documentation : descriptifs des logiciels réalisés, articles scientifiques et mémoire.

Nous pouvons regrouper les tâches en trois catégories :

- les tâches en amont : documentation et veille scientifique et technologique (bibliographie : recherche d'informations, résumé, classification et classement) ;
- les tâches en aval : mise en valeur des résultats par le biais de communications et par l'utilisation du réseau ;
- les tâches courantes : réalisation du projet de recherche (recherche et réalisation technique des logiciels).

C.3.3 Compétences méthodologique en communication

Ces compétences concernent la constitution et l'utilisation d'un réseau, et la communication de résultats scientifiques. J'ai constitué et développé un réseau via la messagerie électronique pour informer et être informé, préparer des collaborations, rédiger des articles. La communication des résultats scientifiques s'est faite de deux manières (en anglais et/ou en français) :

◇ communications écrites :
  - rédaction d'articles scientifiques : *i)* analyse du support et des créneaux de diffusion, *ii)* choix du support, *iii)* rédaction au format demandé. Ils m'ont servi à me faire connaître, à communiquer mes résultats à un public professionnel.
  - diffusion de mon mémoire de thèse à la demande (chercheurs, professeurs, doctorants...) ;

  - réalisation d'un polycopié de cours comme support aux enseignements sur les «Effets et Transformations Sonores »à l'ENSEIRB (Bordeaux, élèves-ingénieurs en troisième année), qui m'a servi à faire connaître mes activités d'enseignements auprès d'autres établissements et universités. Ceci a débouché sur des propositions de vacations et de post-doctorat.
- ◇ communications orales (à destination d'un public professionnel) :
  - séminaires et groupes de travail : présentation de mes travaux sous forme de séminaire (Paris : Sup'Télécom, CICM ; Aix-en-Provence : Laboratoire Parole et Langage, société Génésis S.A. ; Barcelone : Music Technology Group) ou de groupe de travail (Laboratoire d'Informatique de Bordeaux , Workshop DAFx à Madrid). Ces rencontres ont été l'occasion de me faire connaître, et d'échanger des idées et des informations autour de mon sujet de recherche.
  - conférences : ces présentations d'une durée de 20 minutes ont permi à la fois une vulgarisation de mes recherches et une illustration sonore et visuelle des concepts exposés dans l'article des Actes de la conférence.

C.3.4 Méthodes de travail, gestion du temps, travail en équipe

J'ai géré mon temps à l'aide d'un échéancier et d'un planning pluriannuel, dans lequel je notais les principales tâches à effectuer et leur durée estimée. Je découpais chaque tâche en sous-tâches avec plusieurs niveaux de hiérarchie (en mois et semaines) puis leur attribuait des niveaux d'urgence.

Travaillant seul sur mon sujet au LMA, j'ai travaillé principalement en réseau, plutôt qu'en équipe. Le réseau m'a permi d'avoir des avis, des idées, des retours sur ma recherche. J'ai participé à la vie de l'équipe à travers la rédaction d'article commun, des réunions (point d'avancement de chacun, foire aux idées et aux questions), l'encadrement de stagiaires, la gestion du matériel et logistique.

C.3.5 Savoir faire administratifs, organisationnels, linguistiques

J'ai développé des savoir-faires administratifs :
- recherche du financement de ma thèse ;
- demande de subvention pour les missions, auprès du CNRS, de la SFA[12], etc. ;
- gestion du transport et du logement pour ces missions.

Mes savoir-faire organisationnels proviennent de la gestion simultanée de plusieurs projets :
- sujets de recherche ;
- rédaction d'articles ;
- préparation et réalisation d'enseignement ;
- encadrement de stagiaires.

Mes langues de travail sont le français, l'anglais, et l'espagnol (j'ai des contacts permanents avec des anglophones et des hispanophones).

C.3.6 Mes qualités personnelles

Les différentes expériences vécues durant ma thèse m'ont permis de développer et d'utiliser différentes qualités personnelles :
- une grande autonomie : j'étais le seul porteur de mon projet de recherche ;
- une grande curiosité pour l'informatique musicale et ses techniques, pour les musiques contemporaines ;

[12] SFA : Société Française d'Acoustique.

– un esprit d'analyse et de synthèse, notamment durant la rédation de l'article d'équipe ;
– une faculté de créativité et d'innovation : j'ai proposé de nouveaux effets sonores et de nouvelles techniques de traitement ;
– un bon relationnel, notamment grâce à la pratique de la scène et du travail en groupes (spectacles de théâtre d'improvisation, concerts de musique irlandaise).

Suite à différentes conversations avec des chercheurs et des doctorants, il m'a été indiqué que j'ai :

– une bonne écoute et ouverture d'esprit, appréciée lors des commissions et conseils auxquelles j'ai participé (conseil de l'Ecole Doctorale et commission de représentation des Etudiants auprès du laboratoire) ;
– une rapidité de compréhension, ce qui me permet de m'adapter rapidement (notamment lors des collaborations).

C.3.7 Construction d'un réseau personnel

Sans réseau au départ (je suis entré dans le domaine de l'informatique musicale lors du DEA), j'ai construit avec l'aide de Daniel Arfib un réseau relationnel scientifique en France et à l'étranger : j'ai utilisé les rencontres lors des conférences et séminaires, l'inscription à des sociétés savantes, la participation à des *mailing-lists*. J'ai par ailleurs pris contact avec des centres de création tel le Groupe de Recherches Musicales (GRM) de l'INA-Radio France et le Groupe de Musiques Expérimentales de Marseille (GMEM), ainsi qu'avec des compositeurs.

Mon réseau m'a été très utile. En France, j'ai pu réaliser des collaborations avec Anne Sédès du CICM dans le cadre de l'ACI «*Espaces Sonores* », ainsi qu'avec Nicolas Misdariis de l'IRCAM. Grâce à l'appui de Myriam Desainte-Catherine, j'ai donné 22 heures de vacations à l'ENSEIRB[13] (Bordeaux) en Janvier 2003. L'utilisation de mon réseau m'a permi d'effectuer des missions à l'étranger (Espagne) et de rédiger des articles avec des coauteurs étrangers (Espagne, Allemagne).

C.4 Résultats, impact de la thèse

C.4.1 Pour le laboratoire ou l'équipe, pour la recherche

J'ai lié trois domaines dès le début de l'étude : l'analyse-synthèse sonore, la psychoacoustique et les instruments de synthèse, de manière à utiliser des bases de chaque domaine pour développer une approche originale. Ceci a permis une reconnaissance internationale du laboratoire et de l'équipe en particulier dans :

– l'utilisation du geste pour le contrôle adaptatif d'effets audio,
– la réalisation de nouveaux outils de traitements sonores,
– la proposition de nouvelles démarches d'utilisation de ces effets.

Pour la recherche, un pont est jeté entre deux disciplines généralement mises dos-à-dos : sciences et musique. Les outils créés ont pour vocation leur utilisation en composition ; ils répondent à une demande du Ministère de la Recherche sur l'interdisciplinarité, via des réseaux tels «Art, Sciences et Technologies ». Il ne resterait que l'étape de l'industrialisation afin de pouvoir commercialiser ces outils.

C.4.2 Pour le laboratoire et moi-même en termes de communications scientifiques

Toute ma communication scientifique écrite s'est faite en anglais ou en français :

– 2 articles de revue (de rang A), en anglais (parus à ce jour) ;

[13] ENSEIRB : Ecole Nationale Supérieure d'Electronique, Informatique et Radiocommunications de Bordeaux.

– 2 articles en anglais en cours de rédaction, pour des revues de rang A ;
– 4 articles de conférences en anglais (ainsi qu'une présentation orale pour chaque article) ;
– 1 article de conférence en français (ainsi qu'une présentation orale) ;
– 2 chapitres du livre ACI Espaces Jeunes Chercheurs «Espaces Sonores » ;
– un polycopié de cours (115 pages) en français.

C.4.3 Pour moi-même en termes de pistes professionnelles

Après réflexion, je me trouve face à trois possibilités : la recherche, l'enseignement ou l'industrie. Je choisi aujourd'hui de me diriger vers la recherche, publique ou privée, appliquée à l'industrie. Les outils de création musicale gagnent à puiser des idées dans la recherche et à se voir réaliser pour être commercialisés.

Durant tout ce projet, j'ai noué des contacts dans les milieux de la recherche publique français et internationaux, ce qui m'a permis d'avoir l'information pour me positionner sur un poste de chercheur associé au Canada, sur des post-doctorats (Canada, Espagne). Ingénieur de formation, j'ai bénéficié d'une bourse docteur-ingénieur (BDI). Je manque actuellement de maîtrise de langages et d'environnement de programmation. J'ai cependant le souci de ne pas me couper du monde de l'industrie : j'ai noué des contacts avec des entreprises du milieu de l'audionumérique. Je compte approfondir cette démarche si mes candidatures aux post-doctorats échouent.

Réflexion, synthèse et rédaction effectués entre avril et juin 2003
avec l'aide gracieuse de Christian Puech,
responsable de l'Espace Jeunes Diplômés (EJD)
de l'Association Pour l'Emploi des Cadres (APEC) de Marseille.

Notations

$a_i(n)$: amplitude de l'harmonique numéro i à l'instant n dans le modèle additif
$A_N(x,k)$: amplitude instantanée du signal $x(n)$ sur une fenêtre de taille N autour de l'index k
$A_s(x,k)$: amplitude instantanée de la partie harmonique du signal $x(n)$ autour de l'index k
$A_r(x,k)$: amplitude instantanée de la partie résiduelle du signal $x(n)$ autour de l'index k
$b_{g/a}(x,k)$: balance grave/aigü du signal $x(n)$ autour de l'index k
B_{Aures} : brillance selon le modèle d'Aures
$B_{Z\&F}$: brillance selon le modèle de Zwicker et Fastl
$C_N(y,k)$: fonction d'autocorrélation du signal à court-terme $y(n)$ sur une fenêtre de taille N autour de l'index k
$\mathcal{D}_i(t)$: descripteur du signal
ΔL : différence de niveau inter-canaux
$\mathcal{E}(f)$: enveloppe spectrale
$E_N(x,k)$: énergie du signal $x(n)$ sur une trame temporelle de taille N centrée autour de l'index k
$\mathcal{E}_{cep}(f)$: enveloppe spectrale calculée par le Cepstre
$\mathcal{E}_{lpc}(f)$: enveloppe spectrale calculée par la LPC
$f_i(n)$: fréquence de l'harmonique numéro i à l'instant n dans le modèle additif
$\mathsf{Flux}(X, r+1)$: flux spectral de signal de TFCT $X(f)$ entre la trame r et la trame $r+1$
$\mathsf{HZCRR}_N(x,k)$: variation du taux de passage par zéros du signal $x(n)$ mesuré sur une fenêtre de taille N autour de l'index k
$HFC(x,k)$: contenu en hautes fréquences du signal $x(n)$ autour de l'index k
$I_{note}(n)$: indicateur de début de note
$ICCC$: fonction d'intercorrélation entre canaux d'un son stéréophonique
L_N : isosonie
$LSTER(x,k)$: taux de basses énergies du signal $x(n)$ autour de l'index k
$\mathcal{N}_{env}(x,k)$: norme du résidu de la resynthèse cepstrale
N : sonie
N' : sonie spécifique
N_{max} : maximum de sonie
R_A : pas d'analyse
R_S : pas de synthèse
$\mathcal{R}_{spec}(x,k)$: point de roulement spectral du signal $x(n)$ autour de l'index k
$\mathcal{S}(f)$: source
SPL_A, SPL_B, SPL_C : niveau de pression sonore avec différentes pondérations
$S_{tilt}(x,k)$: pente spectrale du signal $x(n)$ autour de l'index k
$T_{BE}(x,k)$: Taux de trames de basses énergies
$\mathsf{Vois}_N(x,k)$: indice de voisement du signal $x(n)$ autour de l'index k
τ_{ICCC} : temps de meilleure intercorrélation entre canaux d'un son stéréophonique
$w_A(n)$: fenêtre d'analyse
$w_S(n)$: fenêtre de synthèse
$x(n)$: signal d'entrée (temporel)
$X(f)$: TFCT du signal d'entrée (fréquentiel)

$y(n)$: signal de sortie (temporel)
$Y(f)$: TFCT du signal de sortie (fréquentiel)
$\mathbf{ZCR}(x, k)$: taux de passage par zéros du signal $x(n)$ autour de l'index k
$\mathbb{1}$: fonction indicatrice ou fonction de Heaveside

Table des figures

Liste des tableaux

Index

Bibliographie

[Abromont and de Montalembert, 2001] C. Abromont and E. de Montalembert. *Guide de la théorie de la musique*. Fayard – Henri Lemoine, 2001. 14

[AFNOR, 1977] AFNOR. *Recueil des normes francaise de l'acoustique, Tome 1 (vocabulaire)*. Association Francaise de Normalisation, 1977. 16

[Albèra, 1997a] P. Albèra. *Musiques en création*, chapter Entretien avec Iannis Xenakis, pages 114–21. Contrechamps, 1997. 155

[Albèra, 1997b] P. Albèra. *Musiques en création*, chapter Entretien avec Luigi Nono, pages 87–102. Contrechamps, 1997. 251

[Allen and Rabiner, 1977] J. B. Allen and L. R. Rabiner. A unified approach to Short-Time Fourier analysis and synthesis. *Proc. of the IEEE*, 65(11) :1558–64, 1977. 167

[Allen, 1977] J. B. Allen. Short Term Spectral Analysis, Synthesis and Modification by Discrete Fourier Transform. *IEEE Trans. on Ac., Speech and Sig. Proc.*, 25(3) :235–8, 1977. 167

[Amatriain *et al.*, 2001] X. Amatriain, J. Bonada, A. Loscos, and X. Serra. Spectral Modeling for Higher-level Sound Transformations. In *MOSART Wkshp on Current Research Directions in Computer Music, Audiovisual Institute, Pompeu Fabra University, Barcelona*, 2001. 68, 83, 90, 119, 124, 125, 130, 131, 184, 202

[Amatriain *et al.*, 2002] X. Amatriain, J. Bonada, A. Loscos, and X. Serra. *DAFX - Digital Audio Effects*, chapter Spectral Processing, pages 373–438. U. Zoelzer ed., John Wiley & Sons, 2002. 83, 202

[Amatriain *et al.*, 2003] X. Amatriain, J. Bonada, A. Loscos, J. L. Arcos, and V. Verfaille. Content-based transformations. *J. New Music Research*, 2003. 149, 156, 178

[ANSI, 1960] ANSI. *USA Standard Acoustic Terminology*. American National Standards Institute, american national standards institute edition, 1960. 19

[Antares, 2003] Antares. Autotune, http ://www.antarestech.com/, 2003. 156

[Aphex Twin, 2001] Aphex Twin. ♩ Druqks, 2001. 270

[Aramaki *et al.*, 2002] M. Aramaki, J. Bensa, L. Daudet, P. Guillemain, and R. Kronland-Martinet. Resynthesis of coupled piano string vibrations based on physical modeling. *J. New Music Research*, 30(3), 2002. 32

[Arfib and Delprat, 1998] D. Arfib and N. Delprat. Selective transformations of Sound using Time-frequency representations : An Application to the Vibrato Modification. In *104th Convention of the Audio Engineering Society, Amsterdam*, 1998. 133, 178, 205

[Arfib and Verfaille, 2003] D. Arfib and V. Verfaille. Driving pitch-shifting and time-scaling algorithms with adaptive and gestural techniques. In *Proc. of the Int. Conf. on Digital Audio Effects (DAFx-03), London, England*, 2003. 208, 243, 246

[Arfib and Zoelzer, 2002a] D. Arfib and U. Zoelzer. *DAFX - Digital Audio Effects*, chapter Source-Filter Processing, pages 299–372. U. Zoelzer ed., John Wiley & Sons, 2002. 68

[Arfib and Zoelzer, 2002b] D. Arfib and U. Zoelzer. *DAFX - Digital Audio Effects*, chapter Time-Frequency Processing, pages 237–97. U. Zoelzer ed., John Wiley & Sons, 2002. 68, 167, 178, 181

[Arfib *et al.*, 2002a] D. Arfib, J.-M. Couturier, and L. Kessous. Gestural Strategies for Specific Filtering Processes. In *Proc. of the Int. Conf. on Digital Audio Effects (DAFx-02), Hamburg, Germany*, pages 1–6, 2002. 200

[Arfib *et al.*, 2002b] D. Arfib, J.-M. Couturier, L. Kessous, and V. Verfaille. Strategies of mapping between gesture parameters and synthesis model parameters using perceptual spaces. *Organised Sound, Mapping Strategies Issue*, 2002. 109, 152, 156

[Arfib, 1979] D. Arfib. Digital synthesis of complex spectra by means of multiplication of non linear distorded sine waves. *J. Audio Eng. Soc.*, 27 :250–65, 1979. 89, 216

[Arfib, 1998a] D. Arfib. Different Ways to Write Digital Audio Effects Programs. In *Proc. of the COST-G6 Wkshp on Digital Audio Effects (DAFx-98), Barcelona, Spain*, 1998. 253

[Arfib, 1998b] D. Arfib. *Recherches et applications en informatique musicale*, chapter Des courbes et des sons, pages 277–86. Hermès, 1998. 156

[Arons, 1992] B. Arons. Techniques, Perception, and Applications of Time-Compressed Speech. In *Proc. American Voice I/O Society Conference*, pages 169–77, 1992. 178

[Augoyard and Torgue, 1995] J.-F. Augoyard and H. Torgue. *Répertoire des effets sonores*. Parenthèses, 1995. 27

[Avanzini and Rocchesso, 2000] F. Avanzini and D. Rocchesso. Modeling Collision Sounds : Non-linear Contact Force. In *Proc. of the COST-G6 Wkshp on Digital Audio Effects (DAFx-01), Limerick, Ireland*, 2000. 32

[Bartlett, 1970] B. Bartlett. A Scientific explanation of Phasing (Flanging). *J. Audio Eng. Soc.*, 18(6) :674–5, 1970. 86

[Beauchamp, 1982] J. W. Beauchamp. Synthesis by spectral amplitude and "brightness" matching of analyzed musical instrument tones. *J. Audio Eng. Soc.*, 30(6) :396–406, 1982. 109, 126, 145

[Bennett and Rodet, 1991] G. Bennett and X. Rodet. *Current Directions in Computer Music Research*, chapter Synthesis of the singing voice, pages 19–44. Max V. Mathews and John R. Pierce, Eds., the MIT Press, Cambridge, Massachusetts, 1991. 131

[Bernardini and Rudi, 2002] N. Bernardini and J. Rudi. Compositional use of digital audio effects. *J. New Music Research*, 2002. 1

[Beuf, 2002] S. Beuf. ♩ Quartet, RCD Records, 2002. 270

[Blauert, 1983] J. Blauert. *Spatial Hearing : the Psychophysics of Human Sound Localization*. MIT Press, 1983. 189

[Bobrek and Koch, 1998] M. Bobrek and D. B. Koch. Music Signal Segmentation Using Tree-Structured Filter Banks. *J. Audio Eng. Soc.*, 46(5), 1998. 113

[Bons, 1997] J. Bons. *Musiques en création*, chapter Entretien avec Brian Ferneyhough, pages 59–64. Contrechamps, 1997. 265

[Bregman, 1990] A. Bregman. *Auditory Scene Analysis*. MIT Press, 1990. 189, 191, 258

[Brown and Zhang, 1991] J. C. Brown and B. Zhang. Musical frequency tracking using methods of conventional and "narrowed" autocorrelation. *J. Acoustic Soc. Am.*, 92(3) :1394–402, 1991. 112

[Brown, 1991] J. C. Brown. Calculation of a Constant Q Spectral Transform. *J. Acoustic Soc. Am.*, 89(1) :425–34, 1991. 112

[Brown, 1992] J. C. Brown. Musical fundamental frequency tracking using a pattern recognition method. *J. Acoustic Soc. Am.*, 92, 1992. 112

[Brun, 1979] M. Le Brun. Digital waveshaping synthesis. *J. Audio Eng. Soc.*, 27 :757–68, 1979. 89

[Bruno and Leroux, 1999] C. Bruno and P. Leroux. *La Création après la musique contemporaine*, chapter Musique contemporaine : une solution de continuité, pages 23–40. L'Itinéraire – L'Armattan, 1999. 265

[Buzan and Buzan, 2003] T. Buzan and B. Buzan. *Mind Map : dessine-moi l'intelligence*. d'Organisation, 2003. 98, 212, 267

[Cabrera, 1999a] D. Cabrera. PsySound : a Computer Program for Psychoacoustical Analysis. In *Proc. Australian Ac. Soc. Conf., Melbourne*, pages 47–53, novembre 1999. 114

[Cabrera, 1999b] D. Cabrera. "PsySound" : a Computer Program for the Psychoacoustical Analysis of Music. In *Proc. Australasian Computer Music Conference, MikroPolyphonie*, volume 5, Wellington, New Zealand, 1999. 114

[Cabrera, 2000] D. Cabrera. PsySound 2 : Psychoacoustical Software for Macintosh PPC. Technical report, 2000. 114, 133, 135, 136, 139, 140, 141

[Cadoz, 1999] C. Cadoz. *Les Nouveaux gestes de la musique*, chapter Musique, geste, technologie, pages 47–92. H. Genevois and R. de Vivo, ed. Parenthèses, 1999. 97, 98

[Camilo and Tomatito, 2000] M. Camilo and Tomatito. ♩ Spain, Lola Records, 2000. 270

[Cappé, 1993] O. Cappé. *Techniques de réduction de bruit pour la restauration d'enregistrements musicaux*. PhD thesis, Ecole Nationale Supérieure des Télécommunications, Paris, 1993. 94

[Chowning, 1971a] J. Chowning. The Simulation of Moving Sound Sources. *J. Audio Eng. Soc.*, 19(1) :1–6, 1971. 72, 73

[Chowning, 1971b] J. Chowning. The Synthesis of Complex Audio Spectra by Means of Frequency Modulation. *J. Audio Eng. Soc.*, 21 :526–34, 1971. 37

[CNRS éditions, 2002] CNRS éditions. Ce Chant venu des dunes. *Journal du CNRS*, 153–4 :12–3, 2002. 270

[CNRS, 2002] CNRS. ♩ ...entendu sous le sable, CNRS éditions, 2002. 270

[Coppens, 2001] Y. Coppens. Il était une fois l'homme. *Le Monde de l'éducation*, 294 :44–9, 2001. 213

[Corea and Kujala, 1985] C. Corea and S. Kujala. ♩ Voyage, ECM Records, 1985. 270

[Couturier, 2003] J.-M. Couturier. *Espaces Sonores – Actes de Recherches*, chapter Espaces interactifs visuels et sonores pour le contrôle des sons musicaux. CICM - Editions Musicales Transatlantiques, 2003. 259

[Cristo, 1982] A. Di Cristo. *Prolégomènes à l'étude de l'intonation*. Editions du CNRS, 1982. 22

[Cycling'74, 2003] Cycling'74. Max/msp, http ://www.cycling74.com/, 2003. 261

[Daniel and Weber, 1997] P. Daniel and R. Weber. Psychoacoustical roughness : implementation of an optimized model. *Acustica, acta acustica*, 83 :113–23, 1997. 141

[Dattoro, 1997] J. Dattoro. Effect design, part 2 : Delay-line modulation and chorus. *J. Audio Eng. Soc.*, pages 764–88, 1997. 84, 86

[Desain and Honing, 1996] P. Desain and H. Honing. Modeling continuous aspects of music performance : Vibrato and portamento. In *Proc. 4th Int. Music Perception and Cognition Conf., Montreal*, 1996. 133, 152

[Desainte-Catherine and Marchand, 1999] M. Desainte-Catherine and S. Marchand. Structured Additive Synthesis : Towards a Model of Sound Timbre and Electroacoustic Music Forms. *Proc. Int. Computer Music Conf. (ICMC'99), Beijing*, pages 260–3, 1999. 50, 123

[Dixon, 1996] S. Dixon. Multiphonic note identification. In *Proc. of the 19th Australasian Computer Science, Melbourne, Australia*, 1996. 112

[Dolson, 1986] M. Dolson. The phase vocoder : a tutorial. *Computer Music J.*, 1986. 47

[Doval and Rodet, 1991] B. Doval and X. Rodet. Fundamental frequency estimation using a new harmonic matching method. In *Proc. Int. Computer Music Conf. (ICMC'91), Montréal*, 1991. 112

[Doval, 1994] B. Doval. *Estimation de la fréquence fondamentale des signaux sonores.* PhD thesis, Laforia 94/02, 1994. 112

[Drame *et al.*, 1998] C. Drame, D. Wessel, and M. Wright. Removing the Time Axis from Spectral Model Analysis-Based Additive Synthesis : Neural Networks vs. Memory-Based Machine Learning. In *Proc. Int. Computer Music Conf. (ICMC'98), Ann Arbor*, 1998. 107, 109, 146, 228, 250

[Dubnov and Rodet, 1997] S. Dubnov and X. Rodet. Statistical modeling of sound aperiodicities. In *Proc. Int. Computer Music Conf. (ICMC'97), Thessaloniki*, 1997. 143

[Dubnov and Tishby, 1996] S. Dubnov and N. Tishby. Testing For Gaussianity and Non Linearity In The Sustained Portion Of Musical Sounds. In *Proc. Journées Informatique Musicale (JIM'96)*, 1996. 124, 143

[Dubnov and Tishby, 2002] S. Dubnov and N. Tishby. *Recherches et applications en informatique musicale*, chapter Testing for Gaussianity and Non Linearity in the Sustained Portion of Musical Sounds, pages 315–25. M. Chemillier and F. Pachet Eds., Hermès, 2002. 143

[Dufour, 1999] H. Dufour. *Les Nouveaux gestes de la musique*, chapter Prolégomènes à la simulation du geste instrumental, pages 10–7. éditions Parenthèses, 1999. 149

[Dutilleux and Zoelzer, 2002] P. Dutilleux and U. Zoelzer. *DAFX - Digital Audio Effects*, chapter Nonlinear Processing, pages 93–135. U. Zoelzer ed., John Wiley & Sons, 2002. 93, 162

[Dutilleux, 1991] P. Dutilleux. *Vers la machine à sculpter le son, modification en temps-réel des caractéristiques fréquentielles et temporelles des sons.* PhD thesis, University of Aix-Marseille II, 1991. 210

[Dutilleux, 1992] P. Dutilleux. ♩ Lalula, 1992. 270

[Earth Wind & Fire, 2001] Earth Wind & Fire. ♩ Live, 2001. 270

[Evangelista, 1997] G. Evangelista. *Music Signal Processing*, chapter Wavelet representations of musical signals, pages 127–57. Swets & Zeitlinger Publishers, 1997. 56

[Fairbansk *et al.*, 1954] G. Fairbansk, W. L. Everitt, and R. P. Jaeger. Method for time or frequency compression-expansion of speech. *IEEE Trans. on Audio and Electroacoustics*, AU-2 :7–12, 1954. 40

[Favreau, 2001] E. Favreau. Phase Vocoder Applications in GRM Tools Environment. In *Proc. of the COST-G6 Wkshp on Digital Audio Effects (DAFx-01), Limerick, Ireland*, 2001. 199, 241

[Fletcher and Rossing, 1998] N. H. Fletcher and T. D. Rossing. *The Physics of Musical Instruments.* 2nd ed. Springer-Verlag, Berlin and New York, 1998. 131

[Fraser and Fujinaga, 1999] A. Fraser and I. Fujinaga. Toward real-time recognition of acoustic musical instruments. In *Proc. Int. Computer Music Conf. (ICMC'99), Beijing*, pages 175–7. ICMA, 1999. 111

[Gerzon, 1976] M. A. Gerzon. Unitary (energy preserving) multichannel networks with feedback. *Electronic Letters V*, 12(11) :278–9, 1976. 72

[Geslin, 2002] Y. Geslin. Digital Sound and Music Transformation Environments : a Twenty-year Experiment at the "Groupe de Recherches Musicales". *J. New Music Research*, 2002. 1

[Gibiat and Jardin, 1992] V. Gibiat and P. Jardin. Suivi du fondamental de signaux musicaux à haute résolution temporelle. *Journal de Physique III*, 2, 1992. 112

[Grey, 1975] J. M. Grey. *An exploration of musical timbre.* PhD thesis, Stanford University, 1975. 20, 109

[Griffin and Lim, 1984] D. W. Griffin and J. S. Lim. Signal estimation from modified Short-Time Fourier Transform. *IEEE Trans. on Ac., Speech and Sig. Proc.*, 32(2) :236–43, 1984. 90

[Guillemain, 1994] P. Guillemain. *Analyse et modélisation de signaux sonores par des représentations temps-fréquences linéaires.* PhD thesis, University of Aix-Marseille II, 1994. 56

[Harris, 1978] F. Harris. On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform. *Proceeding of the IEEE*, 66(1) :51–8, 1978. 167

[Hartmann, 1978] W. M. Hartmann. Flanging and Phasers. *J. Audio Eng. Soc.*, 26 :439–43, 1978. 86

[Haykin, 1996] S. Haykin. *Adaptive Filter Theory.* Prentice Hall, Third Edition, 1996. 160, 192

[Herrera and Bonada, 1998] P. Herrera and J. Bonada. Vibrato extraction and parameterization in the Spectral Modeling Synthesis framework. In *Proc. of the COST-G6 Wkshp on Digital Audio Effects (DAFx-98), Barcelona, Spain*, 1998. 205

[Herrera *et al.*, 1999a] P. Herrera, X. Serra, and G. Peeters. A proposal for the description of audio in the context of MPEG-7. In *Proc. CBMI'99 European Wkshp on Content-Based Multimedia Indexing*, 1999. 105

[Herrera *et al.*, 1999b] P. Herrera, X. Serra, and G. Peeters. Audio Descriptors and Descriptor Schemes in the Context of MPEG-7. In *Proc. Int. Computer Music Conf. (ICMC'99), Beijing*, 1999. 105

[Hervé, 1999] J.-L. Hervé. *La Création après la musique contemporaine*, chapter Pourquoi écrire de la musique aujourd'hui ?, pages 41–50. L'Itinéraire – L'Armattan, 1999. 9

[Hunt *et al.*, 2000] A. Hunt, M. Wanderley, and R. Kirk. Towards a Model for Instrumental Mapping in Expert Musical Interaction. In *Proc. Int. Computer Music Conf. (ICMC'00), Berlin*, pages 209–12, 2000. 152

[Hutchinson and Knopoff, 1978] W. Hutchinson and L. Knopoff. The Acoustical component of western consonance. *Interface*, 1978. 141

[INA-GRM, 2003] INA-GRM. GRM Tools, http ://www.grmtools.org/, 2003. 156

[IRCAM, 2000] IRCAM. Studio On Line, http ://sol.ircam.fr, 2000. 111

[Iverson, 1995] P. Iverson. Auditory stream segregation by musical timbre : Effects of static and dynamic acoustic attributes. *J. Exp. Psych. : Human Perception and Performance*, 21 :751–63, 1995. 20

[Jehan and Schoner, 2001a] T. Jehan and B. Schoner. An Audio-Driven Perceptually Meaningful Timbre Synthesizer. In *Proc. Int. Computer Music Conf. (ICMC'01), Havana*, 2001. 109

[Jehan and Schoner, 2001b] T. Jehan and B. Schoner. An Audio-Driven, Spectral Analysis-Based, Perceptually Meaningful Timbre Synthesizer. In *AES 11Oth convention*, Amsterdam, Netherland, 2001. 109, 250

[Jehan, 1997] T. Jehan. Musical Signal Parameter Estimation. Master's thesis, University of Rennes I, France - CNMAT, Berkeley, USA, 1997. 128

[Jensen, 1999a] K. Jensen. Enveloppe Model of Isolated Musical Sounds. In *Proc. of the COST-G6 Wkshp on Digital Audio Effects (DAFx-99), Trondheim, Norway*, 1999. 107

[Jensen, 1999b] K. Jensen. *Timbre Models of Musical Sounds.* PhD thesis, University of Copenhagen, 1999. DIKU Report 99/7. 109

[Jensen, 2001] K. Jensen. The Timbre Model. In *Wkshp on Current Research Directions in Computer Music, Barcelona, Nov 15-16-17, 2001*, 2001. 124

[Jot and Chaigne, 1991] J.-M. Jot and A. Chaigne. Digital Delay Networks for Designing Artificial Reverberators. *Audio Eng. Soc. Convention*, 1991. 72

[Jot and Warusfel, 1995] J.-M. Jot and O. Warusfel. A Real-Time Spatial Sound Processor for Music and Virtual Reality Applications. In *Proc. Int. Computer Music Conf. (ICMC'95), Banff*, 1995. 19

[Jot, 1992] J.-M. Jot. *Etude et Réalisation d'un Spatialisateur de Sons par Modèles Physiques et Perceptifs*. 92 e 019, Télécom Paris, 1992. 72

[Jullien *et al.*, 1993] J.-P. Jullien, E. Kahle, M. Marin, O. Warusfel, G. Bloch, and J.-M. Jot. Spatializer : a perceptual approach. In *Proc. 94th AES Convention, Berlin, preprint 3465*, 1993. 4

[Kaiser, 1990] J. F. Kaiser. On a simple algorithm to calculate the 'energy' of a signal. *Proc. of the IEEE Int. Conf. Acou. Speech and Signal Proc. (ICASSP'90)*, pages 381–4, 1990. 116, 126

[Kaiser, 1993] J. F. Kaiser. Some Useful Properties of Teager's Energy Operators. *Proc. of the IEEE Int. Conf. Acou. Speech and Signal Proc. (ICASSP'93)*, pages 149–52, 1993. 116, 126

[Karplus and Strong, 1983] K. Karplus and A. Strong. Digital Synthesis of Plucked-String and Drum Timbres. *Computer Music J.*, 7(2) :43–55, 1983. 32

[Kedem, 1986] B. Kedem. Spectral analysis and discrimination by zero-crossings. In *Proc. of the IEEE*, pages 1477–93, 1986. 122

[Kelly and Lochbaum, 1962] J. L. Kelly and C. C. Lochbaum. Speech Synthesis. In *Proc. of the Fourth International Congress on Acoustics*, 1962. 32

[Kessous, 2003] L. Kessous. *Espaces Sonores – Actes de Recherches*, chapter Instruments bimanuels et espaces sonores. CICM - Editions Musicales Transatlantiques, 2003. 109

[Krimphoff, 1994] J. Krimphoff. *Analyse acoustique et perception du timbre*. PhD thesis, Université du Mans, France, 1994. 139

[Kronland-Martinet *et al.*, 1997] R. Kronland-Martinet, Ph. Guillemain, and S. Ystad. Modelling of Natural Sounds Using Time-Frequency and Wavelet Representations. *Organised Sound*, 2(3) :179–91, 1997. 32

[Kronland-Martinet, 1988] R. Kronland-Martinet. The use of the wavelet transform for the analysis, synthesis and processing of speech and music sounds. *Computer Music J.*, 12(4) :11–20, 1988. 56

[Krumhansl, 1989] C. L. Krumhansl. *Structure and perception of electroacoustic sound and music Amsterdam : Elsevier*, chapter Why is musical timbre so hard to understand ?, pages 43–53. S. Nielzenand and O. Olsson (eds.), 1989. 110, 132

[Laakso *et al.*, 1996] T. I. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine. Splitting the Unit Delay. In *IEEE Signal Processing Magazine*, pages 30–60, 1996. 31, 33, 70

[Lacasse, 2000] S. Lacasse. *'Listen to My Voice' : The Evocative Power of Vocal Staging in Recorded Rock Music and Other Forms of Vocal Expression*. PhD thesis, University of Liverpool, Canada, 2000. 2

[Lakatos, 2000] S. Lakatos. A common perceptual space for harmonic and percussive timbres. *Perception & Psychophysics*, 62 :1426–39, 2000. 110

[Landy, 1991] L. Landy. Sound Transformations in Electroacoustic Music, 1991. 24, 188

[Laroche and Dolson, 1997] J. Laroche and M. Dolson. About this Phasiness Business. In *Proc. Int. Computer Music Conf. (ICMC'97), Thessaloniki*, 1997. 47, 65

[Laroche and Dolson, 1999] J. Laroche and M. Dolson. New Phase Vocoder Technique for Real-Time Pitch-Shifting, Chorusing, Harmonizing and Other Exotic Audio Modifications. *J. Audio Eng. Soc.*, 47(11), 1999. 47, 65

[Laroche, 1995] J. Laroche. *Traitement des Signaux Audio-Fréquences*. département TSI, Sup'Télécom Paris, 1995. 57

[Laroche, 1998] J. Laroche. *Applications of Digital Signal Processing to Audio & Acoustics, In Mark Kahrs and Karlheinz Brandenburg, eds.*, chapter Time and Pitch Scale Modification of Audio Signals, pages 279–309. Kluwer Academic Publishers, 1998. 41

[Lee, 1972] F. F. Lee. Time compression and expansion of speech by the sampling method. *J. Audio Eng. Soc.*, pages 738–42, 1972. 41

[Leman, 2000] M. Leman. Visualization and calculation of the acoustical musical signals using the Synchronization Index Model (SIM). In *Proc. of the COST-G6 Wkshp on Digital Audio Effects (DAFx-00), Verona, Italy*, 2000. 141

[Légeret, 2003] K. Légeret. *Espaces Sonores – Actes de Recherches*, chapter Interface Danse/Musique : une Poétique de l'Energie Percussive, pages 91–104. CICM - Editions Musicales Transatlantiques, 2003. 256

[Lienard *et al.*, 1977] J.-S. Lienard, D. Teil, C. Choppy, and G. Renard. Diphone synthesis of French : Vocal response unit and automatic prosody from text. In *Proc. of the IEEE Int. Conf. Acou. Speech and Signal Proc. (ICASSP'77)*, pages 560–3, 1977. 202

[Lindsay *et al.*, 2003] A. T. Lindsay, A. P. Parkes, and R. Fitzgerald. Description-Driven Context-Sensitive Effects. In *Proc. of the Int. Conf. on Digital Audio Effects (DAFx-03), London, England*, 2003. 149

[Lu *et al.*, 2001] L. Lu, H. Jiang, and H. Zhang. A Robust Audio Classification and Segmentation Method. In *ACM Multimedia*, 2001. 114, 117, 122, 130

[Maher and Beauchamp, 1994] R. C. Maher and J. Beauchamp. Fundamental frequency estimation of musical signals using a two-way mismatch procedure. *J. Acoustic Soc. Am.*, 95(4) :2254–63, 1994. 49, 112, 251

[Maher, 1990] R. C. Maher. Evaluation of a method for separating digitized duet signals. *J. Acoustic Soc. Am.*, 38(12) :956–79, 1990. 112

[Makhoul and El-Jaroudi, 1986] J. Makhoul and A. El-Jaroudi. Time-scale modification in medium to low rate coding. In *Proc. of the IEEE Int. Conf. Acou. Speech and Signal Proc. (ICASSP)*, pages 1705–8, 1986. 42

[Makhoul, 1977] J. Makhoul. Linear prediction : a tutorial review. *IEEE Trans. on Ac., Speech and Sig. Proc.*, 1977. 53

[Malloch, 1997] S. Malloch. *Timbre and technology.* PhD thesis, University of Edinburgh, 1997. 140

[Mani and Nawab, 1995] R. Mani and S. H. Nawab. Integration of DSP algorithms and musical constraints for the separation of partials in polyphonic music. Technical report, Report for the ECE Department, Boston University, 1995. 112

[Markel and Gray, 1976] J. D. Markel and A. H. Gray. *Linear Prediction of Speech.* Springler-Verlag, 1976. 53

[Marthouret, 2003] A. Marthouret. *Espaces Sonores – Actes de Recherches*, chapter A propos de *Proposition 2*, pages 115–6. CICM - Editions Musicales Transatlantiques, 2003. 256

[Martin and Kim, 1998] K. Martin and Y. E. Kim. Musical instrument identification : A pattern-recognition approach, 1998. 111, 113

[Masri and Bateman, 1996] P. Masri and A. Bateman. Improved Modelling of Attack Transients in Music Analysis-Resynthesis. In *Proc. Int. Computer Music Conf. (ICMC'96), Hong Kong*, pages 100–3, 1996. 128

[Mathews and Moore, 1970] M. V. Mathews and F. R. Moore. GROOVE - a program to compose, store, and edit functions of time. *Communications of the ACM (CACM)*, 13(12), 1970. 3

[Mathews and Pierce, 1989] M. V. Mathews and J. Pierce. *Current Directions in Computer Music.* MIT Press, 1989. 57

[Mathews, 1969] M. V. Mathews. *The Technology of Computer Music.* M.I.T. Press, Cambridge, 1969. 1

[Mathworks, 2003] Mathworks. Matlab, http ://www.mathworks.com/, 2003. 251

[McAdams and Bigand, 1994] S. McAdams and E. Bigand. Penser les sons. In *Psychologie cognitive de l'audition.* P.U.F., 1994. 106

[McAdams and Cunibile, 1992] S. McAdams and J. C. Cunibile. Perception of timbral analogies. In *Philosophical Trans. of the Royal Soc.*, volume series B 336, pages 383–89, London, 1992. 20, 109, 110

[McAdams *et al.*, 1995] S. McAdams, S. Winsberg, G. de Soete, and J. Krimphoff. Perceptual scaling of synthesized musical timbres : common dimensions, specificities, and latent subject classes. *Psychological Research*, 58 :177–92, 1995. 107, 139

[McAulay and Quatieri, 1986] R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. on Ac., Speech and Sig. Proc.*, 34(4), 1986. 47, 48, 251

[Meunier and Canévet, 2002] S. Meunier and G. Canévet. Psychoacoustique musicale et psychoacoustique appliquée. Technical report, D.E.A Acoustique, Module perception auditive et musique, 2002. 131

[Meyer, 1995] P. Meyer. *Les Progrès du progrès.* Points, 1995. 103

[Misdariis *et al.*, 2001] N. Misdariis, F. Nicolas, O. Warusfel, and R. Caussé. Radiation control on multi-loudspeaker device : La Timée. In *Proc. Int. Computer Music Conf. (ICMC'01), Havana*, 2001. 79

[Moore and Glasberg, 1996] B. C. J. Moore and B. R. Glasberg. A Revision of Zwickers loudness model. *Acustica, acta acustica*, 82 :3335–45, 1996. 18, 133

[Moorer, 1979] J. A. Moorer. About this reverberation business. *Computer Music J.*, 3(2) :13–8, 1979. 72

[Moulines and Charpentier, 1990] E. Moulines and F. Charpentier. Pitch Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones. *Speech Communication*, 9(5/6) :453–67, 1990. 42

[Métois, 1996] E. Métois. *Musical Sound Information : Musical gesture and Embedding synthesis.* PhD thesis, Massachusetts Institute of Technology, 1996. 109, 150

[Music Tech. Group, 2000] Music Tech. Group. SMS, http ://www.iua.upf.es/sms/, 2000. 50

[Music Tech. Group, 2002] Music Tech. Group. CLAM, http ://www.iua.upf.es/mtg/clam/, 2002. 50

[Native Instrument, 2002] Native Instrument. Spektral Delay, http ://www.nativeinstruments.de/, 2002. 156

[Noll, 1964] A. M. Noll. Short-time Spectrum and "Cepstrum" Techniques for Vocal Pitch Detection. *J. Acoust. Soc. Am.*, 36(2) :296–302, 1964. 54

[Noll, 1967] A. M. Noll. Cepstrum pitch determination. *J. Acoust. Soc. Am.*, 41 :293–309, 1967. 56

[Orfanidis, 1996] S. Orfanidis. *Introduction to Signal Processing.* Prentice Hall International Editions, 1996. 26, 57, 84, 86

[Orio, 1999] N. Orio. The timbre space of the classical guitar and its relationship with the plucking techniques. In *Proc. Int. Computer Music Conf. (ICMC'99), Beijing*, 1999. 132

[Pallone *et al.*, 1999] G. Pallone, P. Boussard, L. Daudet, P. Guillemain, and R. Kronland-Martinet. A wavelet based method for audio-video synchronization in broadcasting applications. In *Proc. of the COST-G6 Wkshp on Digital Audio Effects (DAFx-99), Trondheim, Norway*, 1999. 65

[Pallone, 2003] G. Pallone. *Dilatation et transposition sous contraintes perceptives des signaux audio : application au transfert cinéma-vidéo*. PhD thesis, University of Aix-Marseille III, 2003. 178

[Parncutt, 1989] R. Parncutt. *Harmony : a psychoacoustical approach*. Springler, Berlin, 1989. 135, 136

[Peeters *et al.*, 2000] G. Peeters, S. McAdams, and P. Herrera. Instrument sound description in the context of MPEG-7. In *Proc. Int. Computer Music Conf. (ICMC'00), Berlin*, 2000. 110

[Platon, 1997] Platon. *Apologie de Socrate*. L. Brisson, Flammarion, 1997. 149

[Pollard and Jansson, 1982] H. Pollard and E. Jansson. A tristimulus method for the specification of musical timbre. *Acustica, acta acustica*, 51 :162–71, 1982. 140

[Portnoff, 1976] M. R. Portnoff. Implementation of the Digital Phase Vocoder Using the Fast Fourier Transform. *IEEE Trans. on Ac., Speech and Sig. Proc.*, 24(3) :243–8, 1976. 44

[Pressnitzer and McAdams, 1999] D. Pressnitzer and S. McAdams. Acoustics, psychoacoustics and spectral music. *Contemporary Music Review*, 19(2) :33–60, 1999. 141

[Pressnitzer, 1998] D. Pressnitzer. *Perception de la rugosité psychoacoustique : d'un attribut élémentaire de l'audition à l'écoute musicale*. PhD thesis, University of Paris VI, 1998. 20

[Renard, 2003] J. Renard. *Journal 1893–1898*. ABU, http ://abu.cnam.fr, 2003. 103

[Risset and Wessel, 1999] J.-C. Risset and D. L. Wessel. *Exploration of timbre by analysis and synthesis*, pages 113–69. D. Deutsch, Academic Press, New York, 1999. 20, 106

[Risset, 1971] J.-C. Risset. Paradoxes de hauteur : le concept de hauteur sonore n'est pas le même pour tout le monde. In *Proc. 7^{th} Int. Congres of Acoustics, Budapest*, 1971. 21

[Risset, 1986] J.-C. Risset. *Les Instruments de l'Orchestre*, chapter Son musical et perception auditive, pages 149–65. Pour la Science, 1986. 1

[Risset, 2002] J.-C. Risset. Examples of the Musical Use of Digital Audio Effects. *J. New Music Research*, 2002. 2

[Roads, 1998] C. Roads. *L'audionumérique*. Dunod, Paris, 1998. 29, 57

[Roads, 1999] C. Roads. *Synthèse et transformation des microsons*. PhD thesis, University of Paris VIII, France, 1999. 28, 96

[Roads, 2002] C. Roads. *Microsound*. MIT Press, 2002. 28, 96

[Rochebois, 1997] T. Rochebois. *Méthodes d'analyse/synthèse et représentations optimales des sons musicaux basées sur la réduction de données spectrales*. PhD thesis, University of Paris XI, France, 1997. 146, 250

[Rochesso, 2002] D. Rochesso. *DAFX - Digital Audio Effects*, chapter 6 - Spatial Effects, pages 137–200. U. Zoelzer ed., John Wiley & Sons, 2002. 75, 77, 78, 189

[Rossignol *et al.*, 1998a] S. Rossignol, P. Depalle, J. Soumagne, X. Rodet, and J.-L. Collette. Vibrato : Detection, Estimation, Extraction, Modification. In *Proc. of the COST-G6 Wkshp on Digital Audio Effects (DAFx-98), Barcelona, Spain*, 1998. 128, 133, 205

[Rossignol *et al.*, 1998b] S. Rossignol, X. Rodet, J. Soumagne, J.-L. Collette, and P. Depalle. Feature Extraction and Temporal Segmentation of Acoustic Signals. In *Proc. Int. Computer Music Conf. (ICMC'98), Ann Arbor*, 1998. 111, 204

[Rossignol, 2000] S. Rossignol. *Segmentation et indexation des signaux sonores musicaux*. PhD thesis, IRCAM Paris, Supélec Metz, 2000. 114, 124, 129

[Roucos and Wilgus, 1985] S. Roucos and A. M. Wilgus. High Quality Time-Scale Modification for Speech. In *Proc. of the IEEE Int. Conf. Acou. Speech and Signal Proc. (ICASSP)*, pages 493–6, 1985. 42

[Schaeffer and Reibel, 1998] P. Schaeffer and G. Reibel. ♩ Solfège des Objets Sonores, Seuil/INA-GRM, 1998. 270

[Schaeffer, 1966] P. Schaeffer. *Le Traité des Objets Musicaux*. Seuil, Paris, 1966. 269

[Scheirer and Slaney, 1997] E. Scheirer and M. Slaney. Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator. In *Proc. ICASSP '97*, pages 1331–4, Munich, Germany, 1997. 111, 114, 117, 126

[Schrœder and Logan, 1961] M. R. Schrœder and B. Logan. "colorless" artificial reverberation. *J. Audio Eng. Soc.*, 9 :192–7, 1961. 72

[Schrœder, 1968] M. R. Schrœder. Period histogram and product spectrum : new methods for fundamental-frequency measurements. *J. Acoustic Soc. Am.*, 1968. 112

[Scutenaire, 1984] L. Scutenaire. *Mes Inscriptions (1945-1963)*. Allia, 1984. 7, 57

[Sédès *et al.*, 2003] A. Sédès, B. Courribet, J.-B. Thiébaut, and V. Verfaille. *Espaces Sonores – Actes de Recherches*, chapter Visualisation de l'Espace Sonore, vers la Notion de Transduction : une Approche Intéractive Temps-Réel, pages 125–43. CICM - Editions Musicales Transatlantiques, 2003. 261

[Sédès, 2003] A. Sédès. *Espaces Sonores – Actes de Recherches*, chapter Espaces Sonores, Espaces Sensibles, pages 105–14. CICM - Editions Musicales Transatlantiques, 2003. 256

[Serra and Smith, 1990] X. Serra and J. O. Smith. A Sound Decomposition System Based on a Deterministic plus Residual Model. *J. Acoustic Soc. Am, Supp. 1*, 89(1) :425–34, 1990. 45, 47, 48, 50, 123

[Serra, 1996] X. Serra. *Musical Signal Processing*, chapter Musical Sound Modeling with Sinusoids plus Noise. G. D. Poli and A. Picialli and S. T. Pope and C. Roads Eds. Swets & Zeitlinger, 1996. 47, 111, 119, 123, 125, 130, 131

[Sethares, 1993] W. Sethares. Local consonance and the relationship between timbre and scale. *J. Acoustic Soc. Am.*, 93(3) :1218–28, 1993. 141

[Siron, 1992] J. Siron. *La Partition Intérieure. Jazz, Musiques Improvisées*. Outre Mesure, 1992. 14

[Slawson, 1985] W. Slawson. *Sound Color*. Berkeley : University of California Press, 1985. 109, 201

[Smith *et al.*, 2002] J.O. Smith, S. Serafin, J. Abel, and D. Berners. Doppler Simulation and the Leslie. In *Proc. of the Int. Conf. on Digital Audio Effects (DAFx-02), Hamburg, Germany*, 2002. 74

[Smith, 1984] J. O. Smith. An allpass approach to digital phasing and flanging. In *Proc. Int. Computer Music Conf. (ICMC'84), Paris*, 1984. 72, 86

[Smith, 1987] J. O. Smith. Waveguide Filter Tutorial. In *Proc. Int. Computer Music Conf. (ICMC'87), Champaign-Urbana*, pages 9–16, 1987. 32, 72

[Stautner and Puckette, 1982] J. Stautner and M. Puckette. Designing multichannel reverberators. *Computer Music J.*, 6(1), 1982. 72

[Strawn, 1985] J. Strawn. *Modeling Musical Transitions*. PhD thesis, Center for Computer Research in Music and Acoustics (CCRMA), 1985. 106

[Strawn, 1987] J. Strawn. Analysis and Synthesis of Musical Transitions Using the Discrete Short-Time Fourier Transform. *J. Audio Eng. Soc.*, 35(1) :3–13, 1987. 106, 152

[Sussman and Khars, 1996] R. B. Sussman and M. Khars. Analysis of Musical Instrument Sounds Using Discrete Energy Separation Algorithms. In *Preprint of the AES Convention*, 1996. 126

[TC-Helicon, 2002] TC-Helicon. Voice one, voice prisme, http ://www.tc-helicon.tc/, 2002. 69, 156

[Terhardt, 1979] E. Terhardt. Calculating virtual pitch. *Hearing Research*, 1 :155–82, 1979. 114, 123, 136

[The Sound Guys, 2003] The Sound Guys. Sfx machine RT, http ://www.sfxmachine.com/, 2003. 156, 203

[Todoroff, 2003] T. Todoroff. *Espaces Sonores – Actes de Recherches*, chapter Installations Sonores Intéractives, pages 75–90. CICM - Editions Musicales Transatlantiques, 2003. 256

[Tzanetakis and Cook, 1999] G. Tzanetakis and P. Cook. Multifeature audio segmentation for browsing and annotation. In *Wkshp on Applications of Sig.Proc. to Audio & Acoustics (WASPAA99), New York*, 1999. 111

[Varèse, 1958] E. Varèse. ♩ Poème Electronique, Ricordi-Studio Utrech, 1958. 269

[Vega, 1993] S. Vega. ♩ Tom's Diner in Tom's Album, A & M Records, 1993. 270

[Verfaille and Arfib, 2001] V. Verfaille and D. Arfib. ADAFx : Adaptive Digital Audio Effects. In *Proc. of the COST-G6 Wkshp on Digital Audio Effects (DAFx-01), Limerick, Ireland*, 2001. 149, 156, 178, 201

[Verfaille and Arfib, 2002] V. Verfaille and D. Arfib. Implementation Strategies for Adaptive Digital Audio Effects. In *Proc. of the Int. Conf. on Digital Audio Effects (DAFx-02), Hamburg, Germany*, 2002. 156, 235

[Verfaille, 2002] V. Verfaille. Réalisation d'effets audionumériques adaptatifs en temps réel et hors temps réel. In *Journées d'Informatique Musicale, 9e édition, 29-31 mai 2002*, 2002. 149, 156

[Verfaille, 2003] V. Verfaille. *Espaces Sonores – Actes de Recherches*, chapter Utilisation d'espaces perceptifs pour la synthèse et la transformation, pages 39–54. CICM - Editions Musicales Transatlantiques, 2003. 156, 256, 259

[Verma *et al.*, 1997] T. Verma, S. Levine, and T. Meng. Transient Modeling Synthesis : a flexible analysis/synthesis tool for transient signals. In *Proc. Int. Computer Music Conf. (ICMC'97), Thessaloniki*, 1997. 47, 50, 123, 131, 178

[von Aures, 1985] W. von Aures. Der sensorische Wohlklang als Funktion psychoakustischer Empfindungsgröfsen. *Acustica*, 58 :282–90, 1985. 139

[von Helmholtz, 1954] H. L. F. von Helmholtz. On the Sensations of Tone as a Physiological Basis for the Theory of Music. *Dover Publications*, 1954. 104

[Walmsley *et al.*, 1999] P. Walmsley, S. Godsill, and P. Rayner. Bayesian modelling of harmonic signals for polyphonic music tracking. In *Proc. Cambridge Music Processing Colloquim*, 1999. 112

[Wanderley and Depalle, 1999] M. Wanderley and P. Depalle. *Interfaces homme - machine et création musicale*, chapter Contrôle Gestuel de la Synthèse Sonore, pages 145–63. H. Vinet and F. Delalande, Paris : Hermès Science Publishing, 1999. 3, 99, 152

[Wanderley *et al.*, 1998] M. Wanderley, N. Schnell, and J. B. Rovan. Escher - Modeling and Performing Composed Instruments in Real-Time. In *Proc. of the 1998 IEEE Int. Conf. on Systems, Man and Cybernetics (SMC'98), San Diego, CA USA*, pages 1080–4, 1998. 152

[Wanderley, 2001] M. Wanderley. *Intéraction Musicien-Instrument : application au contrôle gestuel de la synthèse sonore*. PhD thesis, Université Paris VI, IRCAM, 2001. 99

[Warusfel and Misdariis, 2001] O. Warusfel and N. Misdariis. Directivity synthesis with a 3D array of loudspeakers - Application for stage performance. In *Proc. of the COST-G6 Wkshp on Digital Audio Effects (DAFx-01), Limerick, Ireland*, 2001. 79

[Wessel, 1979] D. Wessel. Timbre space as a musical control structure. *Computer Music J.*, 3(2) :45–52, 1979. 109

[Wold *et al.*, 1996] E. Wold, T. Blum, and D. Keislar. Content-based classification, search and retrieval of audio. *IEEE Multimedia*, 3(3) :27–36, 1996. 111

[Wright and Scheirer, 1999] M. Wright and E. D. Scheirer. Title Cross-Coding SDIF into MPEG-4 Structured Audio. In *Proc. Int. Computer Music Conf. (ICMC'99), Beijing*, 1999. 109

[Wright *et al.*, 1999] M. Wright, R. Dudas, S. Khoury, R. Wang, and D. Zicarelli. Supporting the Sound Description Interchange Format in the Max/MSP Environment. In *Proc. Int. Computer Music Conf. (ICMC'99), Beijing*, 1999. 109

[Wyse, 1997] L. Wyse. Flexible Sound Effects. In *Proc. Int. Computer Music Conf. (ICMC'97), Thessaloniki*, 1997. 2

[Zero G, 2000a] Zero G. ♩ Creative Essentials, Vol. 14 - Live Bass Grooves, 2000. 270

[Zero G, 2000b] Zero G. ♩ Creative Essentials, Vol. 24 - Funk Construction, 2000. 269

[Zoelzer, 1997] U. Zoelzer. *Digital Audio Signal Processing*. John Wiley & Sons, 1997. 57, 162, 236

[Zoelzer, 2002] U. Zoelzer, editor. *DAFX - Digital Audio Effects*. John Wiley & Sons, 2002. 26, 33, 57, 97, 114

[Zwicker and Fastl, 1999] E. Zwicker and H. Fastl. *Psychoacoustics : Facts and Models*. Springler, Berlin, 1999. 133, 139, 168

[Zwicker, 1977] E. Zwicker. Procedure for Calculating Loudness of Temporally Variable Sounds. *J. Acoust. Soc. Am.*, 62(3) :675–82, 1977. 114