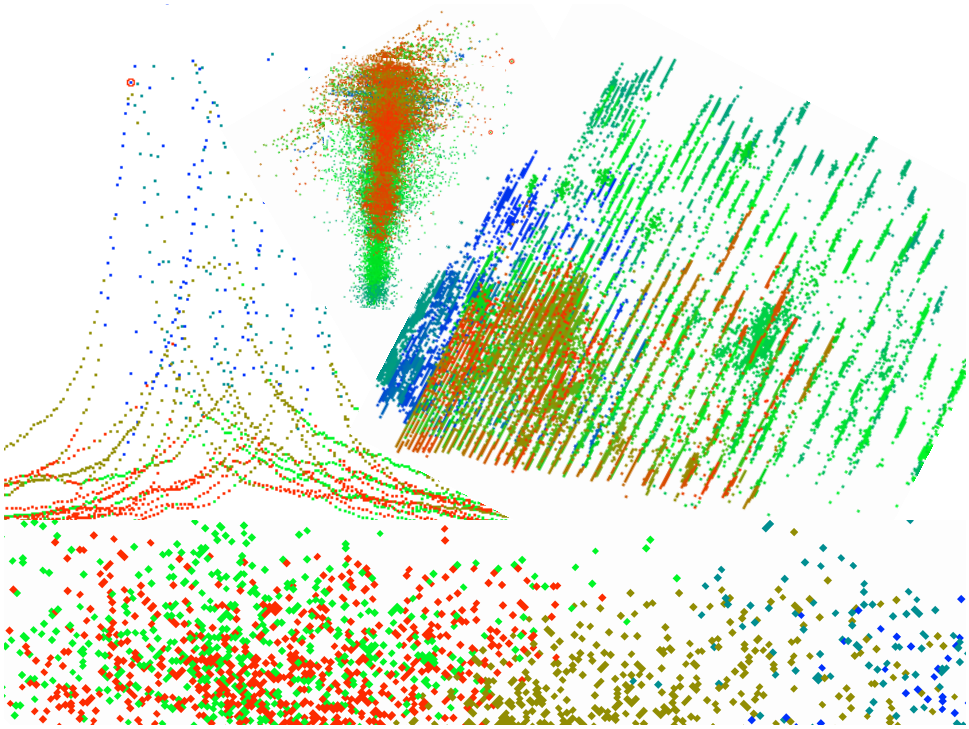
Stratégies de visualisation et de navigation pour le contrôle gestuel de la synthèse par corpus sonore

Rapport de Stage de Master 2 - ATIAM
Sous la direction de Diemo Schwarz



Bruno Verbrugghe

IRCAM - Université Paris 6
Equipe Applications Temps-Réel

Septembre 2006

Table des matières

Table des figures

Remerciements

Je voudrais remercier Norbert Schnell de m'avoir accueilli dans son équipe.

Je remercie vivement Diemo Schwarz, qui m'a encadré tout au long de ce stage, pour le temps qu'il a bien voulu me consacrer et les longues discussions pendant lesquelles parfois notre imagination sortait du cadre strict de ce stage.

Pour les petits et grands moments qu'ils m'ont accordés, je remercie Roland Cahen, Jean-Philipe Lambert, Rémy Muller, Gregory Beller, Frédéric Bevilacqua, Geoffroy Peeters et Nicolas Rasamimanana.

Merci à Julien Bloit pour son accueil et son soutien.

Merci aux Bœufeurs Fous d'Odéon pour les nuits de transe salvatrices à la cave, à tout le Balbazar de PavéJazz pour ces concerts pendant lesquels la folie était reine, et à Dunga pour la tournée et les bouts d'Afrique.

Tu me dis, j'oublie.
Tu m'enseignes, je me souviens.
Tu m'impliques, j'apprends.

Benjamin Franklin.

Chapitre 1

Introduction

Le sujet de ce stage traite de lutherie électronique et s'inscrit dans le cadre des recherches menées à l'Ircam sur la synthèse sonore dite concaténative, basée sur l'utilisation de bases de données sonores.

Dans les instruments de musique acoustiques, le processus de génération du son est indissociable du contrôle gestuel opéré par le musicien sur l'instrument lui même. Par contre, dans un instrument de musique électronique le moteur de synthèse sonore n'est pas lié aux caractéristiques intrinsèques du contrôleur (contrôleur MIDI, capteurs, ...), c'est-à-dire qu'il n'existe pas de *mapping* implicite entre l'un et l'autre. Nous définissons ici le terme *mapping* comme la manière dont sont liés des paramètres de contrôle provenant de l'artiste avec des paramètres de synthèse sonore. Dans le cas des instruments de musique acoustiques, les possibilités d'interaction sont très grandes grâce aux relations complexes et non linéaires des paramètres d'entrée entre eux. A contrario, le mapping adopté dans le cas d'un instrument de musique électronique fait bien souvent correspondre un seul paramètre du contrôleur à un seul paramètre du moteur de synthèse, ce qui réduit les possibilités d'expressivité et ne permet pas un contrôle total du moteur de synthèse. Ainsi il apparaît très important d'étudier les différentes stratégies de mapping lors de l'élaboration d'un instrument de musique électronique.

L'idée première de ce stage était de définir un modèle général de mapping pour la synthèse concaténative pouvant mettre en correspondance n'importe quel contrôleur existant ou non avec n'importe quelle base de données sonores. En effet, il n'existe à l'heure actuelle aucun dispositif de contrôle gestuel temps-réel utilisant pleinement les capacités de la synthèse concaténative. Nous montrerons par la suite qu'au cours de ce stage nous avons restreint notre étude à une petite partie seulement de ce que serait un modèle général de mapping pour la synthèse concaténative, ce sujet s'étant avéré beaucoup trop vaste pour un stage.

D'après [PD04] le sujet du contrôle gestuel d'un instrument de musique électronique est divisé en quatre parties :

- définition et typologie du geste
- acquisition et captation des paramètres gestuels
- algorithme de synthèse
- stratégies de mapping entre les variables gestuelles et les variables de synthèse

Dans la première partie de ce rapport (*cf* chapitre 2 page 13) nous nous appuyons sur cette segmentation afin de présenter un état de l'art des théories et travaux antérieurs.

Il y est tout d'abord défini une taxonomie du geste, et plus particulièrement celle du geste musical (*cf* section 2.1). L'étude en détail des systèmes de captation des gestes a été volontairement occultée dans l'idée d'élaborer un modèle générique pouvant s'adapter à tous types de contrôleur. Ensuite les caractéristiques de fonctionnement de la synthèse concaténative sont étudiées afin d'en déterminer les principaux paramètres de contrôle (*cf* section 2.2). Enfin différentes stratégies de mapping sont présentées (*cf* section 2.3).

À la suite de cette étude bibliographique, nous avons retenu le modèle général de mapping proposé par D. Arfib [ACKV02] (voir figure 1.1). Celui-ci comporte trois couches :

- des données gestuelles à un espace perceptuel des gestes (couche relative à *l'interprétation*),
- d'un espace perceptuel des sons aux paramètres du moteur de synthèse (couche relative à *la définition de l'instrument*),
- le lien entre ces deux espaces perceptuels.

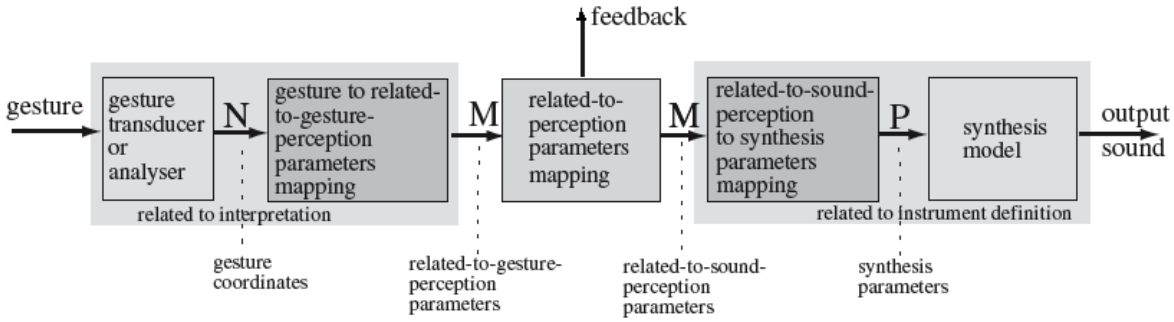


FIG. 1.1 – Modèle de mapping multicouche du geste au son. Le geste est traduit en données physiques par le capteur; les données gestuelles sont ensuite transformées en données *relatives à la perception gestuelle*. Puis un deuxième mapping transforme ces données en données *relatives à la perception du son*. Enfin, un mapping transforme les données *relatives à la perception du son* en données pour le modèle de synthèse. M, N et P représentent le nombre de paramètres.

Dans le cadre de ce stage nous nous sommes concentrés sur la couche relative à la définition de l'instrument, laissant de côté l'étude de la couche relative à l'interprétation des paramètres gestuels. Le mapping retenu pour la couche intermédiaire est un mapping très simple car nous avons souhaité nous contraindre à un environnement de test ne comportant que deux dimensions de contrôle. Le retour instrumental est constitué du résultat sonore de la synthèse, du retour physique du contrôleur lui-même —une tablette graphique—, ainsi que du retour physique d'un écran permettant de visualiser jusqu'à trois paramètres simultanément.

Problématique

Dans le cas de la synthèse concaténative, la difficulté d'élaborer une stratégie de mapping tient tout d'abord aux caractéristiques propres du moteur de synthèse (*cf* section 2.2). La synthèse concaténative fonctionne par concaténation d'unités sonores constituant une base de données, l'accès aux unités dans la base se faisant grâce aux descripteurs

sonores caractérisant chaque unité. Le paramètre principal de contrôle est donc celui du choix des unités concaténées.

La problématique de recherche de ce stage est de trouver le moyen de naviguer et de visualiser une base de données de sons, que l'on a choisi de voir comme un espace de sons. L'agencement de cet espace multidimensionnel est défini par la valeur des descripteurs de chaque unité sonore. Cette navigation n'est pas triviale si l'on souhaite qu'elle soit perceptivement compréhensible car d'une part il existe de très nombreux descripteurs sonores et d'autre part, dans le cas d'une base de données contenant des sons très hétérogènes, certains descripteurs n'ont aucun sens selon la nature de l'unité (les descripteurs harmoniques pour des sons environnementaux par exemple). Dans le cadre de cette étude nous avons choisi de considérer une base de données de sons environnementaux et seulement six descripteurs ont été utilisés. Ces descripteurs ont été choisis de façon heuristique comme étant les plus génériques et perceptifs par rapport à la nature de la base.

Afin de proposer une navigation et une visualisation intuitives, nous avons cherché à réduire la complexité de l'espace des sons tout en gardant ses caractéristiques, c'est-à-dire en trouvant un moyen de représenter les différences entre les unités sonores. Il sera montré que l'utilisation de la méthode d'*Analyse par Composantes Principales* permet de réduire l'espace des sons à un espace à deux dimensions, qui plus est intuitif. Puis nous étudierons le comportement d'une méthode statistique de classification, l'algorithme (*Kmeans clustering*, *cf* chapitre 5), afin de déterminer des sous-ensembles d'unités sonores au sein de la base. Pour valider les résultats de ces deux méthodes, nous considérerons des sous-ensembles de sons de plus en plus complexes au cours de nos expériences. Enfin, à partir des résultats obtenus, nous étudierons différentes stratégies de navigation et de visualisation permettant d'exploiter au mieux la base de données sonores. Nous montrerons ainsi que l'utilisation de méthodes d'exploration de données permet un accès aux données de façon intuitive, sans connaissances préalables sur les descripteurs sonores, et facilite l'élaboration de stratégies de mapping simples et efficaces entre le moteur de synthèse sonore concaténative et le contrôleur gestuel considéré.

Le premier chapitre de ce rapport est constitué de cette longue introduction. Le deuxième chapitre présente l'état de l'art des domaines de recherche considérés : Taxonomie du geste musical, caractéristiques de fonctionnement de la synthèse concaténative et stratégies de mapping. Le troisième chapitre présente les contraintes et problématiques du stage, et expose les résultats des études préliminaires. Les quatrième et cinquième chapitres présentent les algorithmes statistiques d'exploration de données utilisés : l'Analyse en Composantes Principales et la classification *Kmeans*, et expose les résultats obtenus. Le sixième chapitre montre les stratégies de navigation et de visualisation déduites de l'utilisation des deux algorithmes étudiés. Enfin le dernier chapitre expose les perspectives et les travaux futurs.

Chapitre 2

Etat de l'art

Ce premier chapitre expose les théories et travaux antérieurs sur lesquels s'appuie le travail effectué au cours de ce stage. Son articulation repose sur la sub-division proposée par M. Wanderley et Ph. Depalle dans [PD04] pour définir le sujet du contrôle gestuel d'un instrument électronique de musique (cf chapitre 1). La première partie présente une taxonomie du geste, en se focalisant surtout sur le geste musical. La seconde partie montre le fonctionnement du moteur de synthèse considéré, la synthèse concaténative, en introduisant la notion d'espace de sons. Enfin la troisième partie s'intéresse aux différentes stratégies de mapping qui permettent d'expliciter le lien entre un geste et un moteur de synthèse.

2.1 Taxonomie du geste

Afin d'élaborer un modèle de contrôle gestuel de la synthèse concaténative, il est important de caractériser les *gestes de l'instrumentiste*, c'est-à-dire les actions physiques effectuées par le musicien en situation de jeu instrumental traditionnel.

2.1.1 Catégorisation du geste

Dans le but de considérer tout type de contrôleur nous avons cherché dans la littérature différentes classifications du geste, en particulier les gestes d'interaction dans un contexte musical. Dans [CW00] et [WD99], on trouvera une étude quasi-exhaustive des gestes musicaux. Nous avons retenu trois classifications significatives du geste :

Gestes primitifs : D'après Insook Choi [Cho00][Cho03], les gestes primitifs sont « les mouvements fondamentaux d'un sujet humain comme réponses dynamiques à un environnement ». Chaque primitive peut incorporer d'autres mouvements, tant qu'elle caractérise l'intention première de l'exécutant. Il existe trois types de gestes primitifs :

- basés sur des trajectoires (exemple : changements d'orientation),
- basés sur la force (exemple : gradient de mouvements),
- basés sur des motifs (exemple : mouvements quasi-periodiques),

Classification des gestes en trois classes proposée par François Delalande [Del88] :

- Geste effecteur : regroupe l'ensemble des mouvements effectués pour produire mécaniquement un son,
- Geste accompagnateur : rend compte des mouvements qui engagent le corps dans son entier : mimiques, gestes des épaules, ...,

- Geste figuratif : expressions purement symboliques perçues par l'auditeur comme des articulations dans une mélodie : par exemple, un mouvement du corps vers le bas pour signifier la fin d'une phrase mélodique. Canazza et al. ont effectué une étude du geste figuratif sur l'expressivité du jeu de la clarinette [CPV96, CPRV96].

Nous nous intéressons particulièrement dans notre étude au geste effecteur, que Claude Cadoz définit comme le *geste instrumental*.

Typologie du geste instrumental : Dans son étude du geste comme canal de communication [CW00], Claude Cadoz décrit les trois fonctions du geste comme la fonction *épistémique* (toucher pour acquérir de l'information), la fonction *ergotique* (transformer les objets par action physique) et la fonction *sémiotique* (communiquer de l'information). A partir de ces trois fonctions, il sub-divise le geste instrumental en trois classes :

- geste d'excitation : percussif, continu ou entretenu
- geste de modification : structurel ou paramétrique
- geste de sélection : séquentiel ou simultané

A ces trois classes M. Wanderley en ajoute une quatrième regroupant les gestes consistant à assurer des conditions normales de fonctionnement à l'instrument [CW00]. Dans le cas d'une cornemuse par exemple, le geste du bras qui assure un niveau de pression suffisant pour le jeu entre dans cette quatrième catégorie. Ce type de geste est appelé geste de *polarisation* ou de *maintien* [WD99]. Il se distingue des trois catégories ci-dessus dans le sens où il constitue un préalable essentiel à leur existence et à leur signification. Dans un instrument complexe, une phrase mélodique est une combinaison des différents gestes instrumentaux.

2.1.2 Interaction homme-machine

Quel que soit le contrôleur utilisé, le contrôle gestuel de la synthèse concaténative s'inscrit dans cette discipline très large qu'est l'*interaction homme-machine* (*IHM*). Bien que notre contexte soit avant tout musical, il nous a semblé aussi intéressant de considérer le geste sous l'angle de l'IHM [BL][GKP04][CV04] [LVV$^+$03a][WS02][Bon00]. On remarquera notamment certaines propriétés comme la loi de Fitts, qui est l'une des très rares lois quantitatives que fournit la psychologie pour la réalisation de systèmes interactifs. Elle montre qu'un temps de pointage typique est compris entre 1/2 et 1 seconde et que la relation entre le temps t (en seconde) de pointage d'une cible de taille L à une distance D est proportionnelle au logarithme de $2D/L$. D'autre part il est prouvé l'utilité et le gain de temps des équivalents-clavier et accélérateurs divers dans le cadre d'une utilisation experte. Enfin on notera l'intérêt que présente une interaction bimanuelle, tant du point de vue ergonomique qu'en terme de performance, surtout dans le cas d'interactions bimanuelles non-symétriques [CA03, KA03].

2.1.3 Retour instrumental

Dans [WD99], le *retour instrumental* est défini comme les actions physiques renvoyées par l'instrument sur son utilisateur. Il existe deux types de *retour instrumental* : *la rétroaction primaire*, liée essentiellement au contrôleur gestuel, et la *rétroaction secondaire* liée à la production sonore. La *rétroaction primaire* est constituée du retour *visuel*, *tactilo-kinesthésique* et *auditif primaire*, ce dernier étant le bruit résultant du fonctionnement de l'instrument. Dans le cas d'un instrument à vent, il s'agit du bruit des clés. Quant à la *rétroaction secondaire*, elle est constituée par le retour *du signal sonore de l'instrument*

aux oreilles du musicien. Notons que le retour tactilo-kinesthésique (parfois appelé tactilo-propriokinesthésique) revêt d'une grande importance dans le cas du jeu expert (utilisation experte d'un instrument, à opposer à une utilisation amateur n'exploitant pas toutes les subtilités offertes par l'instrument).

La rétroaction primaire du système considéré dans notre étude sera composée par le retour physique du contrôleur -une tablette graphique et son stylo- et par un écran/graphisme permettant de visualiser trois paramètres simultanément : deux dimensions sont représentées par la position du curseur sur deux axes orthogonaux et une troisième dimension est représentée par un dégradé de couleurs (voir figure 2.1). La position du stylo sur la tablette graphique indique la position du curseur sur l'écran.

2.1.4 Acquisition des gestes

Sans rentrer dans le détail d'une énumération des systèmes de captation des gestes (on trouvera de telles études dans [WD99] et dans [PD04]), on distingue deux modes d'acquisition :

- Acquisition *directe* : utilisation de différents types de capteurs qui mesurent l'ensemble des mouvements impliqués dans un même geste. On mesure alors le plus souvent des grandeurs physiques comme des forces, des déplacements, des accélérations, etc. Le plus souvent, chaque type de grandeur nécessite l'utilisation d'un type de capteur particulier.
- Acquisition *indirecte* : le geste est déduit dans ce cas du son produit par l'instrument. Le capteur utilisé est alors un microphone (capteur de gradient de pression) placé devant l'instrument. Les paramètres gestuels sont extraits du son par des techniques d'analyse et de traitement de signal temps-réel, ce qui nécessite une grande puissance de calcul. Pour un exemple de cette approche appliqué à la guitare, voir [TDW03].

Le système considéré dans notre étude fonctionne par l'acquisition *directe* des paramètres gestuels, bien que nous souhaitons définir un modèle qui puisse s'adapter à tout type de contrôleur.

2.2 La synthèse concaténative

Nous allons expliquer dans cette partie le mode de fonctionnement de la synthèse concaténative (CS) et en expliciter les paramètres de contrôle. Il existe plusieurs applications musicales reposant sur la synthèse concaténative. Nous nous intéresserons dans le cadre de ce stage au système développé par Diemo Schwarz (Ircam/CNRS) ([Sch06] et [SBVB06]) pour le logiciel Max/Msp de Cycling'74, appelé *CataRT* (*Caterpillar Real-Time* (voir figure 2.1), basé sur le système *Caterpillar* [Sch04]).

2.2.1 Fonctionnement général

La synthèse sonore concaténative est une méthode de synthèse de sons musicaux promettante qui s'est fortement développée depuis les 5 dernières années, générant de nombreuses publications scientifiques et artistiques. La synthèse concaténative (CS) utilise une grande base de données de sons segmentés en *unités* et un algorithme de *sélection d'unités* qui trouve les unités qui correspondent le mieux au son ou à la phrase que l'on souhaite

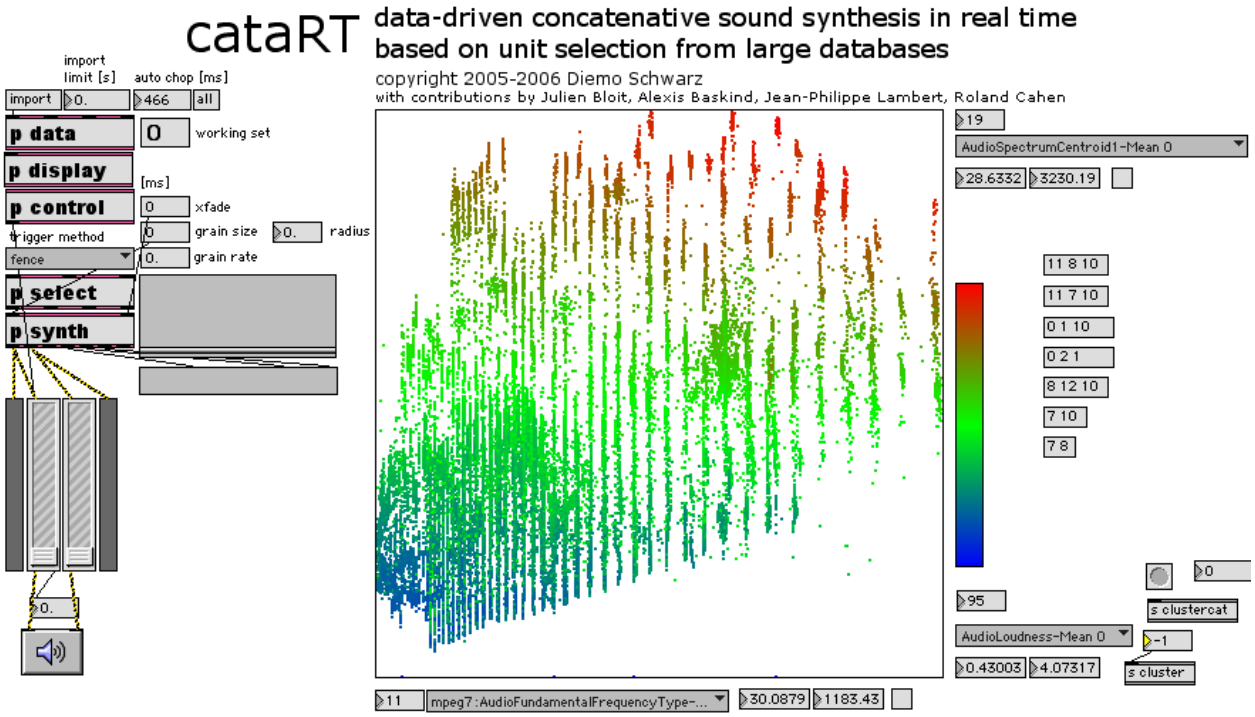


FIG. 2.1 – L'interface du système CataRT. L'écran au centre permet de visualiser et de naviguer dans la base. Ici on voit représentée une base de données contenant plus de 600 sons de saxophone, de trombone et de trompette, découpés en unités de 80ms. L'axe des abscisses montre la fréquence fondamentale des unités, l'axe des ordonnées montre le centroïde spectral et la couleur montre le volume.

synthétiser, que l'on appelle *cible*. La sélection des unités est effectuée grâce aux *descripteurs sonores* des unités. On appelle descripteurs les caractéristiques extraites d'une unité. Celles-ci peuvent être de bas niveau, comme le volume sonore (énergie) ou la hauteur (fréquence fondamentale), ou de plus haut niveau comme la classe d'instrument ou le tempo. Les axes principaux de recherche et de développement concernent à l'heure actuelle la segmentation, les descripteurs, la sélection par l'analyse de motifs et de relations (*data minig*) et l'interaction temps-réel. Les principales applications de la synthèse concaténative sont :

Synthèse haut niveau d'instrument : grâce à la constitution d'une base de données contenant par exemple l'intégralité des sons émis par un instrument et grâce à la connaissance des propriétés du son *cible*, la CS est capable de resynthétiser de façon très naturelle une phrase mélodique par la sélection précise d'*unités* en fonction du contexte.

Resynthèse audio : souvent appelée *audio mosaicing* (construction de mosaïque audio) en référence aux *mosaïques photographiques*, cette technique consiste à resynthétiser un son ou une phrase à partir d'une séquence d'unité qui correspond le mieux aux descripteurs de la cible, par exemple la hauteur, le volume, les caractéristiques timbrales, etc.

Synthèse d'ambiances et de textures : principalement utilisée pour des installations ou de la production de films. Le but est de générer des pistes sonores à partir d'une bibliothèque de sons ou d'ambiances préenregistrées ou d'étendre des paysages sonores (*soundscape*), en en gardant le caractère global et en jouant sur une grande gamme de paramètres.

Synthèse libre à partir de base de données de sons hétérogènes offrant au compositeur un contrôle efficace du résultat par l'utilisation de descripteurs perceptuellement pertinents. La cible est alors vue comme une courbe multidimensionnelle dans un espace de descripteurs. Si cette sélection se fait en temps-réel, cela permet une navigation et une exploration interactives d'un corpus sonore.

C'est la synthèse libre qui est l'application considérée dans le cadre de ce stage.

Pour une description technique du système *CataRT* et un aperçu des différents systèmes de synthèse concaténative et leurs applications, voir [Sch05], [Sch00], [Sch06] et [SBVB06]. Notons aussi une application spécifique du système proposé par D. Schwarz pour la synthèse de la voix (*Talkapillar*) [BSHR05][Hue05].

2.2.2 Constitution de la base de données : segmentation des sons et calcul des descripteurs

La puissance de la synthèse concaténative tient en premier lieu à la base de données sonores utilisée, qui est la matière première du moteur de synthèse. En effet, cette base de données peut par exemple être constituée de sons d'instrument seul, découpés en notes, et couvrant par exemple l'intégralité des sonorités possibles de ces instruments, ce qui permet une synthèse haut niveau d'instrument, très expressive [1]. Elle peut évidemment être constituée de tout type de sons, plus ou moins long, et aussi de morceaux de musique divers, segmentés d'après une reconnaissance de tempo ou aveuglément (toutes les 500 ms par exemple). Cette étape de segmentation est importante dans le sens où les descripteurs qui seront extraits des segments sonores auront plus ou moins de sens selon la nature du segment. Ainsi, si le volume sonore (énergie) est un descripteur qui fonctionne pour tout

[1] Voir notamment le logiciel *Synful*, www.synful.com

type de son, le descripteur exprimant la hauteur ou la fréquence fondamentale n'aura pas beaucoup de sens pour un segment d'enregistrement de batterie de 3 secondes de long. Dans ce cas, la valeur du descripteur pourra être vue comme une sorte de moyenne, mais n'aura pas toujours un sens perceptif (c'est-à-dire en relation avec la perception humaine du son). La pertinence des descripteurs est nécessaire car c'est à partir de ceux-ci que s'opérera la sélection des unités.

D'autre part, puisque chaque unité est la matière première du moteur de synthèse, il peut être souhaitable, selon le type de musique ou d'ambiance sonore que l'on veut générer, que l'unité soit cohérente intrinsèquement : une boucle rythmique, un son de cymbale complet et non coupé en plein milieu par exemple. L'utilisation d'outils de segmentation basés sur la reconnaissance d'*onset* (début de note) ou de tempo permet ainsi d'augmenter la cohérence de la base. Voir [AP05] pour le cas spécifique de la batterie.

Il existe à l'heure actuelle une très grande littérature concernant la définition, le calcul et l'évaluation des descripteurs sonores, notamment dans le contexte de la norme *MPEG-7* et celui du projet *CUIDADO*. Citons [PMH00] pour la description du son des instruments et [Pee04] pour une vue d'ensemble des descripteurs sonores. On notera aussi qu'à partir des descripteurs instantanés de chaque unité on peut calculer les *valeurs caractéristiques* de chaque unité [Sch04]). Celles-ci représentent l'évolution temporelle de chacun des descripteurs instantanés au sein d'une unité : moyenne, déviation standard, indices de pente, de courbure, etc.

Dans [Pee04], il est proposé plus de 150 descripteurs audio. L'utilisation d'une base de données de sons hétérogènes introduisant forcément des erreurs dans le calcul des descripteurs, et pour éviter la redondance d'information, nous souhaitons nous limiter à quelques descripteurs importants. Nous avons donc cherché quels sont les descripteurs qui ont un sens perceptif, ce qui nous a conduit à nous intéresser aux études concernant la définition d'un espace des timbres des instruments de musique.

Dans le cadre de ce stage, nous nous sommes limités à l'utilisation de six descripteurs (voir section 3.3.3). De plus, une petite étude préliminaire nous a permis de choisir la taille des unités, que nous avons fixée à 200ms (voir section 3.3.2). Enfin, la base de données considérée est composée de sons environnementaux (voir section 3.3).

2.2.3 Espace de timbre

L'ensemble des descripteurs sonores utilisé dans la CS peut être vu comme un espace multidimensionnel dans lequel sont placés les segments sonores selon les valeurs de leurs descripteurs. Afin de proposer une navigation et une exploration intuitive de l'espace des sons, nous nous sommes intéressés à l'agencement des unités dans l'espace et donc à la manière de caractériser perceptivement les sons à partir de leurs descripteurs.

Les premières études cherchant à représenter des sons dans un espace psycho-acoutique se sont basées sur la notation musicale écrite : hauteur, volume, timbre, spacialisation. La définition de timbre, dans le cas d'un instrument, a été l'objet de nombreuses recherches [LPY04, MM99, PMH00, Wes78]. Nous avons ainsi pu recenser une quinzaine de descripteurs qui ont du sens perceptivement. On notera cependant que ces études ont été effectuées sur des enregistrements d'instruments seuls et normalisés afin d'être comparés. La pertinence de ces descripteurs (hormis deux ou trois) sur une base de données de sons hétérogènes, notamment polyphoniques et multi-instruments, est donc une hypothèse qui n'a pas été vérifiée scientifiquement jusqu'à présent, et qui au dire d'un expert

en sciences cognitives interrogé est un sujet extrêmement complexe puisque dépendant en grande partie de la culture musicale et sonore de chaque personne. De plus, nous n'avons pas trouvé d'étude validant l'existence de descripteurs perceptifs généraux pour les sons environnementaux.

Afin de proposer un espace perceptuel réduit à deux ou trois dimensions, plusieurs techniques de réduction d'espace telle que *l'Analyse par Composantes Principales (PCA)*, le Positionnement multidimensionnel (*Multidimensional Scaling (MDS)*), les Cartes Auto-organisées (*Self-Organized Map (SOM)*) et la Classification *(k-means clustering [LVV03b])* ont été utilisées, démontrant la pertinence d'une telle réduction dans le cas d'une navigation et d'une représentation intuitive de l'espace des sons [ACKV02, MSP$^+$98, SM95, TB05, Ver94, MM99].

2.3 Mapping

Dans les deux parties précédentes nous avons explicité une catégorisation du geste et compris que le fonctionnement de la synthèse concaténative s'apparente à une navigation dans un espace de son. Le contrôle gestuel de la synthèse concaténative peut donc être vu comme la stratégie adoptée pour faire correspondre un geste à un déplacement dans l'espace des sons, ce que l'on appelle un mapping. Nous avons alors cherché à connaître les différentes stratégies de mapping utilisées dans l'élaboration d'un instrument de musique électronique.

2.3.1 Un modèle de mapping multicouche

Le choix d'un mapping revêt d'une importance capitale car, pour une même interface et un même moteur sonore, il est responsable en grande partie du caractère de l'instrument, de ses qualités. De plus, les réactions psychologiques et émotionnelles provoquées chez l'instrumentiste sont presque entièrement dues au mapping [HWP02]. Dans [HWK00], A. Hunt et M. Wanderley exposent une vue d'ensemble des différentes stratégies de mapping et démontrent pourquoi un mapping complexe est requis pour maximiser les possibilités d'interaction humaine en situation de manipulation experte. Cette affirmation est corroborée par de nombreuses études [WSR98, Got00, HWP02, RWDD97].

Un modèle général admis par tous est de séparer le mapping en deux parties au moins : l'une relative au contrôleur, l'autre relative au moteur de synthèse. Cette séparation permet l'utilisation de plusieurs interfaces de contrôle pour un même moteur de synthèse, ou l'inverse. Un modèle plus complexe et plus modulaire, étendant ce principe et s'appuyant sur des espaces perceptuels est décrit dans [ACKV02] (voir figure 1.1). Ce modèle propose trois couches :

- des données gestuelles à un espace perceptuel des gestes (couche relative à *l'interprétation*),
- d'un espace perceptuel des sons aux paramètres du moteur de synthèse (couche relative à *la définition de l'instrument*),
- le lien entre ces deux espaces perceptuels.

Dans le cadre de ce stage, nous nous concentrerons sur la partie dite *relative à la définition de l'instrument* de ce modèle (voir chapitre 6).

2.3.2 Catégorisation des stratégies de mapping

Il existe tout d'abord trois types de mapping :

- *un-vers-un* : un paramètre d'entrée contrôle un paramètre du moteur de synthèse.
- *un-vers-plusieurs* : un paramètre d'entrée va contrôler simultanément plusieurs paramètres du moteur de synthèse.
- *plusieurs-vers-plusieurs* : plusieurs paramètres d'entrée vont interagir simultanément dans le contrôle de plusieurs paramètres du moteur de synthèse.

On distingue ensuite les mapping *explicites*, dans lesquels la relation entre deux paramètres est clairement définie par une expression mathématique, et les mapping *implicites*, utilisant des mécanismes génératifs (comme des réseaux de neurones) ou des règles générales (base de données) [CCH05, CCH04, MMS03, FH98, Mod00]. L'utilisation d'un mapping *explicite* offre une meilleure visibilité sur les interactions produites et permet de définir en détail l'expressivité de l'instrument. A contrario, un mapping *implicite* est une sorte de «boite noire» dans laquelle on définit un comportement global et non des valeurs précises.

On opposera aussi un comportement statique à un comportement dynamique d'un mapping. Il existe plusieurs niveaux d'interprétation de ce comportement. Le premier est la capacité du mapping à évoluer avec le temps ou d'apprendre à partir des données d'entrée. La deuxième interprétation correspond à l'utilisation par le mapping de paramètres dynamiques de description du geste. Un mapping peut être une combinaison de ces deux idées. Dans le cas où un mapping est adaptatif, il peut correspondre à une évolution d'un mapping à un autre, doucement ou abruptement, ou aussi à l'adaptation d'un mapping.

On remarquera dans [ACKV02] qu'il est démontré qu'un mapping prenant en compte des paramètres dynamiques est très intéressant. En effet, ces paramètres dynamiques prennent en compte plus d'informations à propos du geste et de son intention, permettant par exemple d'évoluer et d'aller d'un espace perceptuel à un autre. L'expressivité est alors codée par des paramètres instantanés ou des dérivées.

Enfin on notera plusieurs études mettant en relation les propriétés géométriques d'un espace avec différentes stratégies de mapping [CBG95, Gou02, NWD04, MW03].

2.4 Conclusion

Afin de proposer un modèle générique de contrôle de la synthèse concaténative, nous avons retenu de cette étude bibliographique les points suivants :

Geste : Un geste musical est assimilable à une combinaison de gestes classés en gestes de sélection, d'excitation, de modification et de polarisation. En fonction du contrôleur considéré, il conviendra alors de définir quels paramètres ou quelles combinaisons de paramètres du contrôleur correspondent à ces différentes classes de geste.

Synthèse concaténative : Le principal paramètre de contrôle de la synthèse concaténative est la sélection des unités dans la base de données. La base de données est vue comme un espace de sons dont l'agencement est défini par les valeurs des descripteurs sonores de chacune des unités. C'est sur cet agencement que nous allons travailler.

Mapping : le modèle de mapping retenu comporte trois couches : l'une relative au contrôleur, définissant un espace perceptuel des gestes, une autre relative au moteur de synthèse, définissant un espace perceptuel des sons et la dernière faisant le lien entre l'espace perceptuel des gestes et l'espace perceptuel des sons.

Les expériences présentées dans les sections suivantes visent à définir la couche relative au moteur de synthèse en proposant pour la synthèse concaténative un espace des sons

perceptuellement compréhensible et des stratégies de navigation et de visualisation de cet espace. C'est donc plus sur la définition de l'instrument lui-même que nous allons travailler.

Chapitre 3

Problématiques et études préliminaires

Nous avons vu dans le chapitre précédent un état de l'art des travaux sur lesquels s'appuie notre étude. Comme nous l'avons expliqué en introduction, nous n'étudierons qu'une petite partie de ce que serait un modèle général de mapping pour le contrôle de la synthèse concaténative, en nous concentrant sur la définition de stratégies de navigation et de visualisation d'une base de données sonores. Dans les deux premières parties de ce chapitre nous allons exposer les contraintes et les problématiques déduites du système de synthèse sonore concaténative et du contrôleur gestuel considérés. Nous éclairerons ainsi l'intérêt de proposer de telles stratégies. Puis nous montrerons les études préliminaires qui ont guidé les choix effectués afin de constituer la base de données sonores.

3.1 Contraintes

Puisque le principal paramètre de contrôle de la synthèse concaténative est la sélection d'unités représentées par un espace multidimensionnel de descripteurs sonores, nous nous sommes concentrés sur différentes stratégies de navigation et de représentation de cet espace. L'environnement de test utilisé est le système CataRT sous Max/Msp, contrôlé à l'aide d'une tablette graphique. Il découle de cette configuration plusieurs contraintes :

- La visualisation de l'espace des sons n'est permise que dans trois dimensions : abscisse, ordonnée et couleur (voir figure 2.1). Chaque dimension permet de visualiser un descripteur sonore.
- La navigation est contrôlée par deux dimensions qui sont les positions en abscisse et en ordonnée du stylo sur la tablette graphique.
- Le mapping entre la tablette et l'interface de CataRT est très simple : il s'agit d'un mapping *one-to-one* entre les coordonnées du stylo sur la tablette graphique et les coordonnées du curseur de sélection des unités dans l'interface de CataRT.
- le mode de synthèse utilisé est le mode *fence* : les unités sélectionnées sont jouées entièrement les unes après les autres, sans répétition d'une même unité. La vitesse de déplacement du curseur de sélection des unités sur l'écran de contrôle affecte donc le caractère continu du résultat sonore.

On remarquera notamment que selon la représentation choisie (les descripteurs sur les différents axes), les unités forment des «tas» compacts qui limitent l'exploration (plusieurs unités étant superposées sur le même pixel) (*cf* figure 3.1).

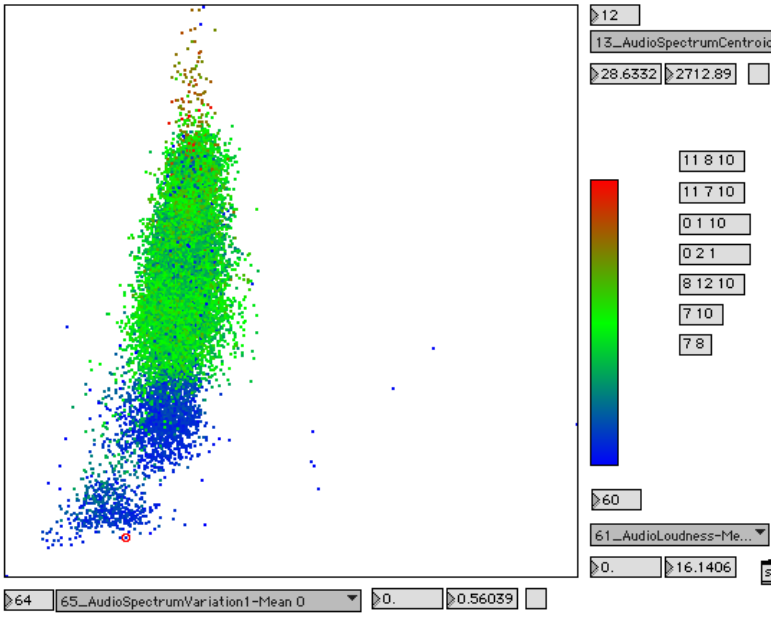


FIG. 3.1 – Cette représentation ne maximise pas les possibilités d'exploration : de nombreuses unités sont représentées sur un même pixel. Comment éclater cet amas tout en gardant des distances qui représentent les différences entre les unités?

3.2 Problématiques

Ces contraintes nous ont donc amenés à trouver des solutions pour les problèmes suivants :

- Comment visualiser toutes les informations contenues dans la base grâce à une représentation utilisant seulement trois dimensions?
- Comment naviguer de façon intuitive dans la base?
- Comment visualiser et naviguer dans plus de trois descripteurs en même temps?
- Comment déterminer des sous-ensembles de sons semblables afin de réduire la complexité de l'espace des sons?
- Quel est le meilleur moyen de naviguer dans un sous-ensemble de sons semblables? i.e comment déterminer les descripteurs qui permettent d'exploiter au mieux un sous-ensemble de sons?
- Comment obtenir une distribution graphique des unités qui évite les «tas» d'unités, tout en respectant les différences entre les unités?

Ces questions cherchent plus généralement à exploiter au mieux un corpus de sons, en simplifiant l'espace des sons. Afin d'y répondre, nous avons tout d'abord sélectionné un petit ensemble de descripteurs sonores considérés comme perceptivement compréhensibles, puis nous avons utilisé un algorithme statistique, l'Analyse par Composantes Principales (ACP) pour simplifier l'espace des sons. La segmentation de l'espace à été réalisée grâce à l'algorithme de classification *Kmeans*.

Les sections suivantes présentent les études préliminaires qui ont été effectuées pour constituer la base de données. Les résultats de l'analyse par composantes principales et résultats de la classification sont présentés respectivement dans les chapitres 4 et 5. Le chapitre 6 montre les différentes stratégies de navigation et de visualisation qui utilisent les résultats décrits.

3.3 Etudes préliminaires

La synthèse concaténative est utilisée dans notre cas pour réaliser de la synthèse sonore libre (voir section 2.2.1). Nous avons choisi de considérer une base de données de sons environnementaux. Les sons ont été choisis de façon heuristique, en gardant cependant en tête deux points importants détaillés ci-dessous. On considérera plusieurs sous-ensembles de sons, qui formeront ensuite notre base de données sonores :

Sons de vagues : Il proviennent de la banque de son SoundIdeas[1] 6035. Ce sont des enregistrements de différents bruits de vagues. Chaque fichier dure entre 20 secondes et 1 minute 30.

Sons de voiture avec effet doppler : Ils proviennent des banques de sons SoundIdeas 6005 à 6008. Ce sont des enregistrements d'une voiture qui s'approche, passe puis s'éloigne. Chaque banque correspond à un modèle de voiture particulier qui diffère notamment par le bruit du moteur. On dispose de quatre enregistrements par type de voiture, chacun correspondant à une vitesse de passage plus ou moins rapide. Chaque enregistrement dure environ 30 secondes.

Sons d'explosions : Ils proviennent de la banque de sons SoundIdeas 6040. Chaque enregistrement dure environ 15 secondes.

Contrairement aux expériences réalisées pour la caractérisation des timbres des instruments de musique (voir section 2.2.3), ces sons ne sont pas normalisés en volume et n'ont pas été enregistrés en chambre anéchoïque.

Bien que notre approche se veuille généraliste sur la nature des sons que peut contenir la base, les sous-ensembles de sons ont été choisis pour nous permettre de considérer deux points importants :

1. En connaissant par avance le déroulement temporel de chaque sous-ensemble de sons, on peut valider facilement les résultats des algorithmes étudiés.
2. Chaque son a un déroulement temporel facilement interprétable par un geste : cela nous sera utile plus tard pour reproduire les sons par le contrôle de la tablette graphique.

3.3.1 Calcul des descripteurs

Pour calculer les descripteurs sur la base de données sonores, nous avons utilisé le programme Matlab *GDClass* développé par G. Peeters dans le cadre du projet européen CUIDADO. Le set de descripteurs calculés dans GDClass est détaillé dans [Pee04]. Deux types de descripteurs sont calculés pour chaque fichier son : des descripteurs instantanés calculés toutes les 10 ms sur une fenêtre de 30 ms et des descripteurs globaux donnant une valeur pour l'ensemble d'un fichier son. Nous ne considérerons que des descripteurs instantanés.

3.3.2 Segmentation des sons

Après cette analyse, nous utilisons la méthode de calcul des *valeurs caractéristiques* décrite dans [Sch04]). Pour mémoire les valeurs caractéristiques décrivent l'évolution temporelle des descripteurs instantanés au sein d'une unité (*cf* section 2.2.2). Il convient donc d'abord de décider quelle sera la taille des unités dans la base de données. Ce choix est important tant sur le point du résultat sonore que sur la pertinence des valeurs caractéristiques :

[1] http://www.sound-ideas.com/6000.html

- Une segmentation en unité courte favorise une description très fine de l'unité, ce qui est intéressant pour de nombreux descripteurs instantanés.
- Une segmentation en unité plus longue permet par contre une meilleure description de l'évolution temporelle d'un descripteur. Cela permet par exemple de reconnaître la présence de vibrato en observant la variation de la fréquence fondamentale. Il est important de noter que la longueur de l'unité est aussi liée à la perception que l'on a d'un son : trop courte, tous les sons sonneront de façon identique. Une unité de l'ordre de 500ms permet de distinguer une évolution dans le son, ou d'entendre un caractère plus ou moins percussif.
- Dans le cadre d'une base de données d'enregistrements de musique polyphonique, une segmentation selon le tempo de chaque morceau pourrait permettre d'optimiser la cohérence des unités. On peut imaginer aussi que la taille des unités dépendrait de leur contenu : segmentation plus fine sur des passages percussifs, segmentation plus longue sur des passages mélodiques lents (notes tenues, accord), chaque élément restant un multiple de l'unité de base choisie (croche, double croche).

Les figures 3.2, 3.3 et 3.4 présentent pour un même son les courbes de quatre descripteurs pour des unités de longueur variable. Pour des unités de 500ms, l'évolution globale est très bien représentée, les courbes des descripteurs sont lissées. Pour des unités de 80ms, le surplus d'information donne des courbes très discontinues, ce qui tend à fausser la description du son. Pour des unités de 200ms, on a un compromis qui décrit bien l'évolution des descripteurs et qui permet d'avoir des unités relativement courtes (pour le calcul des descripteurs) et suffisamment longue pour que l'on puisse entendre une évolution sonore dans l'unité.

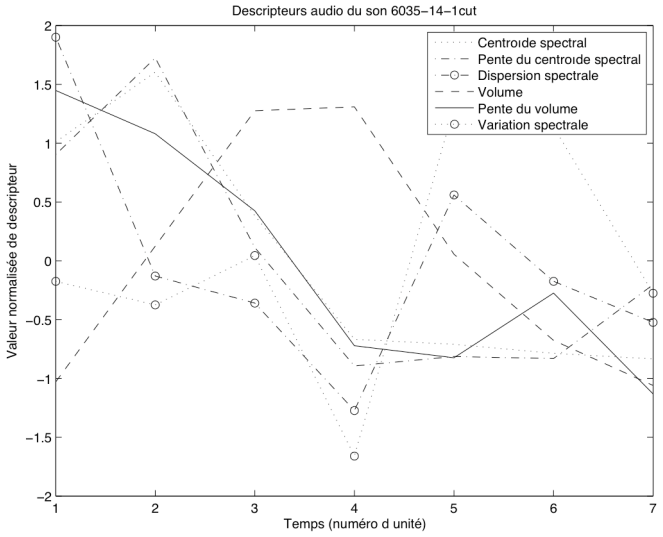


FIG. 3.2 – Représentation graphique de 4 descripteurs pour des unités de 500ms

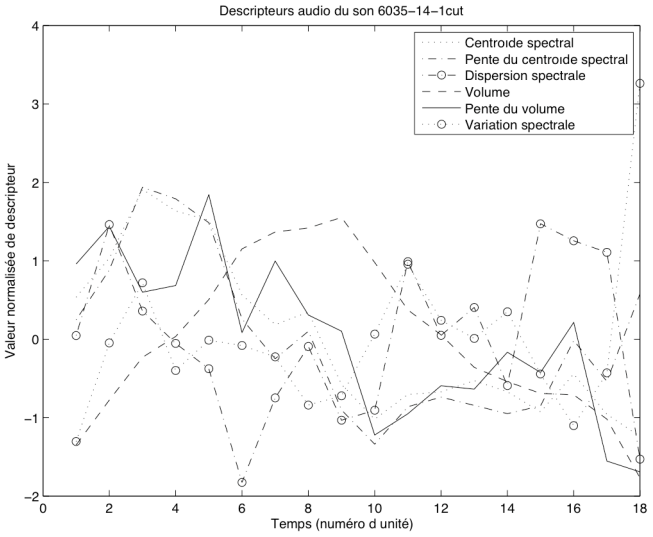


FIG. 3.3 – Représentation graphique de 4 descripteurs pour des unités de 200ms

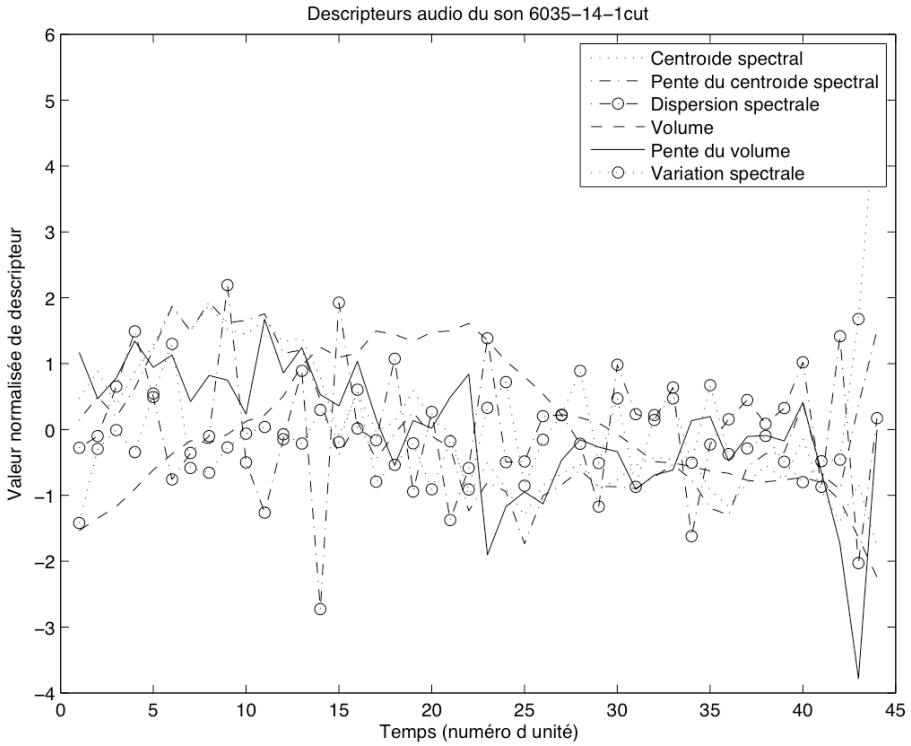


FIG. 3.4 – Représentation graphique de 4 descripteurs pour des unités de 80ms

3.3.3 Sélection des descripteurs

L'ensemble des descripteurs globaux et des valeurs caractéristiques des descripteurs instantanés représente plus de 1500 descripteurs pour chaque unité, ce qui est beaucoup et ne permet pas d'obtenir une représentation des sons. En effet, une étude préliminaire (non-documentée ici) a fait état de l'impossibilité d'obtenir une représentation correcte de l'information portée par ces 1500 descripteurs dans un graphique à deux dimensions. Le nombre de dimensions nécessaires serait plutôt de l'ordre d'une trentaine de dimensions. De plus, en fonction du matériau sonore analysé (par exemple, son harmonique ou non) certains descripteurs n'ont pas d'intérêt et produisent des valeurs fausses (voir section 2.2.2). Ainsi, pour proposer un espace de navigation «perceptif», nous avons souhaité réduire le nombre de descripteurs considérés. D'après [Pee04], nous avons sélectionné un ensemble de descripteurs pertinents d'un point de vue perceptif, espérant ainsi que l'utilisateur sera capable d'identifier rapidement les différences sonores entre deux points éloignés de l'espace (connaître par avance ce que le son va être). Deux indices de pente ont été ajoutés. Ils offrent une description temporelle des descripteurs au sein d'une unité. Ils permettent par exemple de connaître, pour une même valeur de volume, si celui-ci tend à croître, décroître ou rester stable.

Au final, la liste des descripteurs utilisés pour nos expériences est la suivante :

- Centroïde spectral ;
- Indice de pente du centroïde spectral ;
- Variation spectrale
- Volume ;
- Indice de pente du volume ;
- Dispersion spectrale.

Ces descripteurs sont définis par (voir [Pee04]) :

Le centroïde spectral : Le centroïde spectral, ou centre de gravité du spectre. Les études perceptives montrent qu'il est un des paramètres les plus pertinents dans la descriptiton du timbre. Il est calculé par :

$$\mu = \int_0^1 \nu.a(\nu)\, d\nu$$

avec ν la fréquence réduite et $a(\nu)$ le spectre d'amplitude normalisé.

La dispersion spectrale : La dispersion spectrale (ou spectral spread) est une mesure de l'étendue du spectre autour de sa valeur moyenne. Elle est calculée par un moment d'ordre 2 :

$$\sigma^2 = \int_0^1 (\nu - \mu)^2.a(\nu)\, d\nu$$

avec μ la valeur moyenne de $a(\nu)$.

Variation spectrale : La variation spectrale, aussi appelée *spectral flux* représente la quantité de variation du spectre dans le temps. Elle est calculée à partir de la corrélation normalisée entre deux spectres d'amplitude successifs $a(t-1)$ et a(t).

$$variation = 1 - \frac{\sum_k a(t-1,k) \cdot a(t,k)}{\sqrt{\sum_k a(t-1,k)^2}\sqrt{\sum_k a(t,k)^2}}$$

Volume : Le volume, ou plus précisément le *volume total* est calculé à partir du *volume spécifique* qui est le volume associé à chacune des bandes de fréquences *Bark*.

$$Volume\_total = \sum_{z=1}^{nb\_bande} N'(z)$$

où $N'(z)$ est le volume spécifique de la z^{ieme} bande de fréquence Bark.

Indice de pente : Mesure le taux de décroissance d'une valeur par une régression linéaire.

Chapitre 4

Réduction des dimensions de l'espace des sons

Nous avons vu dans le chapitre précédent les choix qui ont été effectués pour constituer la base de données sonores. Dans ce chapitre, nous allons montrer comment visualiser les unités qui composent la base de données en deux dimensions seulement tout en gardant des distances qui représentent les différences entre les unités. Ces distances ne sont ni des distances acoustiques ni des distances perceptives. Pour y parvenir on utilise une méthode statistique, l'Analyse par Composantes Principales. Cette méthode à déjà été utilisée dans de nombreux cas pour la description des espaces de timbre (cf section 2.2.3). La première partie de ce chapitre expose le principe de l'Analyse par Composantes Principales. Les parties suivantes présentent successivement les résultats obtenus sur plusieurs sous-ensembles croissants de sons, afin de mettre en évidence le caractère générique de notre approche.

4.1 Présentation de l'algorithme PCA

L'analyse en composantes principales (ACP) est une technique mathématique permettant de réduire un système complexe de corrélations en un plus petit nombre de dimensions. Lorsqu'on étudie simultanément un nombre important de variables quantitatives (ne serait-ce que 4), il est ardu de faire un graphique global. La difficulté vient en effet de ce que les observations étudiées ne sont plus représentées dans un plan, espace de dimension 2, mais dans un espace de dimension plus importante (par exemple 4). L'objectif de l'Analyse en Composantes Principales (ACP) est de revenir à un espace de dimension réduite (par exemple 2) en déformant le moins possible la réalité. Il s'agit donc d'obtenir le résumé le plus pertinent possible des données initiales.

C'est la matrice des variances-covariances (ou celle des corrélations) qui va permettre de réaliser ce résumé pertinent, parce qu'on analyse essentiellement la dispersion des données considérées. De cette matrice, on va extraire, par un procédé mathématique adéquat, les facteurs que l'on recherche, en petit nombre. Ils vont permettre de réaliser les graphiques désirés dans cet espace de petite dimension (le nombre de facteurs retenus), en déformant le moins possible la configuration globale des observations selon l'ensemble des variables initiales (ainsi remplacées par les facteurs).

Géométriquement, le processus de la mise en facteurs revient à placer des axes dans un ballon de rugby. Si l'on considère 100 dimensions, il est peu probable que nous amassions suffisamment d'informations sur une seule droite, le long de ce grand axe de cet

hyperballon de rugby, droite appelée première composante principale. Nous aurons besoin d'axes supplementaires. Par convention, nous représentons la deuxième dimension par une droite perpendicutaire à la première composante principale. Ce deuxième axe, ou deuxième composante principale, se définit comme la droite qui «explique» (le mot n'a pas ici de signification causale) la plus grande partie de l'information restante (aucune autre droite qui «expliquerait» autant ou davantage ne pourrait être tracée perpendiculairement à la première composante principale. Si, par exemple, l'hyperballon de rugby était aplati comme une limande (petit poisson plat, comme la sole), la première composante principale passerait par le centre, de la tête à la queue, et la deuxième également par le centre de l'animal, d'un côté à l'autre (voir figure 4.1). Toutes les droites suivantes seraient perpendiculaires aux axes précédents et contiendraient de moins en moins d'informations (sur la forme du poisson).

Pour obtenir une description complète de l'Analyse par Composantes Principales voir [BB05].

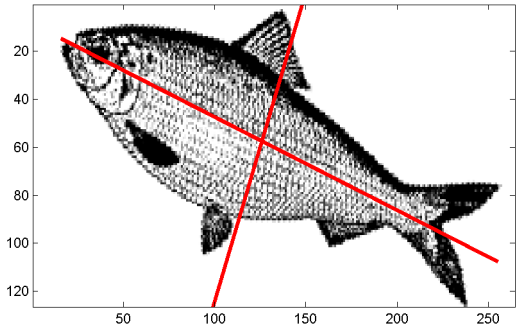


FIG. 4.1 – Résultat d'une ACP pour une image. Les deux axes trouvés sont les deux premières composantes principales : elles expliquent au mieux la dispersion des points de l'image.

4.2 Etude préliminaire sur un son de vague

On considère pour cette étude préliminaire un petit bout d'un enregistrement de sons de vagues, représentant le bruit d'une vague, découpé en unités de 200ms, chacune décrite par les 6 descripteurs choisis. On cherche à savoir s'il est possible de représenter les unités sonores de ce son dans un espace réduit à deux dimensions tout en gardant une représentation des différences entre les unités. On considère uniquement ce bout d'enregistrement, et non l'enregistrement complet, afin de pouvoir visualiser graphiquement le résultat de l'algorithme sur un déroulement temporel connu.
Nous avons en premier lieu normalisé les descripteurs, chacun étant de dimension physique différente. Les valeurs ont été centrées puis normalisées par l'écart-type. La figure 4.3 représente l'évolution des descripteurs choisis sur un fichier contenant le son d'une vague.

On remarque que la courbe du centroïde spectral montre clairement une plus grande quantité de fréquences aiguës au début du son. Cette information est intéressante car elle nous renseigne sur des variations du spectre qui ne sont pas forcément décelables à l'oreille, mais qui permettent de mettre en perspective les différentes unités du son.

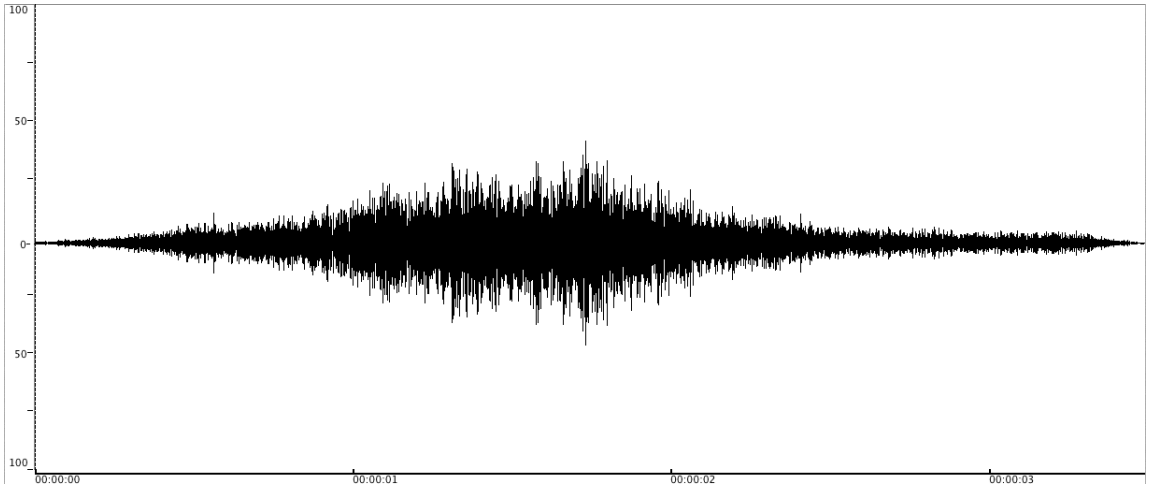


FIG. 4.2 – Forme d'onde d'un son de vague

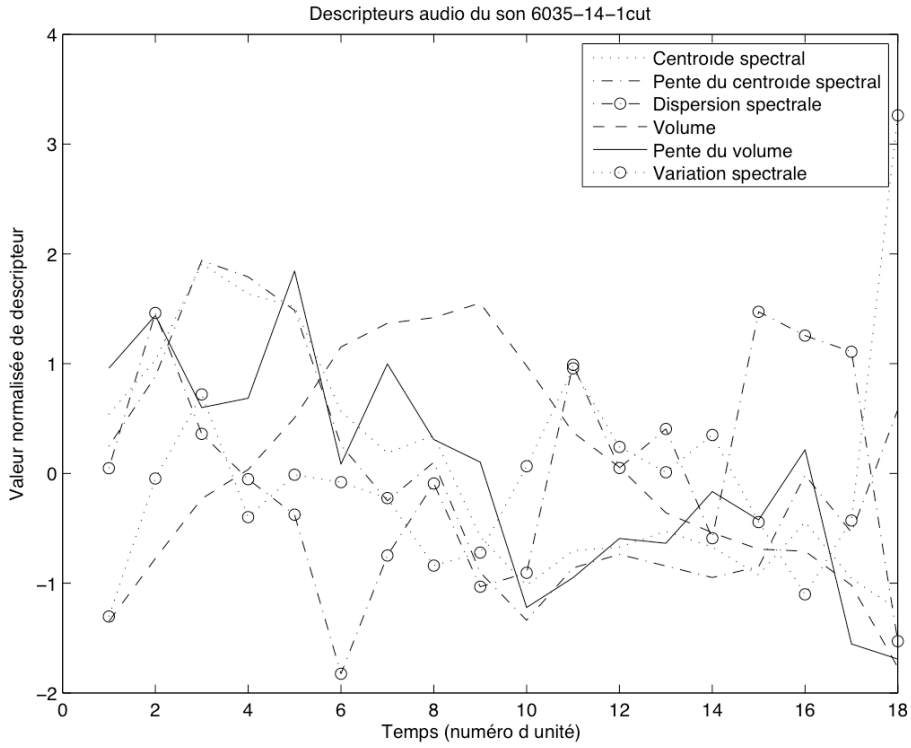


FIG. 4.3 – Représentation graphique des 6 descripteurs choisis pour un son de vague

Voici à présent les résultats obtenus pour l'analyse des composantes principales sur les 6 descripteurs choisis, pour ce même son de vague (voir figure 4.4).

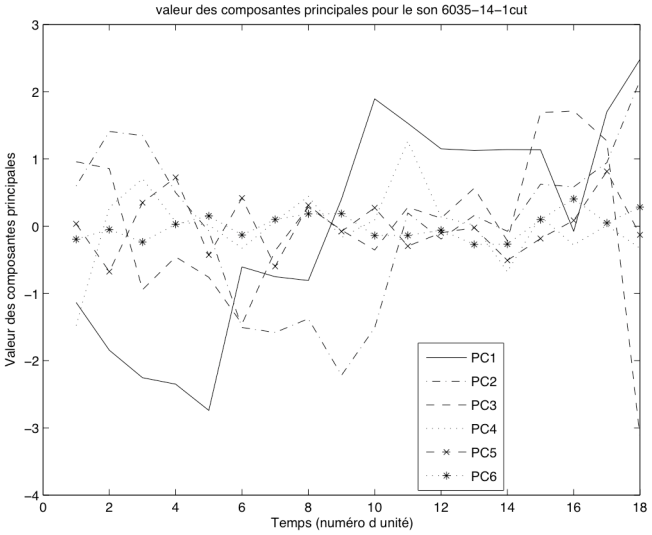


FIG. 4.4 – Représentation graphique des 6 composantes principales pour un son de vague

À partir des valeurs propres de la matrice de covariance, qui nous donnent un indice d'explication de la variance pour chaque composante, on peut calculer le pourcentage de la variation totale expliqué par chaque composante :

$$pourcentage\_explication = 100 \times \frac{variances}{\sum variances}$$

, où *variances* correspond au vecteur de la quantité de variance expliquée par chaque composante. On obtient pour le son de vague le résultat décrit dans le tableau 4.1.

On peut en déduire qu'une représentation en deux dimensions utilisant les deux premières composantes principales explique à 70% la variance des unités, ce qui est tout à fait acceptable pour réduire le nombre de dimensions. On peut remarquer que dans le cas d'une représentation selon les trois premières composantes, on aurait plus de 90%.

1ère composante	43,84 %
2ème composante	24,22 %
3ème composante	22,96 %
4ème composante	5,29 %
5ème composante	3,01 %
6ème composante	0,69 %

TAB. 4.1 – Pourcentage de variabilité expliqué par chaque composante principale pour un son de vague

La matrice de corrélation variables-facteurs (tableau 4.2) montre la contribution de chacun des descripteurs pour les 6 composantes principales. Ainsi on voit que le premier

Composantes	ACP1	ACP2	ACP3	ACP4	ACP5	ACP6
Centroïde spectral	-0.5964	0.0719	-0.1193	0.1495	0.2150	-0.7459
IP du centroïde spectral	-0.4954	0.3554	-0.3017	0.0894	0.3955	0.6106
Dispersion spectrale	-0.0396	0.4038	0.6837	0.5955	-0.1037	0.0506
Volume	-0.1254	-0.7523	-0.1053	0.6128	0.0289	0.1758
IP du Volume	-0.5692	-0.0947	0.0942	-0.2283	-0.7606	0.1660
Variation spectrale	0.2400	0.3613	-0.6382	0.4329	-0.4552	-0.0995

TAB. 4.2 – Matrice de corrélation variables-facteurs pour un son de vague. Les colonnes représentent les composantes principales, les lignes les descripteurs sonores.

facteur est corrélé positivement, et moyennement, avec le centroïde spectral, l'indice de pente (IP) du centroïde spectral et l'IP du volume. Il est par contre corrélé faiblement et négativement avec la variation spectrale. Une corrélation positive signifie que si les valeurs de descripteurs pour une unité sont fortes, alors cette unité aura une valeur forte sur l'axe, et inversement pour une corrélation négative. Le deuxième facteur oppose le volume à l'IP du centroïde spectral, à la variation spectrale et à la dispersion spectrale.

4.3 Résultats sur un ensemble de sons de voiture avec effet Doppler

On considère à présent une partie des sons de voiture avec effet doppler (sous-ensemble de sons 6006 —un type de voiture, quatre vitesses de passage différentes—). On augmente donc le nombre de sons dans la base tout en changeant de type de son. Le résultat de l'algorithme est exposé dans le tableau 4.3.

1ère composante	47,38 %
2ème composante	18,09 %
3ème composante	15,95 %
4ème composante	12,07 %
5ème composante	4,69 %
6ème composante	1,18 %

TAB. 4.3 – Pourcentage de variabilité expliqué par chaque composante principale pour les sons 6006

Ce résultat est bon puisqu'une visualisation avec les deux premières composantes explique à 65% la variance des unités. Une représentation en trois dimensions serait encore une fois bien meilleure. L'analyse de la matrice de corrélation variables-facteurs (tableau 4.4) montre que la première composante est très corrélée avec le centroïde spectral, l'IP du centroïde spectral, le volume et un peu avec la variation spectrale.

Composantes	ACP1	ACP2	ACP3	ACP4	ACP5	ACP6
Centroïde spectral	0.5624	-0.0607	0.0232	-0.1598	-0.2025	0.7829
IP du centroïde spectral	0.5251	-0.0819	0.0211	-0.1801	0.7992	-0.2142
Dispersion spectrale	-0.0129	0.6878	0.6702	-0.2781	-0.0056	-0.0154
Volume	0.5416	-0.1133	0.0327	-0.1830	-0.5657	-0.5825
IP du Volume	0.0572	0.6259	-0.7402	-0.2374	-0.0135	-0.0226
Variation spectrale	0.3335	0.3347	0.0275	0.8803	0.0038	-0.0337

TAB. 4.4 – Matrice de corrélation variables-facteurs pour le jeu de son 6006. Les colonnes représentent les composantes principales, les lignes les descripteurs sonores.

4.4 Résultats sur l'ensemble des sons de la base

Voyons à présent les résultats pour l'ensemble des sons qui constituent notre base. On réunit donc une quarantaine de sons, soit un peu plus de 17800 unités sonores de 200ms de long, provenant de trois types de sons différents —sons de vagues, sons de voitures et sons d'explosions—.

1ère composante	49,20 %
2ème composante	21,10 %
3ème composante	12,21 %
4ème composante	9,79 %
5ème composante	4,68 %
6ème composante	3,09 %

TAB. 4.5 – Pourcentage de variabilité expliqué par chaque composante principale pour tous les sons de la base

Composantes	ACP1	ACP2	ACP3	ACP4	ACP5	ACP6
Centroïde spectral	0.5190	-0.0160	0.0286	-0.1402	-0.8228	0.1811
IP du centroïde spectral	0.5309	-0.0161	-0.0282	0.2860	0.1117	-0.7892
Dispersion spectrale	-0.0177	-0.7072	0.7052	0.0429	0.0188	-0.0045
Volume	0.4472	-0.0296	0.0178	-0.7775	0.4335	0.0804
IP du Volume	0.0257	0.7060	0.7071	0.0147	0.0215	-0.0141
Variation spectrale	0.4978	-0.0012	-0.0272	0.5404	0.3489	0.5811

TAB. 4.6 – Matrice de corrélation variables-facteurs pour tous les sons de la base. Les colonnes représentent les composantes principales, les lignes les descripteurs sonores.

On retrouve encore un très bon résultat puisque 2 dimensions suffisent pour expliquer à 70% la variance des unités (voir tableau 4.5 et figure 4.5). L'analyse de la matrice de corrélation variables-facteurs (tableau 4.6) montre les mêmes résultats que pour le jeu de sons 6006 : la première composante est très corrélée avec le centroïde spectral, l'IP du centroïde spectral, le volume et la variation spectrale. La deuxième composante oppose la dispersion spectrale à l'IP du volume. Ces deux informations se retrouvent de façon intuitive dans CataRT puisque l'on entend bien par exemple que les sons présents dans le haut du graphique sont beaucoup plus forts que ceux présents en bas. Les figures 4.6 et 4.7 mettent en évidence ces deux informations : on remarque bien un dégradé de couleurs qui suit l'un des deux axes.

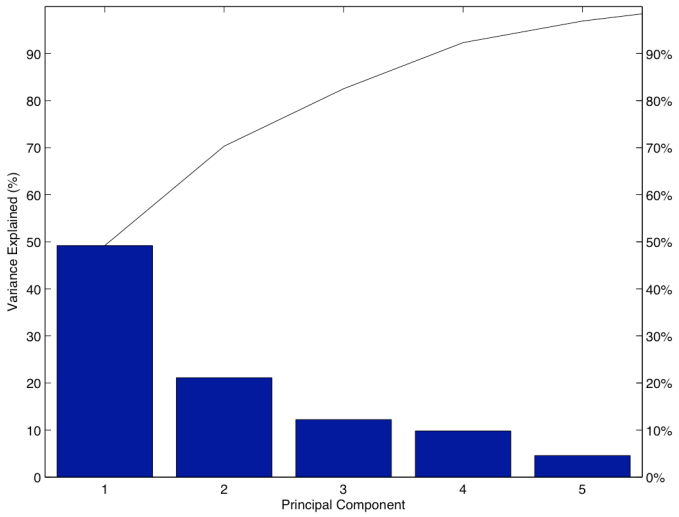


FIG. 4.5 – Pourcentage de variation expliqué par chaque composante principale pour l'ensemble des sons de la base.

Conclusion

L'étude de l'*Analyse par Composantes Principales* sur notre base de données nous a permis de valider l'intérêt d'une telle méthode pour réduire la complexité de l'espace des sons tout en gardant une représentation des différences des unités entre elles. Ainsi, même réduit à deux dimensions, l'espace de navigation et de visualisation proposé reste intuitif. De plus, les composantes principales étant corrélées à plusieurs descripteurs, on navigue dans plusieurs descripteurs en même temps, ce qui est aussi l'une des problématiques que l'on cherche à résoudre. Enfin, les résultats étant probant sur l'ensemble de la base comme sur les deux sous-ensembles considérés, on peut faire l'hypothèse que notre approche pourra être généralisée sur d'autres sons environnementaux.

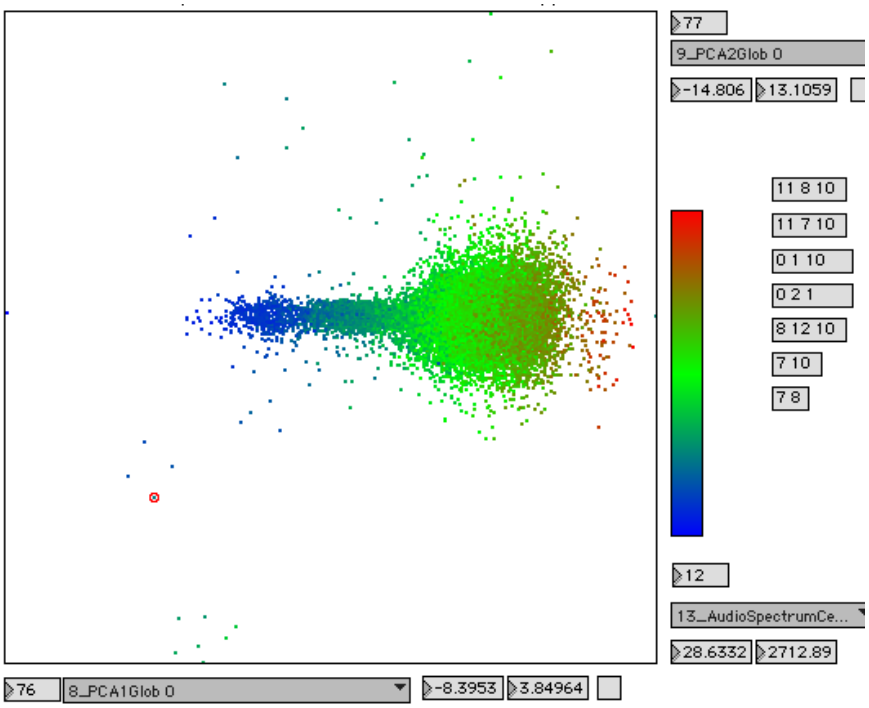


FIG. 4.6 – Visualisation dans CatART de l'ensemble de la base de son. La première composante (axe des abscisses) est très corrélée notamment avec le centroïde spectral (couleur). L'axe des ordonnées correspond à la deuxième composante principale.

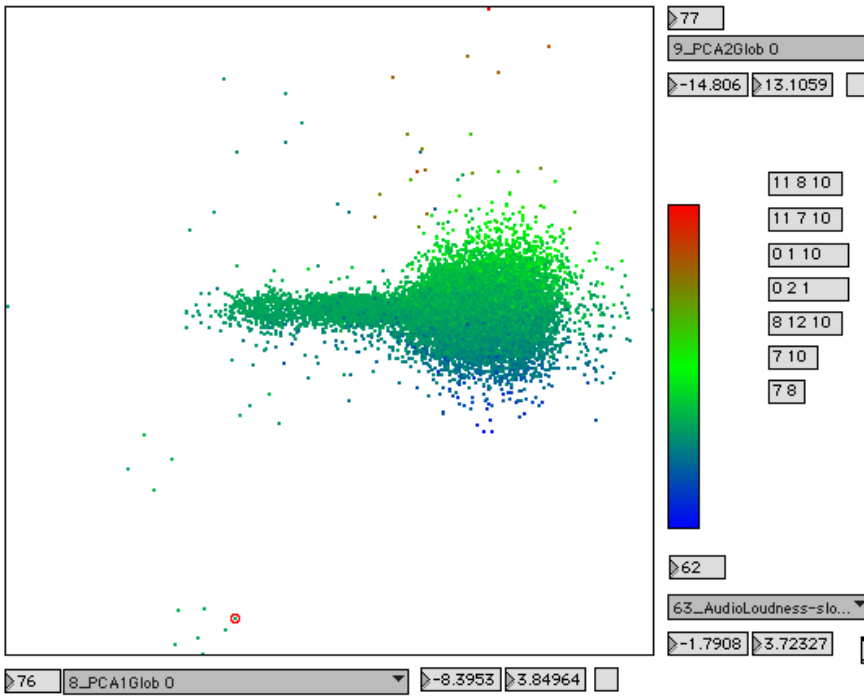


FIG. 4.7 – Visualisation dans CatART de l'ensemble de la base de son. La deuxième composante sur l'axe des ordonnée est très corrélée avec l'indice de pente du volume (couleur). L'axe des abscisses correspond à la première composante principale

Chapitre 5

Catégorisation d'unités sonores

Dans le chapitre précédent (Chapitre 4), nous avons montré comment obtenir une représentation correcte en deux dimensions des informations contenues dans la base de données. L'espace de navigation et de visualisation qui en découle se révèle assez intuitif. Nous voulons à présent réduire encore la complexité des informations contenues dans la base en déterminant des sous-ensembles d'unités sonores semblables et en proposant une représentation visuelle de ces sous-ensembles. La première partie de ce chapitre expose le fonctionnement de l'algorithme Kmeans *qui est utilisé pour réaliser la classification des unités sonores. Les parties suivantes présentent successivement les résultats obtenus sur plusieurs sous-ensembles croissants de sons.*

5.1 Présentation de l'algorithme de classification *Kmeans*

D'un point de vue général, la classification (*clustering* en anglais) peut être considérée comme la recherche automatique d'une structure ou d'une typologie au sein d'une collection de données non-étiquetées. Une définition approximative du clustering pourrait être alors : le processus de classification d'objets en groupes ou classes dont les membres sont d'une certaine manière similaires. Par conséquent, une classe est un ensemble d'objets qui sont « similaires » entre eux et « différents » des objets appartenant aux autres classes. Le but de la classification est donc de déterminer la structure intrinsèque d'un ensemble de données non-étiquetées. De nombreux algorithmes ont été développés depuis un quart de siècle, notamment l'algorithme des K-means et ses nombreuses variantes. Les algorithmes de classification de données ont de nombreuses applications dans des domaines divers et variés : Biologie, Marketing, Bibliothèque, Assurances, Sismologie, Urbanisme... C'est ainsi, par exemple, que les différents êtres vivants ont été classifiés en plusieurs espèces.

L'algorithme K-means, créé par MacQueen en 1967 est l'algorithme de classification le plus connu et le plus utilisé car il s'avère être très simple à mettre en oeuvre et efficace. Il suit une procédure simple de classification d'un ensemble d'objets en un certain nombre K de classes, K fixé à priori.
Dans cet algorithme, chaque classe (aussi appelée prototype) est caractérisée par son centre qui se trouve être le barycentre ou la moyenne des éléments composant la classe —K-means est ainsi généralement traduit par K-moyennes. Soit un ensemble d'objets
$D_n = (x_1, x_2, \ldots, x_n)$, avec pour tout i, x_i réel et soit $\mu_k | 1 < k < K$, les centres des K classes.

L'algorithme K-means s'exécute en 4 étapes :

- **Étape 1** : Choisir aléatoirement K objets qui forment ainsi les K classes initiales. Pour chaque classe k, la valeur initiale du centre est $\mu_k = x_i$, avec x_i l'unique objet de D_n appartenant à la classe.
- **Étape 2** : (Ré-)Affecter les objets à une classe. Pour chaque objet x, le prototype qui lui est assigné est celui qui est le plus proche de l'objet, selon une mesure de distance, (habituellement la mesure euclidienne) :

$$s = argmin_k \|\mu_k - x\|^2$$

- **Étape 3** : Une fois tous les objets placés, recalculer les centres des K classes (les barycentres).
- **Étape 4** : Répéter les étapes 2 et 3 jusqu'à ce que plus aucune réaffectation ne soit faite.

Même si l'algorithme termine toujours, on n'est pas assuré d'obtenir la solution optimale. En effet, l'algorithme est très sensible au choix aléatoire des K centres initiaux. C'est pourquoi, on utilise souvent l'algorithme des K-means de nombreuses fois sur un même ensemble de données pour essayer de minimiser cet effet tout en sachant que des centres initiaux les plus espacés possibles donnent de meilleurs résultats. Dans notre étude, le choix des K centres initiaux est effectué grâce à une phase de classification préliminaire opérée sur un sous-échantillon de 10% des données, ce qui améliore les résultats de l'algorithme. On remarquera cependant que l'un des inconvénients majeurs des K-means, outre le fait qu'il faille exécuter à plusieurs reprises l'algorithme pour ainsi avoir un résultat le plus optimal possible, est le besoin d'initialiser le nombre de prototypes au début de l'exécution. Ceci nuit à l'efficacité de l'algorithme car dans la pratique, on ne connaît à priori pas le nombre de classes finales.
Pour obtenir une description complète de l'algorithme Kmeans, voir [LVV03b] et [BB05].

5.2 Etude d'un son de vague

De la même manière que pour l'étude de l'Analyse par Composantes Principales, on cherche à comprendre comment réagit l'algorithme de classification Kmeans sur des ensembles de sons de plus en plus complexes, afin de savoir si la méthode pourra être généralisée sur de grandes bases de données sonores.

L'étude sans à priori d'un son de vague (voir figure 5.1) sur une classification en trois parties nous montre très clairement une segmentation que l'on peut interpréter comme trois phases temporelles : croissance, pic, et décroissance. Une analyse de classe plus fine -avec 5 classes- fait ressortir une distinction plus nette entre les unités, donc une meilleure description des phases. Il faut remarquer que comme l'algorithme à une initialisation aléatoire, il n'attribue pas toujours la même classe à une unité entre deux exécutions. De plus, le numéro de classe est aussi choisi de façon aléatoire entre deux exécutions : Sur la figure 5.1, l'algorithme a attribué les numéros de classe de telle manière que cette représentation est très visuelle, une autre exécution de l'algorithme aurait peut-être renversé l'ordre des classes.

On va à présent faire l'étude sur plusieurs sons semblables.

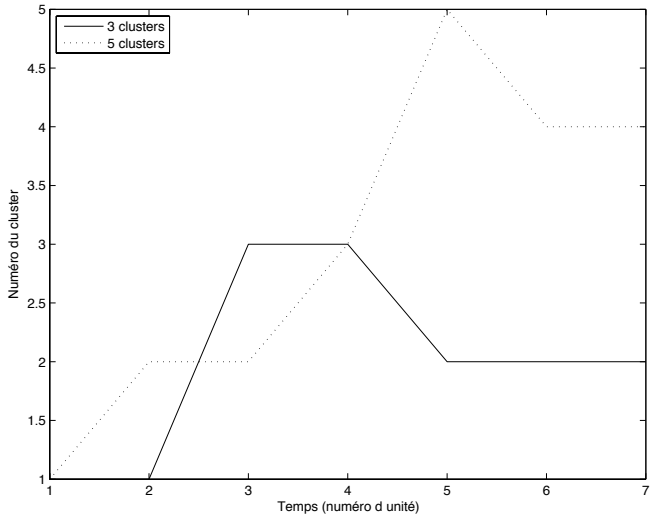


FIG. 5.1 – Représentation graphique du résultat de l'algorithme kmeans pour un son de vague

5.3 Résultats sur un son de voiture avec effet Doppler

On considère à présent le son 6006-27 (son de voiture qui s'approche, passe et s'éloigne, lentement). On cherche à savoir si l'on obtient aussi une classification en phases comme pour le son de vague. On remarque dans la figure 5.2 que le centroïde spectral forme une bosse similaire à celui du volume, ce qui nous informe sur les différents moments du son. Nous retiendrons donc pour les prochains graphiques ces deux descripteurs pour caractériser le déroulement temporel du son. Ainsi on pourra comprendre assez simplement à l'aide d'un graphique comment l'algorithme Kmeans classifie chaque unité du son, et voir s'il attribue à une même classe des unités pourtant éloignées temporellement. Cette information est intéressante car elle prouve que l'algorithme essaye de réunir dans la même classe des unités semblables.

La figure 5.3 montre le résultat de la classification en 3 et 5 classes. La classification suit bien le déroulement temporel du son : on distingue trois phases avec la classification à 3 classes : croissance, climax et décroissance, bien que cette dernière ne soit pas représentée avec une seule classe. Avec une classification à 5 classes, l'algorithme essaye de distinguer la phase montante de la phase descendante du pic central et utilise une classe pour caractériser les deux courtes phases qui entourent le pic central (classe numéro 5 sur la figure 5.3). Bien que le son considéré ici soit assez différent du son de vague étudié ci-dessus, on retrouve une classification similaire dans le sens où celle-ci permet de distinguer des phases temporelles dans le son.

Il faut à présent faire une remarque importante sur les résultats présentés : en effet l'initialisation aléatoire de l'algorithme donne une classification plus ou moins bonne d'une exécution à l'autre, notamment pour une classification à 5 classes. Ce problème est dû au fait que l'algorithme cherche des optimums locaux. Lorsque plusieurs unités consécutives temporellement évoluent sans rupture forte entre elles, l'algorithme a des difficultés à

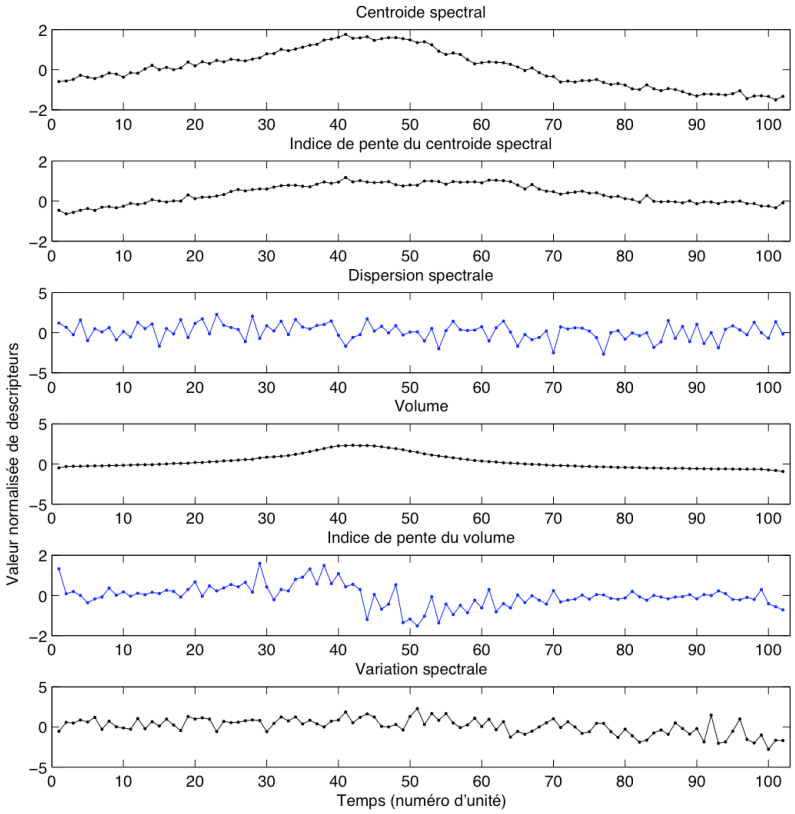


FIG. 5.2 – Représentation graphique des 6 descripteurs choisis pour le son 6006-27

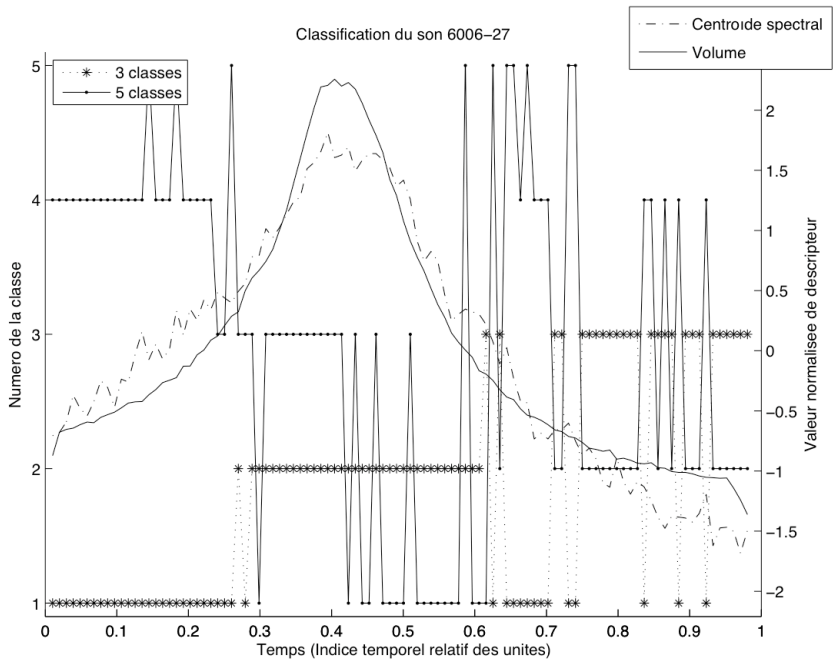


FIG. 5.3 – Représentation graphique du volume, du centroïde spectral et du résultat de l'algorithme kmeans pour le son 6006-27

déterminer à quelle classe appartiennent deux unités qui se suivent. Pour avoir une idée de la bonne séparation ou non des classes, on utilise un graphique nommé *silhouette* (voir figures 5.5 et 5.6). Celui-ci affiche une mesure de la distance entre chaque point d'une classe et les points des classes avoisinantes. Cette mesure va de +1, indiquant que les points sont très distants des classes voisines, passe par 0, indiquant que les points ne sont pas distinctement dans une classe ou une autre, et s'arrête à −1, indiquant que les points sont probablement assignés à une mauvaise classe.

5.4 Résultats sur l'ensemble des sons de voiture avec effet Doppler

Nous avons jusqu'à présent étudié séparément le comportement de l'algorithme Kmeans sur deux sons différents. Afin de valider la généralisation de notre approche sur un grand ensemble de sons, nous allons considérer un ensemble plus complexe de sons, les sons de voitures. Pour rappel, l'ensemble des sons de voiture choisis (jeux de sons 6005 à 6008) présente les caractéristiques suivantes :

- chaque son représente une voiture qui arrive, passe et s'éloigne, produisant un effet *doppler* ;
- chaque banque de sons (6005, 6006, 6007 et 6008) représente un modèle de voiture différent, ayant une signature sonore —le bruit du moteur— bien distincte ;
- chaque modèle de voiture est enregistré quatre fois à des vitesses de passage différentes.

Deux paramètres distinguent donc cet ensemble de sons similaires : la vitesse de passage de la voiture (paramètre temporel) et le bruit de la voiture elle-même (paramètre sonore). Nous avons étudié le comportement de l'algorithme Kmeans en comparant deux approches :

- l'algorithme est calculé sur chaque son séparément ;
- l'algorithme est calculé sur un ensemble de sons.

Cela nous permettra de vérifier que la classification n'est pas perturbée lorsqu'elle est calculée sur un ensemble de sons.

Afin de voir les conséquences d'une variation du déroulement temporel, nous avons tout d'abord analysé séparément les quatre sons de la banque 6006, tel que représenté dans les figures 5.4, 5.5 et 5.6. Nous ne présentons pas le résultat graphique de chaque analyse. Puis nous avons lancé le calcul de l'algorithme sur l'ensemble de ces quatre sons. Une comparaison entre les figures 5.7 et 5.4 montre bien que la classification d'un son est très sensiblement la même, qu'il soit étudié séparément ou dans un ensemble de sons légèrement différents. Enfin, nous avons considéré les quatre banques de sons ensemble pour vérifier que la classification en phases temporelles est similaire même lorsque les sons ont des signature sonores différentes.

Nous pouvons conclure de ces expériences que :

- la classification est sensiblement la même que les sons soient considérés séparément ou tous ensemble ;
- l'algorithme trouve toujours une classification en phases temporelles, qu'il y ait présence ou non de sons ayant une signature sonore différente ;
- l'algorithme réagit bien à une variation temporelle du déroulement des phases dans le son : ici plus la vitesse de passage augmente, plus les classes qui caractérisent le pic central sont courtes ;
- une classification plus fine (ici 5 classes) permet de mieux distinguer les différentes phases.

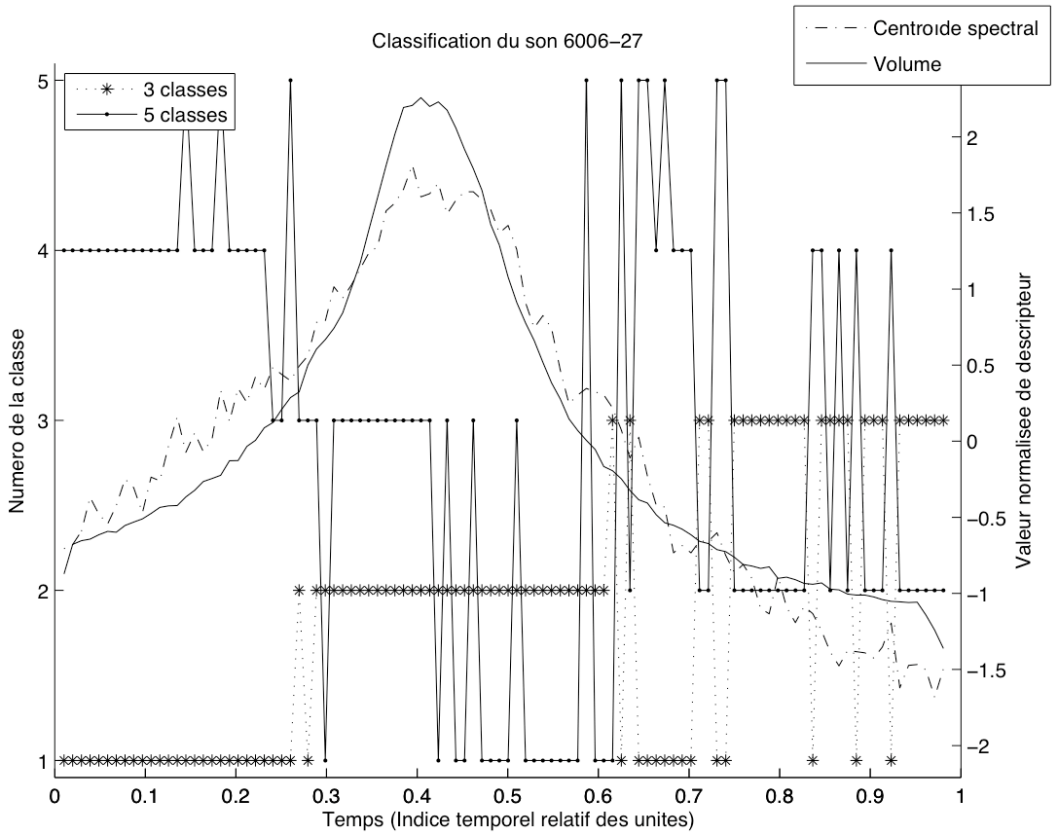


FIG. 5.4 – Représentation graphique du volume, du centroïde spectral et du résultat de l'algorithme kmeans pour le son 6006-27

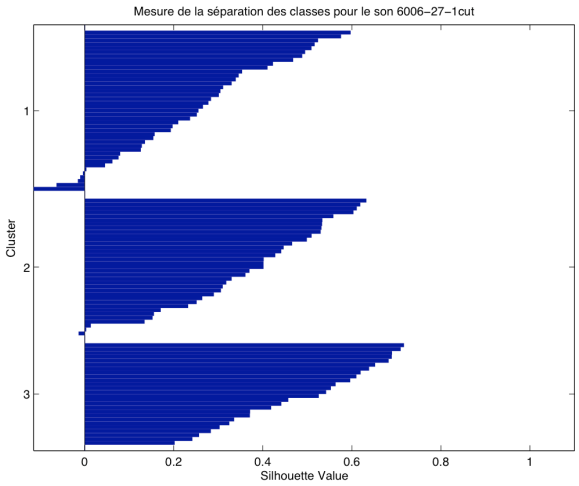


FIG. 5.5 – Représentation de la séparation des 3 classes pour le son 6006-27

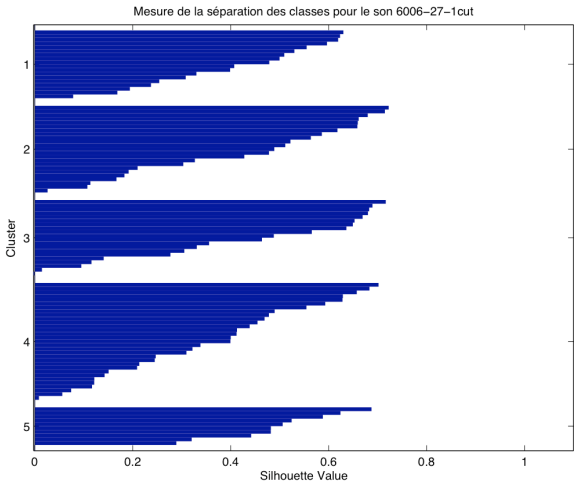


FIG. 5.6 – Représentation de la séparation des 5 classes pour le son 6006-27

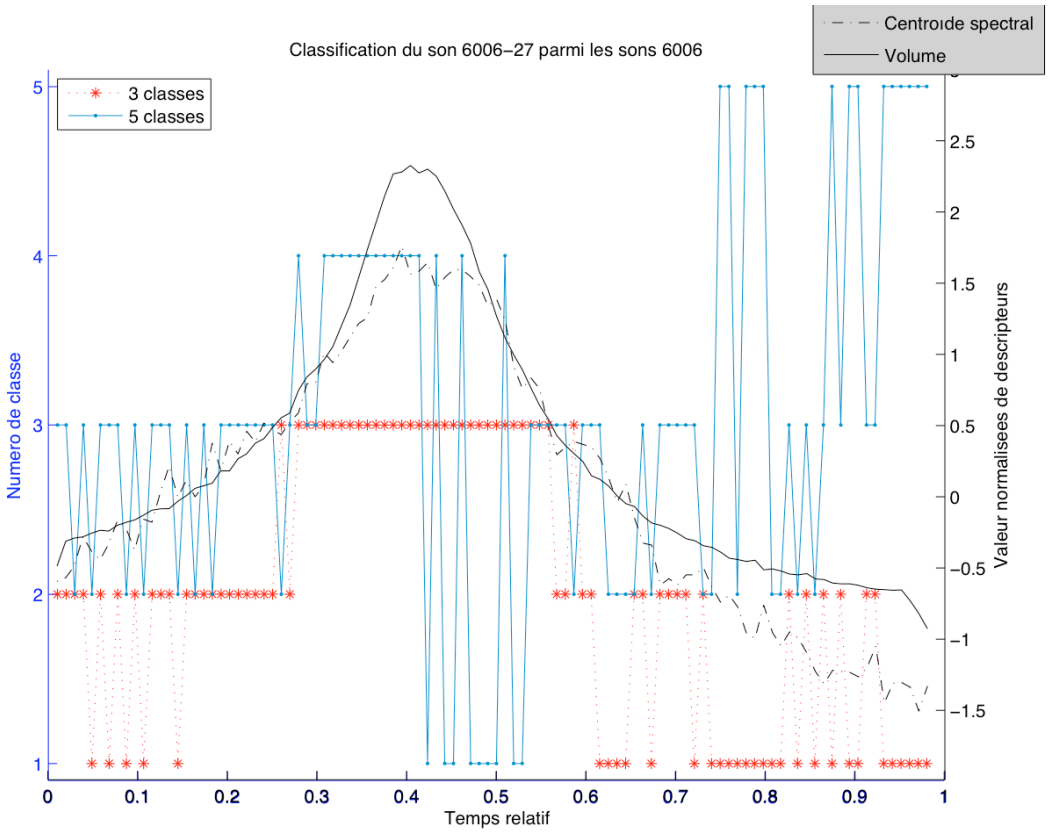


FIG. 5.7 – Classification des unités du son 6006-27 parmi les sons 6006

On remarquera d'ailleurs qu'en augmentant le niveau de classification, nous avons obtenu petit à petit une distinction entre les différents types de voiture, pour une même phase temporelle.

5.5 Résultats sur l'ensemble des sons de la base

On considère à présent l'ensemble des sons de la base.
La figure 5.8 montre le résultat de l'algorithme Kmeans pour le son 6006-27 considéré dans l'ensemble des sons de la base. Le nombre de classes a été augmenté afin d'obtenir une classification plus fine. La décomposition en phases est bien la même que dans les études précédentes. Il est d'ailleurs intéressant de noter que les classification en 24 et en 5 classes donnent globalement le même résultat, puisque les phases de croissance et de décroissance du son commencent aux mêmes unités. La classification en 24 classes attribue cependant plus de classes pour caractériser la partie centrale, ce qui est bien ce à quoi on pouvait s'attendre, puisque que l'on retrouve par exemple une distinction entre les différents types de voiture. La figure 5.9 montre que les classes sont dans l'ensemble assez peu distinctes entre elles. Ceci s'explique comme précédemment par le fait que les unités sonores ont souvent une continuité entre elles.

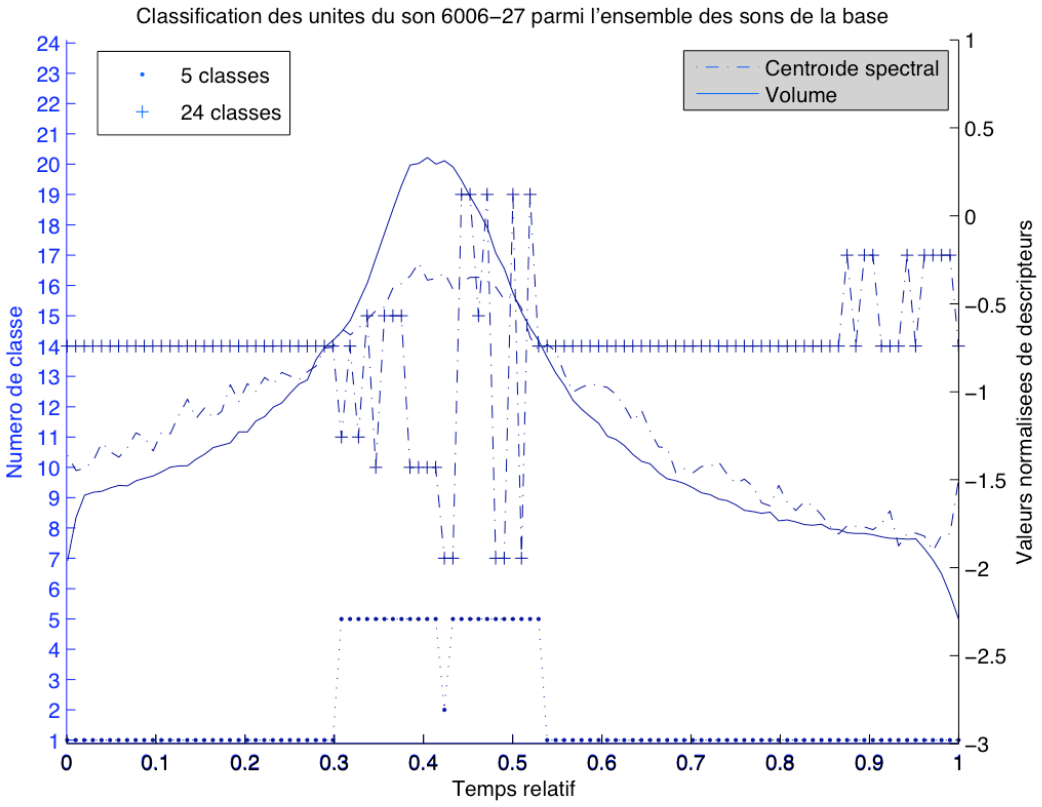


FIG. 5.8 – Classification des unités du son 6006-27 parmi l'ensemble des sons de la base, pour une classification à 5 et 24 classes.

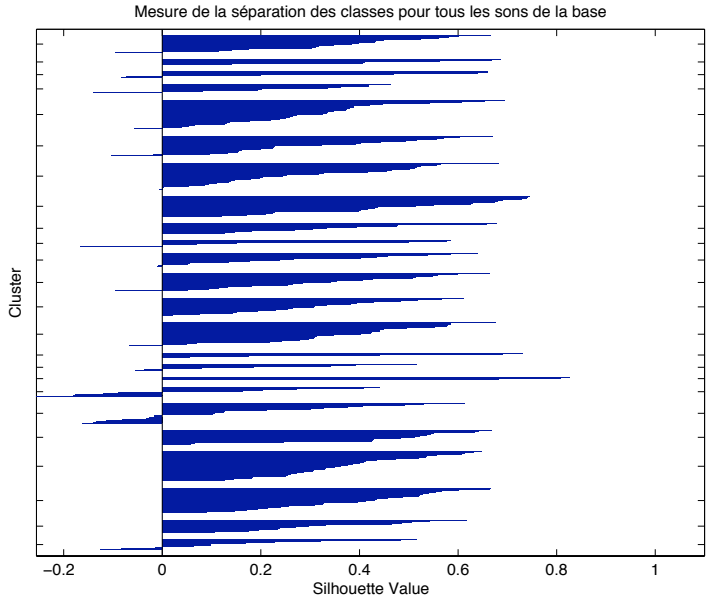


FIG. 5.9 – Mesure de la séparation des classes pour une classification en 24 classes pour l'ensemble des sons de la base.

Conclusion

On a montré que l'algorithme de classification *Kmeans* donne de bons résultats pour la classification des unités sonores de la base de données. Celui-ci permet notamment de caractériser des ensembles d'unités sonores similaires provenant de fichiers sons différents, et comportant cependant des variations de signature sonore. Ainsi, on retrouve dans une même classe un grand nombre des phases d'attaque des sons d'explosions. On remarquera aussi que selon le nombre de classes, on obtient une séparation entre les classes plus ou moins bonne, ce qui se traduit ensuite par une superposition plus ou moins forte entre les classes (voir figures 5.10 et 5.11).

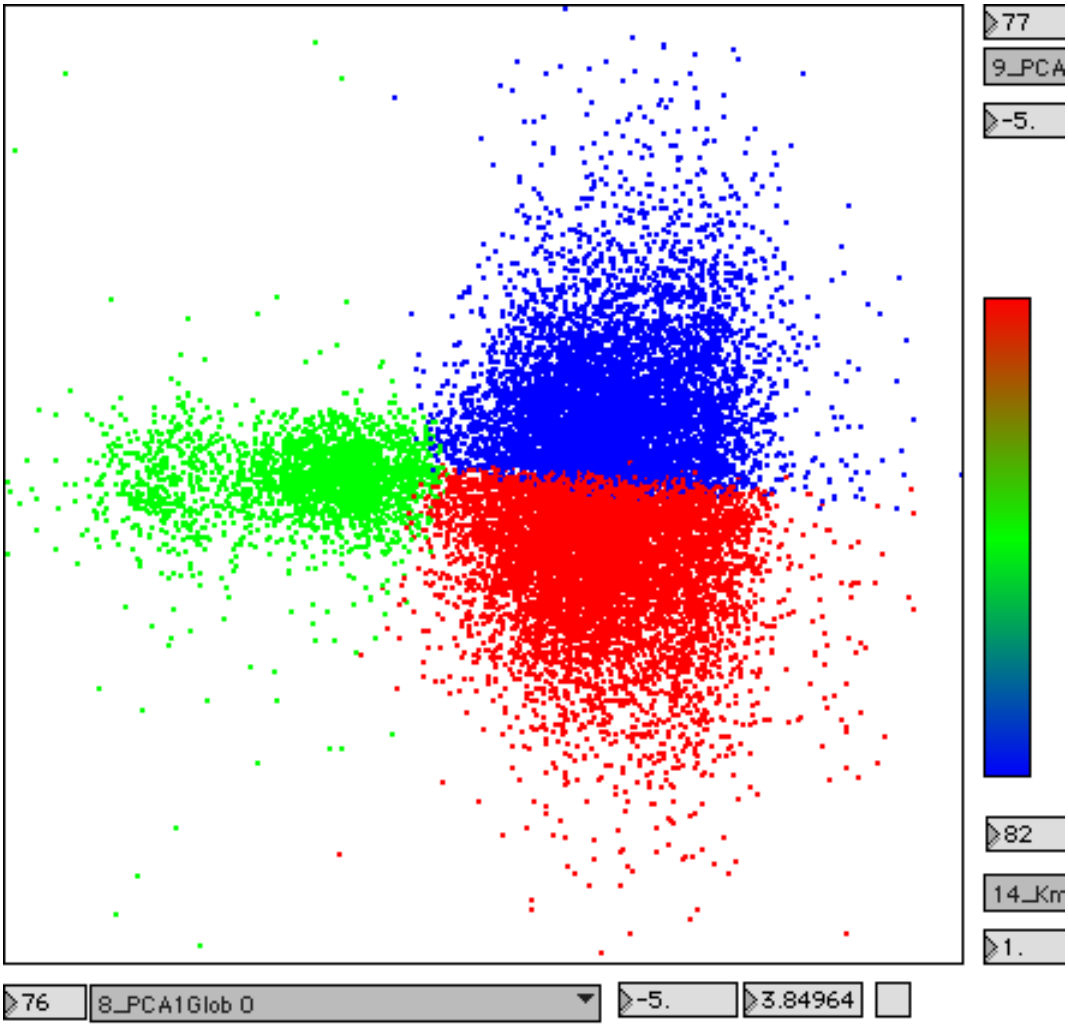


FIG. 5.10 – Représentation graphique de l'ensemble des sons de la base selon les composantes principales. La couleur montre l'appartenance des unités aux différentes classes pour une classification à 3 classes. On remarque une superposition très faible des classes.

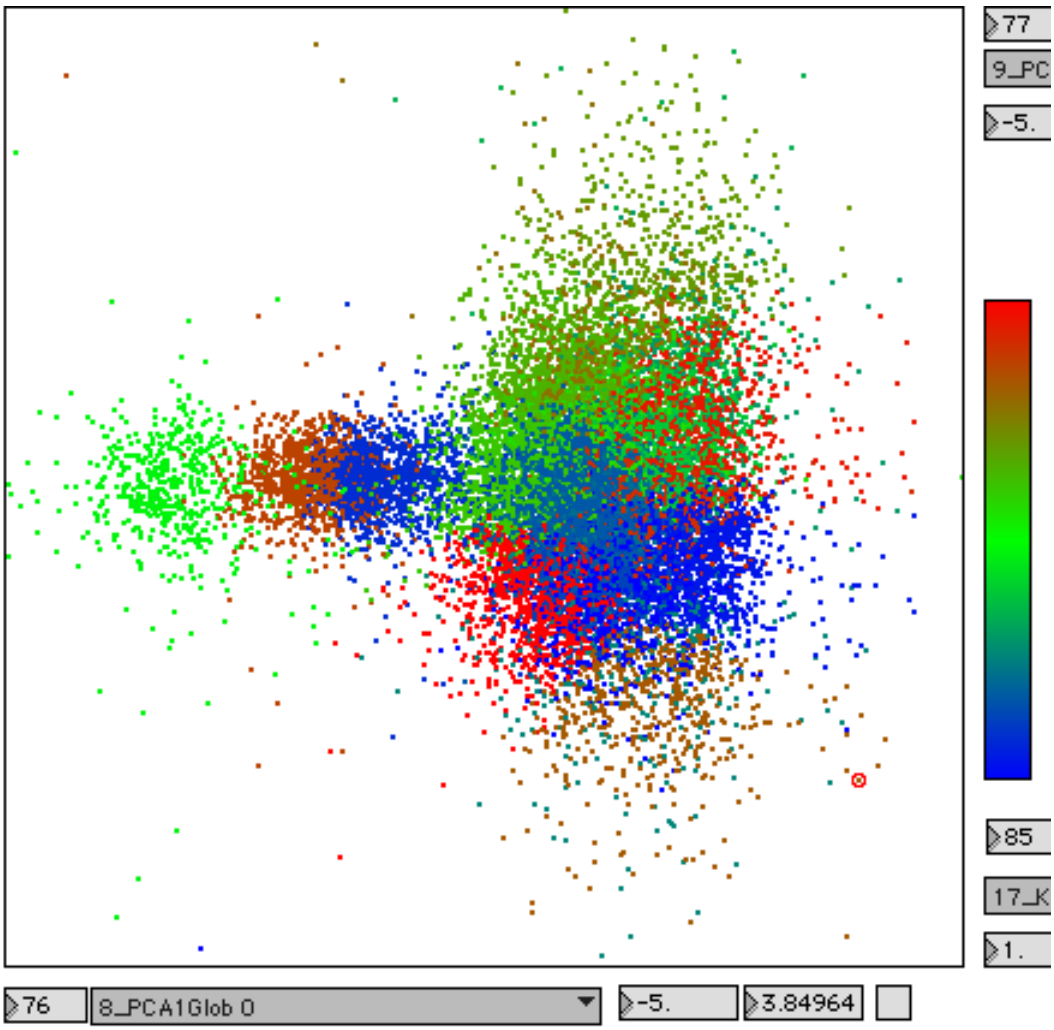


FIG. 5.11 – Représentation graphique de l'ensemble des sons de la base selon les composantes principales. La couleur montre l'appartenance des unités aux différentes classes pour une classification à 24 classes. On remarque une superposition forte des classes sur la partie droite, alors que la partie gauche a été séparée en trois classes assez dictinctes.

Chapitre 6

Stratégies de navigation et de visualisation

Le contrôle gestuel de la synthèse concaténative s'apparente au contrôle de la navigation dans un espace d'unités sonores. Les deux chapitres précédents ont montré l'intérêt d'utiliser des méthodes d'analyse de données telles que l'Analyse par Composantes Principales (chapitre 4) et la classification Kmeans *(chapitre 5) pour réduire la complexité des données contenues dans la base en proposant un espace de navigation et de visualisation intuitif. Ce chapitre résume quelques stratégies de visualisation et de navigation déduites des méthodes étudiées. Une première partie montre l'intérêt de combiner une visualisation par classe avec une représentation en composantes principales. Une deuxième partie expose les études récentes que nous avons effectuées pour trouver une représentation qui maximise la dispersion des unités au sein d'un espace de visualisation, ce que nous appelons un effet* loupe.

6.1 Visualisation et navigation par classe

Nous avons vu au cours des études précédentes que la représentation en composantes principales propose un espace intuitif de navigation au sein de la base de données sonores (voir figures 4.6 et 4.7). De plus, l'utilisation de couleurs pour caractériser les unités nous permet de visualiser un descripteur à la fois, ou encore différentes classes d'unités (voir figures 5.10 et 5.11).

À partir de ces informations, il nous a semblé très utile de pouvoir afficher une seule classe d'unités à la fois, afin de ne visualiser que les unités les plus semblables à celle choisie. De plus, plus le niveau de classification est élevé, plus le nombre d'unités dans la classe est faible, ce qui revient à trouver les unités les plus semblables à celle choisie. Nous avons alors défini un mapping très simple permettant d'explorer les données de la base de façon efficace. Les deux capteurs présents sur le côté du stylet de la tablette graphique ont été mappés (en *one-to-one*) avec le contrôle de la visualisation par classe. Un capteur est utilisé pour incrémenter le niveau de classification, un autre pour le décrémenter. Un curseur rouge visualise l'unité qui sert de référence. Les figures 6.1, 6.2, 6.3 et 6.6 montrent le résultat de la manipulation.

Cependant, on remarque que la visualisation de la classe n'est pas optimale dans le sens où les points sont très superposés les uns sur les autres. En utilisant une fonction qui n'affiche que l'espace occupé par la classe (« zoom » automatique), on améliore un peu la représentation et la navigation comme dans la figure 6.5. Mais cette représentation n'est pas valable dans le cas où beaucoup d'unités sont représentées en même temps. Nous

avons donc cherché un moyen de maximiser la dispersion des unités au sein d'un espace de visualisation.

6.2 L'effet *Loupe*

Nous présentons dans cette section les derniers travaux qui ont été faits dans le but de trouver une représentation qui maximise la dispersion des unités au sein d'un espace de visualisation, ce que nous appelons un effet *loupe*. Plusieurs méthodes ont été étudiées, mais les résultats obtenus ne nous ont pas permis, pour l'instant, de définir la méthode la plus adaptée. Nous ne donnerons donc qu'un aperçu des méthodes étudiées.

En premier lieu, trois approches ont été considérées :

- **ACP** : nous avons regardé le résultat d'une Analyse par Composantes Principales (ACP) sur la classe elle-même (voir figure 6.6). Le résultat est très intéressant car d'une part les unités sont bien éclatées dans tout l'espace de visualisation, et d'autre part on peut toujours visualiser par la couleur l'évolution d'un autre descripteur au sein de la classe. L'ACP va d'ailleurs trouver un espace spécifique à la classe considérée. Ce point peut être un avantage mais aussi un inconvénient, puisqu'en passant d'une classe à une autre on changera aussi d'espace de représentation, ce qui risque de perturber une navigation intuitive.
- **SOM** : *self-organizing map* ou carte d'auto-organisation de Kohonen. L'intérêt de cette méthode est d'avoir un espace continu qui essaye de représenter les différences entre les unités sonores. Dans notre cas l'hétérogénéité de la base de données sonores n'a pas permis d'obtenir une représentation correcte en utilisant un paramètrage simple de la méthode.
- **MDS** : *Multidimensional Scaling* ou Positionnement Multidimensionnel. Cet algorithme est un dérivé de l'ACP, qui utilise des matrices de distance entre des points et non les coordonnées des points. Nous avons souhaité utiliser l'algorithme MDS pour forcer une distance minimale entre deux points. Les résultats obtenus sur un ensemble très petit de points ont été plutôt convaincants. Cependant, en forçant une distance minimale, on crée des aberrations dans la matrice de distance. La version non métrique de MDS permet de s'affranchir de cette difficulté mais alors les temps de calcul deviennent prohibitifs. Dans notre cas, nous n'avons jamais réussi à avoir de résultat pour plus de 150 points, même au bout de 24 heures de calcul.

Les trois approches présentées ci-dessus ont l'inconvénient majeur de ne pas pouvoir contrôler le facteur d'éclatement des points. Nous nous sommes donc tournés vers l'utilisation de *sigmoïdes*. Cette famille de fonction produit une *courbe sigmoïde*, c'est-à-dire une courbe qui a la forme d'un «S», tel que représenté dans la figure 6.7 et définie ainsi :

$$P(t) = \frac{1}{1 + \exp^{-t}}$$

La figure 6.8 représente les unités d'une classe dans un espace éclaté. La fonction de type sigmoïde utilisée pour cette représentation est appelée *fonction d'erreur*. C'est une distribution gaussienne intégrée deux fois avec une moyenne égale à 0 et une variance égale à 1/2. Elle est définie par :

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp^{-t^2} \, dt$$

La fonction sigmoïde a été appliquée sur chacun des axes (les composantes principales). En faisant varier la valeur de la moyenne et celle de la variance, on peut contrôler la forme

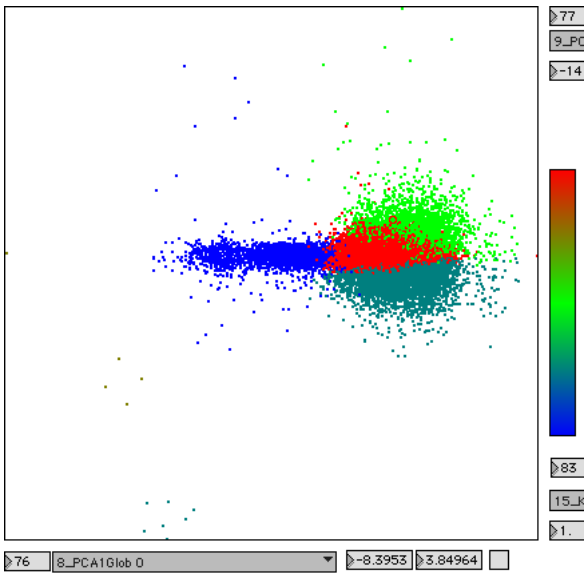


FIG. 6.1 – Représentation des sons de la base selon les deux premières composantes principales. La couleur montre la classification des unités en 5 classes. Le petit rond rouge correspond à l'unité qui nous intéresse.

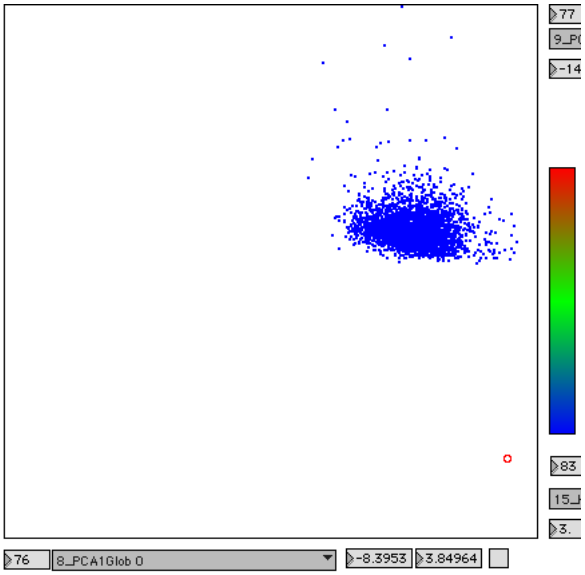


FIG. 6.2 – On voit à présent uniquement la classe de l'unité choisie, toujours dans une représentation en composantes principales.

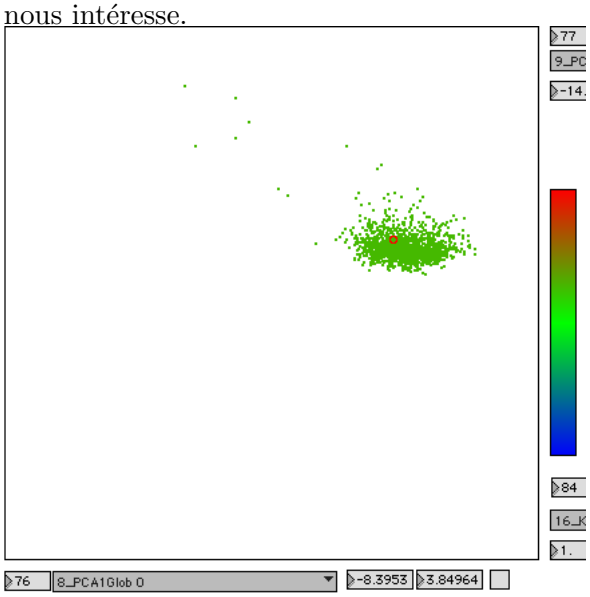


FIG. 6.3 – Avec une classification plus fine (ici 12 classes), on réduit le nombre d'unités similaires. La représentation est toujours en composantes principales.

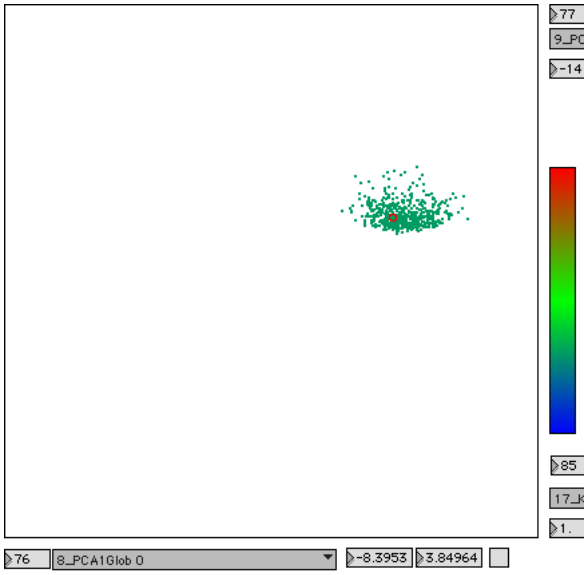


FIG. 6.4 – On augmente encore un peu plus la finesse de la classification en 24 classes. Le nombre d'unités similaire baisse encore. La représentation est toujours en composantes principales.

de la sigmoïde et donc éclater plus ou moins les points qui sont centrés en 0, c'est-à-dire autour du centroïde de la classe d'unités considérées.

Conclusion

L'utilisation de l'*Analyse par Composantes Principales* et de la *classification Kmeans* nous a permis de définir des stratégies de navigation et de visualisation à la fois efficaces et simples. À présent la sélection des unités sonores pour la synthèse concaténative peut se faire par deux accès complémentaires, chacun correspondant à un degré de connaissance des propriétés sonores des unités :

- Si l'utilisateur connaît bien à quoi correspond la signification des descripteurs sonores, il peut visualiser les unités en sélectionnant le descripteur qui l'intéresse.
- Si l'utilisateur ne souhaite pas considérer un descripteur en particulier mais cherche un espace de navigation et de visualisation intuitif, il peut sélectionner une visualisation par composantes principales, ce qui revient à visualiser les unités dans un espace qui *explique* le mieux les unités représentées. L'utilisateur naviguera alors dans un espace à deux dimensions qui est une combinaison de plusieurs descripteurs, sans pour autant savoir quels sont ces descripteurs.

De plus, si l'utilisateur s'intéresse à une unité sonore en particulier, il a à présent la possibilité de n'afficher que les unités les plus semblables à celle-ci avec plusieurs degrés d'affinage, en gardant les deux accès mentionnés ci-dessus. Ensuite, afin de faciliter la navigation et d'exploiter au mieux un ensemble d'unités, l'utilisateur pourra appliquer un effet de *loupe* pour bien distinguer chaque unité.

On remarquera enfin que grâce à la réduction de la complexité des informations contenues dans la base de données sonores, l'élaboration d'un mapping entre le moteur de synthèse et le contrôleur gestuel considéré est grandement facilité et paraît beaucoup plus intuitif. Ici par exemple, les mappings utilisés pour contrôler la synthèse sonore sont des mapping *one-to-one*, comme pour la position du curseur de sélection des unités (voir section 3.1) et le contrôle de la navigation par classe (voir ci-dessus). Un accès aux unités par des caractéristiques plus globales que celles offertes par les descripteurs offre des perspectives d'utilisation très séduisantes, comme nous le verrons dans le prochain chapitre.

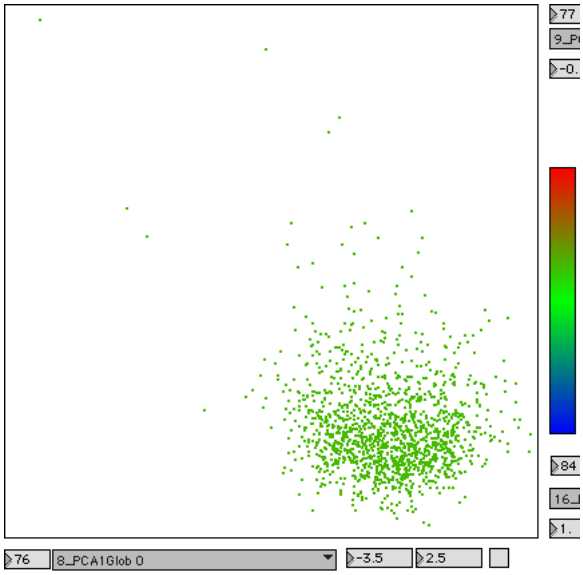


FIG. 6.5 – Voici la même classe que dans la figure 6.3, mais l'espace de visualisation est centré sur la classe. L'éclatement des unités permet une meilleure visualisation et une meilleure navigation.

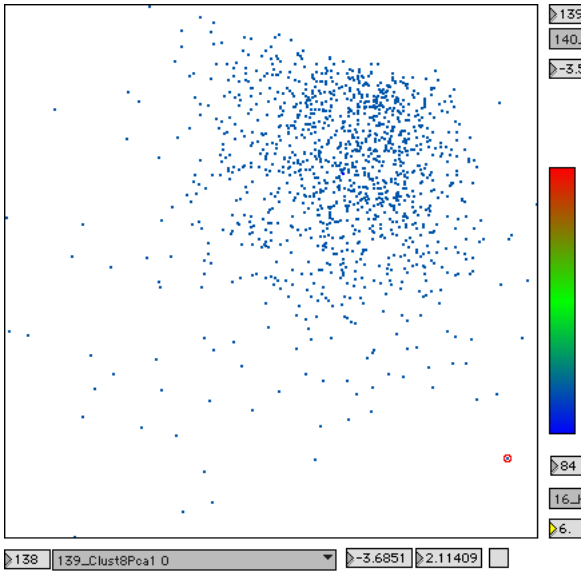


FIG. 6.6 – Voici la classe même classe que dans la figure 6.3. La représentation est toujours en composantes principales, mais celles-ci ont été calculées en ne considérant que la classe choisie.

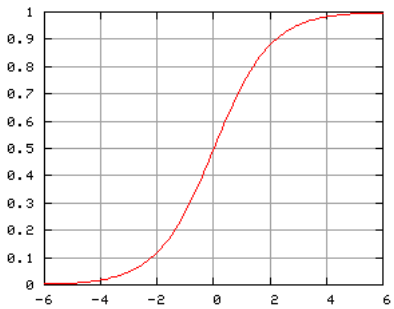


FIG. 6.7 – Une des fonctions de la famille des sigmoïdes.

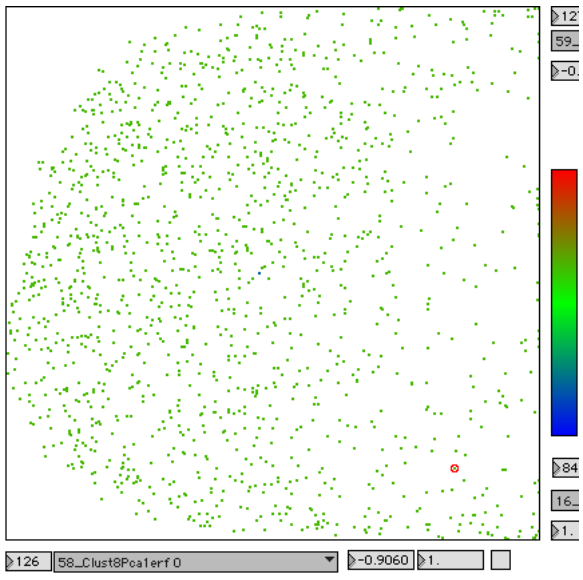


FIG. 6.8 – Voici la classe même classe que dans la figure 6.6. La représentation est toujours en composantes principales, calculées en ne considérant que la classe choisie, mais les points ont été éclatés à l'aide d'une sigmoïde.

Chapitre 7

Perspectives et travaux futurs

*Nous avons montré dans les chapitres précédents que l'utilisation de méthodes statistiques d'exploration de données comme l'*Analyse par Composantes Principales *et la* classification Kmeans *permet de définir des stratégies de navigation et de visualisation d'une base de données sonores efficaces et simples, tout en facilitant l'élaboration d'un* mapping *entre le moteur de synthèse sonore concaténative et un contrôleur gestuel. Nous allons exposer dans ce chapitre les perspectives et les travaux futurs qui s'inspirent des résultats obtenus jusqu'à présent. Une première partie traitera du problème du passage d'une classe à une autre dans le cadre d'une navigation intuitive. Une deuxième partie s'intéressera aux avantages et aux inconvénients qui résulteraient d'une augmentation du nombre de descripteurs considérés pour alimenter les méthodes d'exploration de données présentées. Enfin une troisième partie proposera un modèle de mapping général basé sur l'enchaînement temporel de classes d'unités sonores.*

7.1 Passage continu d'une classe à une autre

En utilisant des descripteurs sonores perceptifs, l'ACP nous offre une représentation intuitive qui maximise sur chaque axe la variabilité entre les unités. On souhaiterait donc trouver un moyen de représenter toutes les classes d'unités sonores les unes à côté des autres dans un espace à deux dimensions, tout en gardant une continuité entre chaque classe. La figure 7.1 permet de visualiser ce concept.

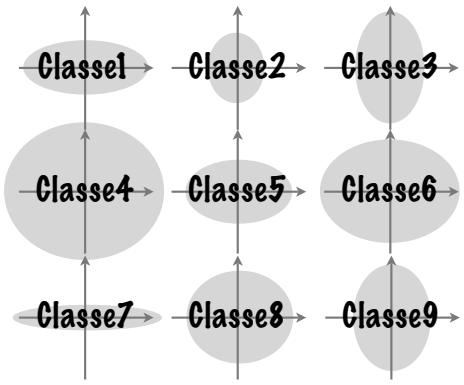


FIG. 7.1 – Cette figure représente un espace à deux dimensions dans lequel sont visualisées 9 classes d'unités côte à côte.

Grâce à l'ACP, on peut *expliquer* chaque classe d'unités en regardant la contribution de chacun des descripteurs sonores. Cette information est donnée par l'analyse de la matrice de corrélation variables-facteurs comme celle représentée par le tableau 4.6. Ainsi, dans cet exemple, on en déduit qu'une unité sonore placée à l'extrémité positive de l'axe formé par la première composante principale a une valeur positive et forte pour le centroïde spectral, l'IP du centroïde spectral et la variation spectrale. A contrario, sur l'axe formé par la deuxième composante principale, une unité placée à l'extrémité positive aura une valeur positive et forte pour la dispersion spectrale, mais négative et forte pour l'IP du volume.

Pour passer de façon continue d'une classe à une autre, l'idée serait donc de trouver un agencement des espaces formés par l'ACP de chaque classe qui garde la continuité des valeurs d'au moins un descripteur sur chaque axe. La figure 7.2 illustre un exemple d'agencement pour quatre classes. Puisque la sélection des unités se fait grâce à la position du curseur de sélection dans le graphique de visualisation, on aurait alors un moyen de passer de façon continue d'une classe à l'autre, chacune étant visualisée par sa propre représentation en composantes principales. On remarquera d'ailleurs que le sens de visualisation de la classe n'a pas d'importance, et qu'afin de trouver le meilleur agencement, on pourrait même effectuer des rotations sur les espaces tel que décrit dans la figure 7.2.

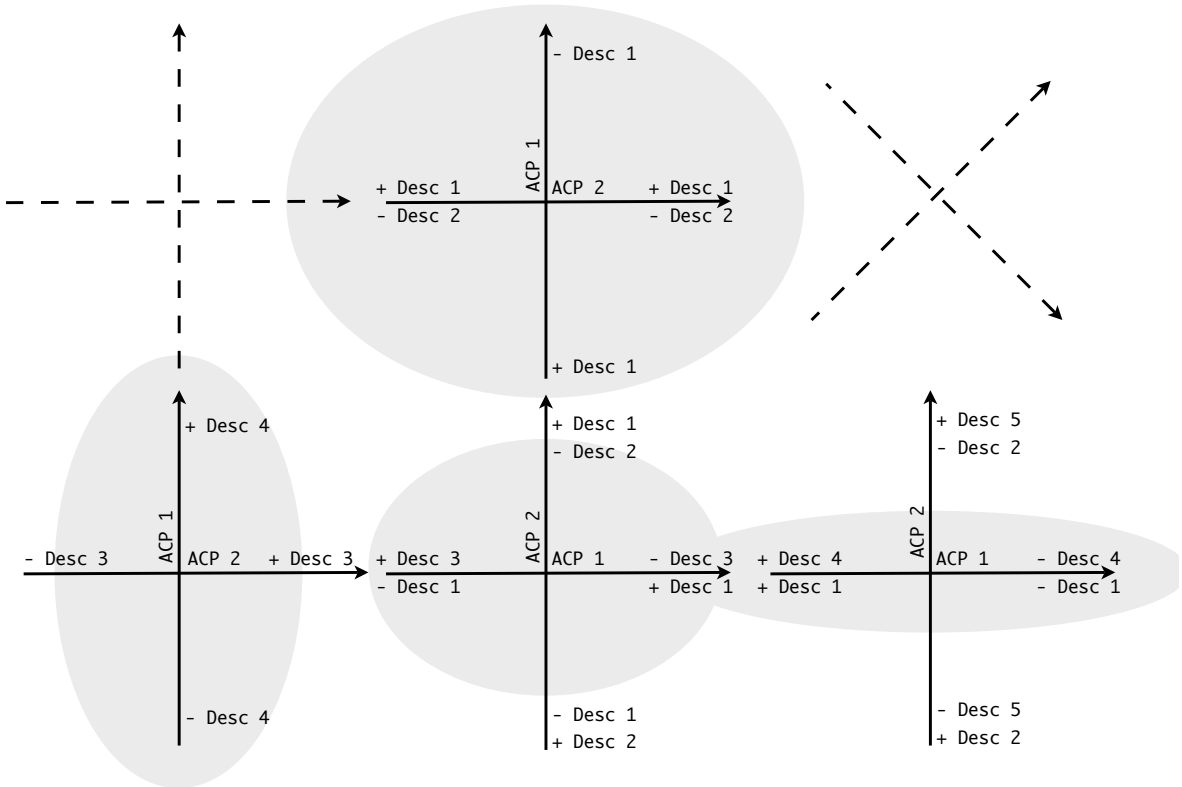


FIG. 7.2 – Cette figure représente un espace à deux dimensions dans lequel sont visualisées 5 classes d'unités côte à côte. Chaque classe est visualisée dans sa propre représentation en composantes principales. L'agencement des classes garde une continuité sur au moins un descripteur. Ici la vue est simplifiée puisque les composantes principales ne sont pas forcément orthogonales.

7.2 Augmentation du nombre de descripteurs

Nous n'avons considéré dans notre étude qu'un tout petit ensemble des descripteurs sonores disponibles pour alimenter les méthodes d'exploration de données présentées. L'augmentation du nombre de descripteurs utilisés présente l'avantage évident d'apporter des informations supplémentaires pour caractériser les unités sonores. On pourrait ainsi obtenir une meilleure classification des sons, peut-être d'ailleurs à l'aide de descripteurs qui n'ont pas forcément des propriétés perceptives, mais qui permettraient de bien distinguer les classes d'unités. Cependant l'utilisation de descripteurs non perceptifs pourraient donner, après une ACP, des espaces de navigation qui n'ont aucun sens perceptivement. De plus, on rappelle que certains descripteurs n'ont aucun sens selon la nature de l'unité sonore (par exemple, provenant d'un son harmonique ou non). Enfin ajoutons que l'utilisation de valeurs caractéristiques permet de mieux décrire l'évolution temporelle d'un descripteur au sein d'une unité, mais que dans ce cas la longueur de l'unité devient un paramètre très important pour la pertinence des valeurs obtenues.

Les stratégies envisagées pour augmenter le nombre de descripteurs sans avoir les inconvénients d'un surplus d'information sont :

- choisir automatiquement des sous-ensembles particuliers de descripteurs en fonction de la nature du son et de sa taille ;
- donner des poids aux descripteurs en fonction des stratégies de visualisation et de navigation choisies.

7.3 Stratégie de mapping basée sur l'enchaînement temporel de classes d'unités sonores

L'idée sous-jacente d'une stratégie de mapping basée sur l'enchaînement temporel de classes d'unités sonores est d'associer un geste au déroulement temporel d'un son tout en cherchant à interpréter la variabilité du geste comme une variation sonore.

Voici un scénario simple qui illustre cette application :

1. On demande à l'utilisateur d'écouter une première fois un son.
2. On demande ensuite à l'utilisateur de faire un geste qui décrit pour lui l'évolution du son, pendant qu'il entend une deuxième fois le son. Le geste est par exemple enregistré à l'aide d'une tablette graphique.
3. Enfin on demande à l'utilisateur de refaire son geste, et on resynthétise un son en exprimant dans le résultat sonore la variation entre le geste de référence et celui de resynthèse.

Ce scénario implique d'une part de trouver un moyen de relier le déroulement temporel du geste à celui du son, et d'autre part de trouver une stratégie de mapping qui permette d'exprimer de façon intuitive la variation du geste en variation sonore.

Pour cela, deux modèles devront être réalisés puis testés :

- **Approche par les paramètres gestuels** : On considérerait dans ce cas les variations du geste d'une part par rapport aux deux dimensions de la tablette graphique, et d'autre part par rapport aux paramètres temporels sur ces deux dimensions (vitesse, accélération).
- **Approche par l'utilisation de Modèles de Markov Cachés (HMM)** : (en anglais, *Hidden Markov Model*). L'idée est d'une part d'utiliser un modèle de Markov pour reconnaître un geste et d'autre part de caractériser temporellement un son par la

succession de classes auxquelles appartiennent les unités du son. On souhaite ainsi relier chaque état du modèle de Markov à une classe d'unités sonores. La variation gestuelle est exprimée par la probabilité d'être dans un état plutôt qu'un autre. On relie cette probabilité à la probabilité d'être dans une classe d'unités sonores ou une autre. Enfin la variation sonore serait exprimée par le choix d'une unité sonore plus ou moins distante de l'unité de référence, au sein d'une même classe. Ainsi on définirait comment passer d'une classe d'unités sonores à une autre et comment interpréter la variabilité au sein d'une classe. Prenons l'exemple d'un geste de resynthèse très similaire au geste de référence, mais rejoué plus lentement. Dans ce cas, tant que le modèle de Markov reste dans un état, le moteur de synthèse sélectionnera automatiquement dans la classe d'unités correspondante, des unités semblables à l'unité de référence. Lorsque le modèle de Markov changera d'état, le moteur de synthèse considérera alors la classe d'unités suivante temporellement. On aura alors un contrôle à la fois sur le déroulement temporel et sur la variabilité des caractéristiques sonores.

Conclusion

Nous avons introduit dans ce rapport des *stratégies de navigation et de visualisation d'une base de données sonores*. En nous basant sur un modèle général de *mapping multicouche* et en étudiant le fonctionnement de la *synthèse sonore* dite *concaténative*, nous avons montré que le principal paramètre de contrôle de la synthèse concaténative porte sur la sélection des unités sonores qui constituent la base de données, auxquelles on accède par leurs descripteurs sonores. La complexité des informations que fournissent ces descripteurs nous a amené à orienter notre étude sur l'utilisation de méthodes statistiques d'exploration de données pour obtenir une sélection efficace et intuitive des unités sonores. Deux méthodes ont été étudiées, l'*Analyse par Composantes principales* et l'algorithme de classification *Kmeans*. Pour valider l'utilisation de telles méthodes nous avons comparé les résultats obtenus sur des sous-ensembles de sons de plus en plus complexes. Nous avons choisi pour ce faire des sons environnementaux, décrits par une petite sélection de descripteurs perceptivement compréhensibles. Enfin, nous avons montré que ces méthodes ont permis d'une part de proposer des stratégies de visualisation et de navigation efficaces et intuitives et d'autre part de faciliter l'élaboration d'un mapping simple entre le moteur de synthèse sonore et le contrôleur considéré. En effet, comme l'a fait remarquer M. Wanderley concernant l'élaboration d'un instrument électronique de musique, meilleure est la conception de l'instrument, plus simple sera le mapping pour le contrôler. Remarquons cependant que l'application de la synthèse concaténative considérée dans notre étude a été celle de la synthèse libre que l'on peut qualifier de synthèse *explorative*, et qu'elle n'utilise qu'une partie des possibilités que propose la synthèse concaténative.

Les résultats obtenus trouvent un domaine d'application assez large puisqu'ils permettent enfin à un utilisateur néophyte de naviguer de façon intuitive dans une base de données sonores. La synthèse sonore concaténative est une technologie récente qui offre de nouveaux horizons pour de nombreuses applications comme le design sonore ou la composition musicale. Il reste cependant beaucoup de recherche à faire pour proposer des modèles pertinents permettant d'exploiter de grandes bases de données sonores très hétérogènes, surtout dans le cadre d'une utilisation en temps-réel.

Bibliographie

[ACKV02] D. Arfib, J. M. Couturier, L. Kessous, and V. Verfaille, *Strategies of mapping between gesture data and synthesis model parameters using perceptual spaces*, Organised Sound **7** (2002), no. 2, 127–144.

[AP05] Jean-Julien Aucouturier and François Pachet, *Ringomatic: A real-time interactive drummer using constraint-satisfaction and drum sound descriptors.*, ISMIR, 2005, pp. 412–419.

[BB05] Alain Bacini and Philippe Besse, *Data minig 1 - exploration statistique*, Publication du Laboratoire de Statistique et Probabilités (2005).

[BL] Michel Beaudouin-Lafon, *Interaction homme-machine - cours d'ihm pour l'ens.*

[Bon00] Bert Bongers, *Physical interaction in the electronic arts: Interaction theory and interfacing techniques for real-time performance*, Trends in Gestural Control of Music, Ircam, 2000.

[BSHR05] Grégory Beller, Diemo Schwarz, Thomas Hueber, and Xavier Rodet, *Hybrid concatenative synthesis in the intersection of speech and music*, JIM **12** (2005), 41–45.

[CA03] Jean-Michel Couturier and Daniel Arfib, *Pointing fingers: using multiple direct interactions with visual objects to perform music*, NIME '03: Proceedings of the 2003 conference on New interfaces for musical expression (Singapore, Singapore), National University of Singapore, 2003, pp. 184–187.

[CBG95] Insook Choi, Robin Bargar, and Camille Goudeseune, *A manifold interface for a high dimensional control space*, 1995, pp. 385–92.

[CCH04] Arshia Cont, Thierry Coduys, and Cyrille Henry, *Real-time gesture mapping in pd environment using neural networks.*, NIME, 2004, pp. 39–42.

[CCH05] ———, *Augmented mapping: toward an intellignet user-defined gesture mapping*, 2005.

[Cho00] Insook Choi, *Gestural primitives and the context for computational processing in an interactive performance system*, Trends in Gestural Control of Music, Ircam, 2000.

[Cho03] I. Choi, *A component model of gestural primitive throughput*, 2003.

[CPRV96] Sergio Canazza, Giovanni De Poli, Stefano Rinaldin, and Alvise Vidolin, *Sonological analysis of clarinet expressivity.*, in Leman [Lem97], pp. 431–440.

[CPV96] Sergio Canazza, Giovanni De Poli, and Alvise Vidolin, *Perceptual analysis of the musical expressive intention in a clarinet performance.*, in Leman [Lem97], pp. 441–450.

[CV04] Antonio Camurri and Gualtiero Volpe, *Gesture-based communication in human-computer interaction, 5th international gesture workshop, gw 2003, genova, italy, april 15-17, 2003, selected revised papers*, Gesture Workshop, Lecture Notes in Computer Science, vol. 2915, Springer, 2004.

[CW00] Claude Cadoz and Marcelo M. Wanderley, *Gesture-music*, Trends in Gestural Control of Music, Ircam, Paris, France, 2000, pp. 1–55.

[Del88] François Delalande, *La gestique de glenn gould*, pp. 84–111, Louise Courteau Editrice, 1988.

[FH98] S. Fels and G. Hinton, *Glove-talkii: A neural network interface which maps gestures to parallel formant speech synthesizer controls*, 1998.

[GKP04] S. Gibet, J.F Kamp, and F. Poirier, *Gesture analysis: Invariant laws in movement*, Gesture-based Communication in Human-Computer Interaction, 5th International Gesture Workshop (G. Volpe A. Camurri, ed.), LNAI 2915, Springer Verlag, 2004, pp. 1–9.

[Got00] Suguru Goto, *Virtual musical instruments: Technological aspects and interactive performance issues*, Trends in Gestural Control of Music, Ircam, 2000.

[Gou02] Camille Goudeseune, *Interpolated mappings for musical instruments*, Organised Sound **7** (2002), no. 2, 85–96.

[Hue05] Thomas Hueber, *Talkapillar système d'analyse, de synthèse et de transformation de la parole à partir du texte*, 2005.

[HWK00] Andy Hunt, Marcelo M. Wanderley, and Ross Kirk, *Towards a model for instrumental mapping in expert musical interaction*, ICMC: International Computer Music Conference (Berlin, Allemagne), Septembre 2000.

[HWP02] A. Hunt, M. Wanderley, and M. Paradis, *The importance of parameter mapping in electronic instrument design*, 2002.

[KA03] Loic Kessous and Daniel Arfib, *Bimanuality in alternate musical instruments*, NIME '03: Proceedings of the 2003 conference on New interfaces for musical expression (Singapore, Singapore), National University of Singapore, 2003, pp. 140–145.

[Lem97] Marc Leman (ed.), *Music, gestalt, and computing - studies in cognitive and systematic musicology*, Lecture Notes in Computer Science, vol. 1317, Springer, 1997.

[LPY04] Mauricio A. Loureiro, Hugo B. De Paula, and Hani C. Yehia, *Timbre classification of a single musical instrument.*, ISMIR, 2004.

[LVV+03a] M. Leman, V. Vermeulen, L. De Voogdt, A. Camurri, B. Mazzarino, and G. Volpe, *Relationship between musical audio, perceived qualities, and motoric responses - a pilot study*, Proc. International Stockholm Acoustic Conference 2003 (SMAC03) (Stockholm, Sweden) (R. Bresin, ed.), August 2003.

[LVV03b] A. Likas, N. Vlassis, and J. Verbeek, *The global k-means clustering algorithm*, 2003.

[MM99] Stephen McAdams and Nicolas Misdariis, *Perceptual-based retrieval in large musical sound databases*, Human Centred Processes '99 (Brest) (P. Lenca, ed.), 1999, pp. 445–450.

[MMS03] Paul Modler, Tony Myatt, and Michael Saup, *An experimental set of hand gestures for expressive control of musical parameters in realtime.*, NIME, 2003, pp. 146–150.

[Mod00] Paul Modler, *Neural networks for mapping hand gesture to sound synthesis parameters*, Trends in Gestural Control of Music, Ircam, 2000.

[MSP+98] Nicolas Misdariis, Bennett K. Smith, Daniel Pressnitzer, Patrick Susini, and Stephen McAdams, *Validation of a multidimensional distance model for perceptual dissimilarities among musical timbres*, ICA and ASA joint meeting (Seattle, USA), vol. 103, Juin 1998.

[MW03] A. Momeni and D. Wessel, *Characterizing and controlling musical material intuitively with graphical models*, 2003.

[NWD04] D. Van Nort, M. Wanderley, and P. Depalle, *the choice of mappings based on geometric properties*, 2004.

[PD04] Marcelo M. Wanderley Philippe Depalle, *Gestural control of sound synthesis*, Proceedings of the IEEE, vol. 92, 2004.

[Pee04] Geoffroy Peeters, *A large set of audio features for sound description (similarity and classification)*, Tech. report, 2004.

[PMH00] Geoffroy Peeters, Stephen McAdams, and Perfecto Herrera, *Instrument sound description in the context of mpeg-7*, ICMC: International Computer Music Conference (Berlin, Allemagne), Septembre 2000, pp. 166–169.

[RWDD97] Joseph B. Rovan, Marcelo M. Wanderley, Shlomo Dubnov, and Philippe Depalle, *Instrumental gestural mapping strategies as expressivity determinants in computer music performance*, KANSEI - The Technology of Emotion (Genes, Italie) (Antonio Camurri, ed.), Octobre 1997, pp. 68–73.

[SBVB06] Diemo Schwarz, Grégory Beller, Bruno Verbrugghe, and Sam Britton, *Real-Time Corpus-Based Concatenative Synthesis with CataRT*, September 2006.

[Sch00] Diemo Schwarz, *A system for data-driven concatenative sound synthesis*, 2000.

[Sch04] Diemo Schwarz, *Data-driven concatenative sound synthesis*, Thèse de doctorat, Université Paris 6 - Pierre et Marie Curie, Paris, 2004.

[Sch05] Diemo Schwarz, *Current research in concatenative sound synthesis*, International Computer Music Conference (ICMC) (Barcelona, Spain), Septembre 2005.

[Sch06] ———, *Concatenative sound synthesis: The early years*, Journal of New Music Research **35** (2006), no. 1, 3–22, Special Issue on Audio Mosaicing.

[SM95] Gredory Sandell and William Martens, *Perceptual evaluation of principal-componenet-based synthesis of musical timbre*, Audio Engineering Society **43** (1995), no. 12.

[TB05] Slaney M. Terasawa, H. and J Berger, *Perceptual distance in timbre space*, Proceedings of International Conference on Auditory Display (ICAD05) Limerick, Ireland, 2005.

[TDW03] Caroline Traube, Philippe Depalle, and Marcelo Wanderley, *Indirect acquisition of instrumental gesture based on signal, physical and perceptual information*, NIME '03: Proceedings of the 2003 conference on New interfaces for musical expression (Singapore, Singapore), National University of Singapore, 2003, pp. 42–47.

[Ver94] R. Vertegaal, *An evaluation of input devices for timbre space navigation*, 1994.

[WD99] Marcelo M. Wanderley and Philippe Depalle, *Controle gestuel de la synthese sonore*, Interfaces Homme-Machine et Création Musicale, Hermès Science Publications, Paris, France, 1999, pp. 145–164.

[Wes78] David L. Wessel, *Timbre space as a musical control structure [low dimensional control of musical timbre]*, Ircam, Paris, France, 1978.

[WS02] Ipke Wachsmuth and Timo Sowa, *Gesture and sign languages in human-computer interaction, international gesture workshop, gw 2001, london, uk, april 18-20, 2001, revised papers*, Gesture Workshop, Lecture Notes in Computer Science, vol. 2298, Springer, 2002.

[WSR98] Marcelo M. Wanderley, Norbert Schnell, and Joseph B. Rovan, *Escher - modeling and performing composed instruments in real-time*, IEEE Int. Conference on Systems Man and Cybernetics (San Diego - CA, USA), Octobre 1998.