Baptiste Caramiaux
Parcours A.T.I.A.M.
Supervisé par **Norbert Schnell**
7 Mars - 10 Juillet 2008

« Gestification » du son : mapping adaptatif geste/son dans un contexte d'écoute et de performance musicale

Master 2 Recherche de l'**U**niversité **P**ierre et **M**arie **C**urie
Mention Informatique, spécialité **SAR** (Systèmes et Applications Réparties)
En collaboration avec **ParisTech** et l'**IRCAM - Centre Pompidou**

Résumé

Notre travail de stage présente une nouvelle approche dans la correspondance entre geste et synthèse sonore (ou *mapping*). Les stratégies de mapping abondent dans la littérature scientifique et peuvent schématiquement être regroupées en deux catégories principales : utilisation d'un mécanisme génératif, ou utilisation d'une correspondance explicite entre paramètres gestuels et paramètres sonores. Notre étude propose une nouvelle approche dans l'élaboration d'un mapping entre geste et son, utilisant un mécanisme statistique génératif pertinent afin de mettre en relation deux domaines distincts, que sont le geste et le son. Notre conviction est qu'il existe une relation fondamentale entre ces entités.

La première étape consistera à identifier leurs similarités par une analyse canonique (ou CCA, *Canonical Correlation Analysis*). Cette méthode engendre deux espaces formels qui représentent la projection des données gestuelles (ou données explicatives) et la projection des données sonores (ou données à expliquer) et son objectif est la minimisation de la distance entre ces deux espaces théoriques. Les nouvelles représentations des données, ou composantes canoniques, seront alignées en temps grâce à un algorithme de programmation dynamique. Enfin on prendra soin de valider les différentes étapes et de montrer la pertinence du mapping obtenu.

Remerciements

J'estime que mon premier remerciement doit aller à mon maître de stage Norbert Schnell, tout d'abord pour m'avoir fait confiance en me proposant de travailler sur ce sujet de stage, et ensuite pour tous les conseils, les discussions agréables et les opportunités qu'il m'a apportés. Ensuite, je voudrais remercier Frédéric Bevilacqua pour l'intérêt porté à mon travail.

L'équipe IMTR de l'IRCAM a cet esprit aventureux et créatif, donnant corps à un formidable dynamisme dans lequel on s'abandonne avec envie.

J'aimerais donc remercier toute la brave équipe IMTR et plus précisément mes compagnons de bureau, à savoir Jean-Philippe Lambert, sans lequel mes patchs Max ne seraient pas ce qu'ils sont, Arshia Cont pour ses conseils sur la langue anglaise et Fabrice Guédy pour ses conseils et sa bonne humeur.

Mes remerciements resteraient incomplets sans mentionner mes chers camarades ATIAM, et tout particulièrement (car le particulier naît du groupe) nos deux expatriés Gonçal Calvo i Perez et Julien Junod, sans oublier Gaetan, Marc, Romain, Lise, Maxime, Tifanie, Sarah, Sophie, Emilien, Nkam et notre cher disparu Jean-Yves.

Table des matières

Table des figures

1 Introduction

« Everything is related with every other thing, and this relation involves the emergence of a relational quality. The qualities cannot be known a priori, though a good number of them can be deduced from certain fundamental characteristics »

- Philosophie Jaïne

« he task is not so much to see what no one yet has seen, but to think what no body yet has thought about that which everyone sees. »

- Arthur Schopenhauer

Il s'avère indiscutable que le XX$^{\text{ème}}$ siècle fut un siècle de changements. Dans le domaine musical, divers esthétiques virent le jour au gré des impulsions créatrices induites par le contexte social et politique d'une part et les avancés de la technique d'autre part. Les connaissances scientifiques et particulièrement le développement de la physique a permis de mieux comprendre l'onde acoustique et le phénomène ondulatoire a pu être observé, stocké, synthétisé. C'est au lendemain de la seconde guerre mondiale, l'après 1945, qu'on parle d'ère numérique : les sons peuvent dorénavant être générés par l'ordinateur sans qu'un instrumentiste ait besoin de transmettre une énergie mécanique au système.

En désolidarisant la cause et l'effet dans la création musicale, la technologie numérique boulverse autant le jeu de l'instrumentiste que la composition musicale. Les méthodes usuelles de composition se muent en processus qui schématisent le comportement de l'oeuvre, ou d'une de ses parties, dans le temps. L'ordinateur possède ce potentiel d'effectuer un aplanissement de la ligne du temps dans laquelle le compositeur peut se promener à son gré. Dans le domaine du jeu instrumental, l'apprivoisement de l'instrument virtuel diffère sur de nombreux aspects de la technicité classique applicable dans le cas des instruments acoustiques. D'un côté il y a le concret, le mécanique, le palpable, l'instrument acoustique dont les caractéristiques mécaniques de vibration sont intimement liées, aux caractéristiques mécaniques du corps de

l'instrumentiste. Ses gestes auront un effet immédiat sur la qualité sonore. De l'autre côté, il y a l'univers numérique où les sons sont créés par des composants invisibles, et l'agencement de ces derniers peut se faire suivant un nombre de configurations incalculable.

Dans le contexte du nouveau jeu instrumental, l'ordinateur devient analogue à une « boîte noire »faisant le lien entre les gestes humains, par l'intermédiaire d'un contrôleur, et le son sortant sur des haut-parleurs. Pourtant malgré la formidable diversification de comportements programmables, il n'y en existe pas encore assurant une telle précision, une telle subtilité et une telle ampleur de jeu que celle des instruments acoustiques.

De nombreuses connaissances ont été rassemblées autour de la définition d'un comportement gestuel en situation de jeu, celles-ci provenant des sciences mathématiques, cognitives, biologiques, informatiques ou encore linguistiques. La liste n'est pas exhaustive mais donne un aperçu de la transversalité de la problématique.

L'étude présentée dans ce rapport s'inscrit dans une volonté d'améliorer les connaissances du jeu instrumental dans le contexte des instruments virtuels. Elle propose une nouvelle approche dans l'élaboration d'une correspondance entre les gestes et la génération du son. Celle-ci est communément appelée *mapping* et on adoptera dès à présent cette dénomination. À mi-chemin entre écoute et performance, l'étude met en exergue le fait qu'un geste tracé lors de la diffusion d'un son enregistré n'est pas arbitraire et cherche à représenter gestuellement et de manière subjective le son. La métaphore du son par le geste relève de l'abstrait ce qui nous amène à devoir définir un mapping à haut-niveau de contrôle.

Le travail présenté dans ce document a comme objectif l'élaboration de cette stratégie de mapping et de sa validation. Dans un contexte de jeu instrumental, le projet se basera sur le schéma de la figure [1.1] qui illustre la structure type de la lutherie des instruments virtuels. Il matérialise les différentes étapes à parcourir pour le développement du mapping. Il s'agira ensuite de valider la méthode proposée et de donner des pistes pour de futures avancées.

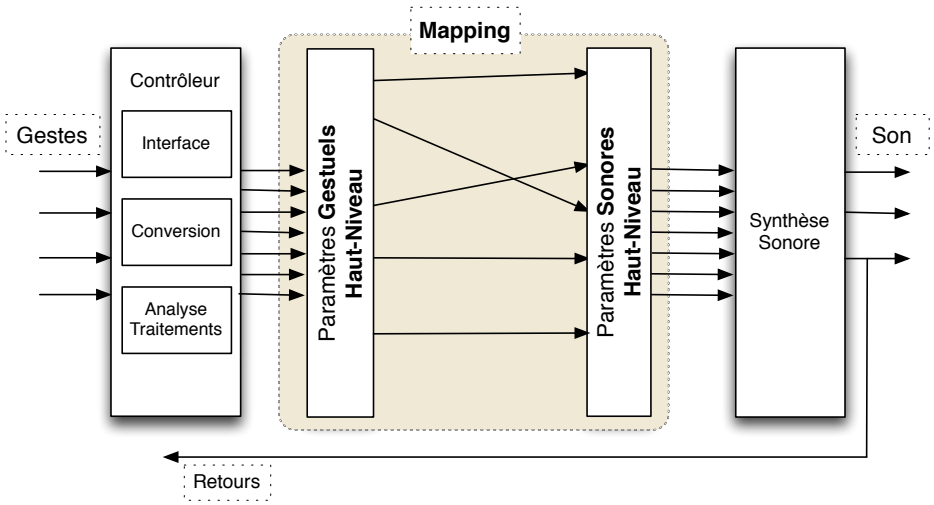


FIG. 1.1 – Schéma de la structure type d'un instrument virtuel

2
Études préliminaires

Comme le montre l'illustration introductive, le mapping se trouve au milieu d'un système plus général de lutherie d'instruments virtuels. L'étude de la correpondance entre geste musical et phénomène sonore possède une caractéristique transversale nous amenant à nous interroger sur divers problématiques telles que le geste, la saisie du geste ou encore la synthèse sonore. Ce chapître présente ces notions préliminaires et nécessaires à une étude sur le mapping. On illustrera ce paragraphe avec les outils disponibles au sein de l'équipe IMTR (Intéraction Musicale Temps Réel) de l'IRCAM.

2.1 Le Geste

La première étape inévitable dans la compréhension du contexte dans lequel s'inscrit l'étude, consiste à définir la notion de "geste". On se rendera rapidement compte qu'il est difficile, pour ne pas dire impossible, de donner une et une seule définition du geste. Pour s'en convaincre, on pourrait recueillir les propositions émanant de ses divers champs de modélisation, de reconnaissance ou encore d'analyse tels que la danse, la musique, le médical ou encore le sport. Les gestes y sont les briques élémentaires amenant à des finalités diverses, et là où on parlera de geste (e.g. instrumental), on utilisera conventionnellement le mot "mouvement" (e.g. dansé). Pourtant, une définition même subjective du geste reste une nécessité afin de pouvoir définir une typologie (relative à cette définition) sans laquelle, l'étude ne serait pas consistante. De là, un ensemble de paramètres gestuels pourront être définis, reflètant les hypothèses restrictives choisies dans le cadre du projet.

2.1.1 Vers une défintion du "geste" dans le domaine musical

Il n'est pas concevable d'aborder le "geste" de manière globale. Comme il a été signalé ci-dessus, il existe autant de définitions différentes, pertinentes mais spécifiques, qu'il existe de champs d'application du "geste". De fait, on se restreint au concept du geste dans le domaine musical. En première approche, (Iazzetta, Fernando 2000) définit le geste comme un mouvement qui exprime quelque chose, c'est à dire un mouvement porteur de sens. L'article

(Kurtenbach, G. and Hulteen, E.A. 1990) développe une définition similaire par le biais du concept plus précis qu'est *l'information* : « *A gesture is a motion of the body that contains information.* ». Les auteurs précisent leur point de vue par cet exemple : « *Waving goodbye is a gesture. Pressing a key is not a gesture because the motion of a finger on its way to hitting the key is neitheir observed nor significant. All that matters is which key was pressed* ». La précision dans le concept du geste apportée par les auteurs Kurtenbach et Hulteen amène inexorablement à une limitation. En effet, l'idée du geste est ainsi réduite sur ce qui est perceptible à la vision et durant le temps de son exécution (cf. (Cadoz, Claude and Wanderley, Marcelo M. 2000)). C'est à dire qu'un mouvement est appelé geste s'il exprime à lui seul une information (par exemple montrer du doigt une personne), alors qu'il ne sera pas un geste s'il n'est qu'un lien vers un contrôleur. Avec cette définition, le mouvement d'une personne jouant du violon ne serait pas un geste.

Le rôle du sens n'est pas de se dissoudre dans le concept d'information. Lorsqu'il est question de geste, il est aussi question d'émotions, de sentiments, d'intention, d'opinion, d'expression, qui constituent l'âme du geste. Bien que de telles notions forment une base instable pour un travail scientifique, on ne peut pas définir le geste sans elles. On se réfère à la définition simple d'un dictionnaire [1], fournissant une première base explicative de l'objet.

Geste : Mouvement du corps (principalement des bras, des mains, de la tête) volontaire ou involontaire, révélant un état psychologique ou visant à exprimer quelque chose. *Le geste moyen d'expression. «Les gestes de l'orateur sont des métaphores.»(Valéry)*

Le geste vise à exprimer quelque chose, et par là constitue un acte indissociable de sa sémantique. (Kurtenbach, G. and Hulteen, E.A. 1990) plaçait le rôle du sens au niveau du rôle informatif du geste. On préfèrera la définition apportée par (Hummels, Caroline et al. 1998) qui estime qu'un geste est un mouvement du corps qui transmet un message sensé à soi-même ou à un partenaire. Cela lui confère un but de "communication". Et l'article précise que ce "partenaire" englobe l'être humain comme le matériel informatique. Le sens est alors l'information qui contribue au but. On se basera sur cette définition en ajoutant que le mouvement effectué peut être, à notre sens, conscient ou inconscient : notre conviction est que certains gestes peuvent se délayer de la pensée consciente afin d'agir indépendamment.

Le contexte général de notre étude pose la problématique de l'intention dans le geste lorsque celui-ci est lié à une séquence sonore enregistrée. D'après les considérations précédentes, notre opinion est que l'intention en question

[1] Le Petit Robert, Dictionnaire alphabétique et analogique de la langue française, 1973

donne l'information sur ce qu'on trouve de significatif dans le son. L'information transmise porte la signification subjective de la séquence sonore sous la contrainte du langage gestuel.

La suite de l'étude nécessite l'élaboration d'une typologie afin de cibler quel geste nous devons prendre en compte dans l'ensemble des mouvements du corps.

2.1.2 Typologie du geste

La définition donnée au geste reste très générale, et il est nécessaire de délimiter différents niveaux : du plus concret au plus métaphorique. Dans (Delalande, François 1988), François Delalande définit trois niveaux différents, du plus fonctionnel au plus symbolique.

1. Le *Geste effecteur* est responsable de la production mécanique du son
2. Le *Geste accompagnateur* engage tout le corps de l'instrumentiste mais n'est pas responsable directement de la production du son. Ils peuvent être des mimiques, des gestes d'épaules, ...
3. Le *Geste figuré* est purement symbolique (représentation métaphorique des gestes précédents) qui est perçu par l'auditeur comme des articulations dans la musique

Comme on se focalise sur la correspondance geste/son, et sur les intéractions homme-machine, nous ciblerons notre travail autour des gestes effecteurs. C'est ce type de gestes que Claude Cadoz appelle *geste instrumental* (cf. (Cadoz, Claude 1988), (Wanderley, Marcelo M. et al. 1999)). Il décrit alors les trois fonctions du geste instrumental. Tout d'abord la fonction épistémique, qui se réfère au toucher pour acquérir de l'information, puis la fonction ergotique qui est la transformation des objets par action physique, et enfin la fonction sémiotique, ou communiquer de l'information. De ces trois fonctions, il propose une catégorisation en trois parties (cf. (Cadoz, Claude 1999)),

1. Le *Geste d'excitation* de type percusionnelle, continue ou entretenue
2. Le *Geste de modification*, structurelle ou paramétrique
3. Le *Geste de sélection*, séquentielle ou simultannée

Enfin, Marcello Wanderley ajoute une quatrième catégorie à celles proposées par Claude Cadoz : celle du geste de *maintien* ou de *polarisation* (cf. (Wanderley, Marcelo M. and Depalle, Philippe 2004)). Celui-ci assure le fonctionnement normal de l'instrument. Dans le cas d'une cornemuse, le bras effectue le maintien d'une pression suffisante pour le jeu de l'instrument. Cette catégorie de gestes se distingue des autres car elle constitue un préalable essentiel à leur existence et à leur signification.

2.2 L'acquisition du geste

Cette section aborde la question du choix des paramètres, et de manière logique, du choix des contrôleurs. Elle fait donc référence à ce qui se trouve en amont du mapping et interroge sur les choix à effectuer et sur les problèmes, pas toujours visibles, posés par le matériel.

Comme il n'existe aucune étude récapitulative sur les paramètres gestuels disponibles dans un contexte de jeu instrumental, chaque application doit définir son propre ensemble de paramètres. De même, chaque application induira un choix de contrôleurs adaptés. On se propose un petit survol des contrôleurs gestuels utilisés à l'IRCAM, au sein de l'équipe IMTR.

2.2.1 Quels contrôleurs pour l'utilisateur ?

On se place dans le contexte des interfaces homme-machine (IHM) pour définir quels outils choisir pour l'utilisateur. Dans le domaine musical, de récents travaux portent sur les interfaces pour le geste instrumental ainsi que sur les pratiques musicales. On peut citer Serge de Laubier et le "Méta-Instrument", la "reactable", la principe d'"enaction" ou encore des travaux où l'interface s'efforce de donner un retour perceptif à l'utilisateur telle qu'un bras haptique (cf (Lambert 2004)).

Dans ce paragraphe, on présentera une série de contrôleurs simples, aboutis, et sans retours perceptifs. Chaque contrôleur présenté ici est lié aux paramètres qu'il peut fournir pour une problématique de mapping. La pertinence de ces-derniers dans le cadre de notre étude sera analysée.

La souris

Un premier contrôleur simple est la souris. C'est un contrôleur naturel car tous les ordinateurs modernes en sont équipés. Une souris acquiert la position (x, y) dans un repère cartésien et envoie l'information à l'ordinateur.

Ces coordonnées définissent deux paramètres continus. En outre, une souris "trois boutons" possède deux boutons binaires (`on` / `off`) et une molette prenant des valeurs discrètes ou pouvant être cliquée. À première vue, la souris est un bon candidat car nous sommes accoutumés à sa prise en main. Un exemple en est l'utilisation quotidienne par les réalisateurs en informatique musicale. Cependant, la souris a un gros désavantage qui est son comportement non linéaire ayant pour conséquence l'impossibilité d'effectuer des calculs simples comme des dérivations.

Notre choix s'est donc dirigé naturellement vers un contrôleur très connu : le clavier maître MIDI.

Le clavier MIDI

Le clavier MIDI est un contrôleur qui a été beaucoup utilisé dans la musique assistée par ordinateur et ceci pour plusieurs raisons. Une de ces raisons est que la norme MIDI induit un protocole de communication (quasi) temps-réel utilisable pour des installations multi-machines (son, lumières, ...) et de manière très simple (par séquence d'événements).

Le clavier utilisé (celui sur la photographie) comporte 25 touches, et plus d'une dizaine de contrôleurs continus, démultipliant les possibilités d'une souris basique. Pourtant, ces paramètres ne sont pas satisfaisants pour l'étude menée. En effet, ils ne reflètent pas un geste réel, car il n'y a aucune notion de position, vitesse, accélération, etc ... Chaque paramètre est indépendant et nécessite un geste pour le contrôler. Dans un contexte de mapping, l'utilisation conventionnelle du clavier MIDI réside dans la mise en relation d'un contrôleur MIDI vers un paramètre de synthèse.

La tablette graphique

La tablette graphique est un outil de saisie de geste composé d'une surface active (la tablette) et d'un outil dans la main de l'utilisateur. Cet outil se présente communément sous la forme d'un stylet. La superficie de la tablette utilisée correspond au format papier A4.

L'application du stylet sur la surface active déclenche l'envoi des données à l'ordinateur. La surface active repère la position. Le stylet possède une mine permettant de capter la pression exercée, un bouton latéral configurable mais non utilisé, et une gomme à l'autre extrémité, configurable, mais non utilisée.

Un avantage important de la tablette graphique se trouve dans les données envoyées à l'unité centrale : contrairement à certains contrôleurs tels que la souris, les données seront faiblement bruitées. Les positions transmises à l'ordinateur peuvent être traitées (typiquement dérivées) sans obtenir des valeurs trop erronées. L'erreur provient plus souvent de la calibration et de l'interprétation des données reçues. La tablette a une géometrie rectangulaire, donc après normalisation des coordonnées, nous obtenons des échelles différentes en abscisses et en ordonnées (typiquement des rectangles élémentaires), devant être pris en compte dans le calcul des dérivées et des distances. Dans notre cas, la cohérence des échelles en abscisses et en ordonnées est assurée par une normalisation.

Deuxièmement, la tablette envoie les données lorsque un changement d'état est capté sur la surface active, c'est à dire lorsque le stylet est en mouvement.

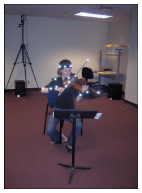
Afin d'obtenir des dérivations correctes, il est nécessaire de ré-échantillonner avec un pas constant, même s'il n'y a aucun changement sur la surface active.

Le module de l'équipe

Ce contrôleur est le fruit d'un projet nommé *The WiSe Box Project*, développé par Emmanuel Fléty. Ce boîtier convertit 16 contrôleurs continus sur 16 bits à travers le protocole OSC (OpenSoundControl). Le boîtier utilise la technologie WiFi ce qui permet d'avoir des temps de latence très faibles. Sur ce boîtier, nous plaçons un ensemble de capteurs (cf. photographie) qui contient trois accéléromètres et trois gyroscopes. Dans le cas du violon, il a été ajouté un capteur de pression sur l'archer.

C'est un outil puissant est bien adapté aux nouvelles technologies de transmission de données, comme le WiFi, grâce à au protocole OSC. L'échantillonnage est beaucoup plus important qu'avec un contrôleur MIDI ce qui permet l'élaboration d'installations plus ambitieuses dans le cadre de la performance musicale temps-réel. Par ailleurs, outre la performance, cet outil a un intérêt pédagogique avéré (cf (Bevilacqua, Frédéric et al. 2007)).

La *motion capture*



La technique de captation de mouvement présenté ici fait référence à un ensemble de sept caméras disposés en cercle autour d'un instrumentiste (ou danseur, acteur, etc ...). Chaque caméra donne une position (x, y) du même objet, et la prise de vue synchronisée de celui-ci permet la reconstruction des coordonnées (x, y, z) de l'objet dans l'espace tridimensionnel.

La captation de mouvement est un dispositif très puissant, et très utilisé dans le monde du cinéma. C'est aussi un dispositif très onéreux et difficilement transportable. Dans le cadre de notre travail, Norbert Schnell a pu effectuer un ensemble de captations de mouvements sur différents sons enregistrés. La saisie a été effectuée à Gratz sur un ensemble de 20 personnes celles-ci devant répéter trois fois le même geste sur un son donné.

2.2.2 Les paramètres choisis

La validation de l'algorithme nécessite un ensemble de paramètres simples et représentatifs alors que le soucis de capter au mieux l'intention d'une geste demande un contrôleur plus avancé. Ces deux points antagonistes nous a fait choisir la tablette graphique comme interface pour les expériences.

Les paramètres envoyés à l'ordinateur sont : la $position_x$, la $position_y$ et la *pression*. De cet ensemble de paramètres on va construire un ensemble plus large de douze paramètres : $position_x$, $position_y$, $vitesse_x$, $vitesse_y$, *accélération*, *norme de la position*, *norme de la vitesse*, *vitesse angulaire*, *accélération angulaire*, *pression*, *dérivée première de la pression*, *dérivée seconde de la pression*.

L'utilisation d'une tablette graphique induit des gestes sur deux dimensions d'espace, c'est pourquoi la vitesse suivant les vecteurs de base du repère cartésien ne sera par toujours pertinente, et qu'une vitesse angulaire reflète plus la dynamique du mouvement de la main. De même, la pression est un témoin important car on sait que plus le geste est rapide moins il tournera vite, et donc on aura besoin de la pression pour témoigner de l'envie d'effectuer des courbes reserrées. Enfin, la surface active a des dimensions finies et donc les bords ont un rôle dans le geste. C'est pourquoi les positions dans le plan sont pertinentes.

2.3 La synthèse sonore

Le mapping englobe les choix « quoi mapper, vers où et comment ». La partie précédente donnait une introduction du "quoi" : les données en entrée sont les paramètres gestuels à choisir minutieusement. Ce paragraphe introduit le "où" : les paramètres audio d'un moteur de synthèse sonore. Alors que le "comment" constitue le coeur de cette étude.

2.3.1 Vers la synthèse granulaire

Le contexte du projet nous incite à abolir les présupposés concernant un fichier audio. Etant donné que nous cherchons à caractériser le son par le geste, celui-ci est notre référence et notre degré de liberté, alors que la séquence audio doit être décrite de la manière la plus générale possible. Pour ce faire, on a choisi de prendre comme description temporelle du son un ensemble de **descripteurs** calculés par trames. Ces descripteurs sont assez nombreux pour offrir une description satisfaisante de la séquence audio. L'abolition des *a priori* sur le son nous conduit à considérer tous les descripteurs bien que cet ensemble contienne un grand nombre de redondances.

Ceci amène à deux remarques fondamentales :

1. les descripteurs peuvent être de natures très différentes : temporelle, spectrale, perceptive, etc
2. les descripteurs sont calculés sur des trames de taille N échantillons (ou T millisecondes) prises en fenêtrant le signal audio

La première remarque n'est pas sans rappeler les travaux de Denis Gabor (sur lesquels nous reviendrons rapidement dans le paragraphe suivant) mettant sur pied les premisses de la synthèse granulaires. La seconde remarque fait surgir la notion de grain par les méthodes de fenêtrage.

2.3.2 Synthèse granulaire : fondements

On attribue à Denis Gabor les fondements de la synthèse granulaire suite à ces travaux sur les quantas acoustiques et la théorie de l'audition. Il écrivit « *La dualité de nos sensations ne s'exprime pas seulement dans la description du son comme un signal en fonction du temps $s(t)$, ni dans sa représentation par composantes de Fourier $S(f)$. Par conséquence, une description mathématique qui tient compte de cette dualité est nécessaire.* »(1947). Pourtant la synthèse granulaire fut proposée comme outil pour la composition musicale par Iannis Xenakis (1971) et Curtis Roads (1978). L'application au temps-réel provient de l'apport de Barry Truax (1988).

L'idée fondamentale est la représentation d'un signal audio sous forme d'amas de grains, eux-mêmes caractérisés par plusieurs paramètres : durée du grain, forme d'onde, enveloppe, données spectrales, amplitude, etc ... Un grain est extrait du signal après multiplication par une fenêtre (cf. figure 2.3.2). L'ensemble des grains crée un corpus sonore.

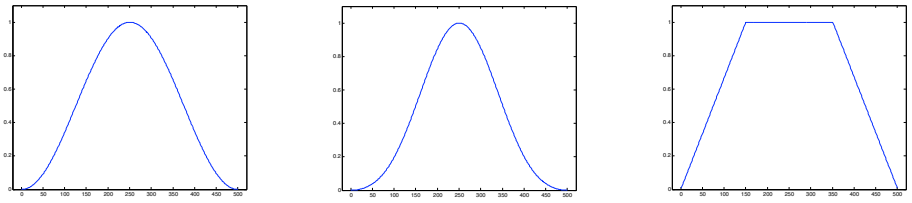


FIG. 2.1 – Différentes fenêtres utilisées pour la découpe des grains dans un signal audio. De gauche à droite : fenêtre de hanning, fenêtre de Blackman et fenêtre trapézoïdale.

Notre corpus de M grains créé, la synthèse sonore est le processus de concaténation de ces grains. Plusieurs paramètres rentrent en jeu, cependant l'idée globale est la suivante. Chaque grain se décrit par un vecteur de valeurs, qui sera le résultat d'une analyse temporelle/fréquentielle. La concaténation proprement dite dépend de deux paramètres : la taille du grain et la fréquence des grains. En effet, le premier paramètre correspond à la taille de la fenêtre utilisée. Le second correspond au recouvrement des grains lors de la lecture.

Typiquement, on aura des grains de 80ms pour une fréquence de 200Hz, soit un grain toutes les 5ms.

2.3.3 Outils disponibles au sein de l'équipe IMTR

Grâce à une coopération entre les équipes *Analyse/Synthèse des sons* et *Intéractions Musciales Temps Réel*, de nombreux outils de synthèse sonore temps réel sont disponibles.

Calcul des descripteurs

Le choix des descripteurs audio utilisés fait référence à l'article de Geoffroy Peeters dans le cadre du projet CUIDADO ((Peeters, Geoffroy 2004)). Cet article n'est pas une liste exhaustive des descripteurs audio (ou taxonomie des descripteurs) mais se veut un récapitulatif des moyens de décrire un signal audio suivant quatre points de vue.

1. La constance ou le dynamisme d'un descripteur : si celui-ci représente une valeur prise à l'instant t ou calculée sur tout la longueur du signal.
2. L'étendue temporelle de la description : si le descripteur correspond à un aspect local du signal (comme une attaque) ou s'il est relatif à la globalité.
3. L'abstraction du descripteur : ce que celui-ci représente.
4. Le processus d'extraction du descripteur : calcul sur le signal temporel ; calcul à partir de la transformée du signal ; calcul à partir d'un modèle du signal ; ou encore à partir d'un modèle d'audition humaine.

Cet article est la base d'un outil d'extraction de descripteurs audio qui se nomme `ircamdescriptor` (version 0.4). Développé au sein de l'équipe analyse/synthèse de l'IRCAM, le programme permet l'extraction de 210 descripteurs représentant le signal audio par sa description temporelle, en terme d'energie, spectrale, harmonique ou encore perceptive.

Le contrôle de la synthèse sonore relative à l'étude utilise des paramètres de haut-niveau, fruits d'une phase d'analyse, qui seront construits grâce à ces descripteurs audio classiques.

Synthèse granulaire et concaténative

Au sein de l'équipe Intéraction Musicale Temps Réel, des modules d'analyse et synthèse sonore temps réel ont été développés pour être utilisables dans `Max/MSP`. L'ensemble de ces modules constitue `GABOR` (cf. (Schnell, Norbert and al. 2005)), en référence à Denis Gabor, dont nous avons parlé précédemment. De manière synthétique, cet outil inclut les processus d'analyse/synthèse sonore suivants :

- ▷ Synthèse Granulaire
- ▷ Analyse/re-synthèse PSOLA
- ▷ Vocodeur de phase et autres techniques basées sur la transformée de Fourier à court terme

- ▷ Analyse/re-synthèse sinusiodale
- ▷ Convolution, corrélation
- ▷ Divers estimations

Le moteur de synthèse temps-réel de `Gabor` sera utilisé pour valider notre mapping. C'est une synthèse basée sur des corpus sonores (ou ensemble de grains). Notons, qu'un autre outil, appelé `cataRT` (et aussi basé sur `Gabor`) et développé par Diemo Schwarz, permet une synthèse concaténative de grains. L'utilisation basique réside en la création d'un corpus sonore constitué de grains de divers fichiers audio. Chaque grain se représente graphiquement par un point dans un plan à deux dimensions, mais est défini par un vecteur N dimensionnel où N est le nombre de descripteurs utilisés pour chaque grain. Dans l'environnement de `cataRT`, le mapping induirait une stratégie de navigation dans le repère à deux dimensions.

3

Problématiques et état de l'art

Dans le premier chapître, une étude a été faite sur les notions environnant le concept de mapping, à savoir, en amont, le geste et l'acquisition de celui-ci, puis, en aval, la synthèse sonore. Les considérations générales abordées ont introduit des éléments fondamentaux de la problématique : intention du geste, écoute et performance musicale, ou descripteurs "haut-niveau". Ces éléments vont permettre, dans ce chapître, de définir la problématique de l'étude qui sera recadrée parmi les travaux relatifs au mapping dans la littérature scientifique.

3.1 Problématique de l'étude

L'étude porte sur les gestes effectués lors de l'écoute d'une séquence sonore enregistrée dans un but de performance musicale. En d'autres termes, on se place à mi-chemin entre écoute et performance. La situation d'expérimentation est celle-ci. Un sujet effectue un geste en écoutant un son diffusé. La captation de son geste est synchronisée avec la lecture de la séquence sonore. Cette captation se fait à travers la tablette graphique. Ensuite, les données recueillies sont analysées afin de générer un mapping adapté.

La difficulté réside dans le processus d'analyse, i.e. la définition du mapping. Dans le contexte d'écoute active, un geste cherche à souligner des dynamiques sensibles dans un son écouté, et par là trouve des caractéristiques subjectives dans le son. L'étude de ces caractéristiques est ardue, et ne peut être menée de front, c'est à dire de manière directe : du geste vers le son.

La problématique générale de l'étude est la recherche d'une méthode inverse afin de mettre en lumière ces caractéristiques. Une motivation du projet est la conviction qu'on ne peut pas avoir une telle impression de cohérence entre deux phénomènes appartenant à des univers physiques différents sans qu'on puisse tenter de l'approcher par une méthode formelle.

La méthode inverse cherchée correspond à ce qu'on nomme la **gestification** du son. Par cette méthode, des informations plus abstraites que le suivi

temporel peuvent-être définies. L'objectif est d'identifier des récurrences intrinsèques, des formes temporelles implicites de certains descripteurs audio, etc... et ainsi définir des descripteurs haut-niveau pour le contrôle de la synthèse sonore à partir des données initiales. Ne connaissant pas a priori ces nouveaux descripteurs, on aimerait utiliser un mécanisme génératif afin de les calculer.

Enfin, le travail de resynthèse permet une vérification auditive et sensible du mapping obtenu. Il nous renseigne sur le pouvoir expressif de la correspondance, et valide la possibilité d'identifier les récurrences et autres informations de haut-niveau. La synthèse d'un son marqué par le geste s'identifie à une phase d'analyse inverse.

En résumé, la problématique se divise en deux étapes

1. Etape d'analyse ou définition du mapping (processus de gestification du son)
2. Etape de synthèse sonore ou analyse inverse

Revenons dès à présent sur les différents travaux effectués autour du concept de mapping.

3.2 Travaux précédents relatifs au Mapping

Conventionnellement, le terme "mapping" fait référence à un opérateur mathématique mettant en relation les éléments d'un ensemble aux éléments d'un autre ensemble. Lorsqu'on utilise les interfaces homme-machine pour la création musicale, ce terme fait plus spécifiquement allusion à la mise en relation de paramètres gestuels, provenant d'un contrôleur, avec des paramètres de synthèse sonore. Dans cette partie nous allons revenir sur les motivations du mapping, puis nous reviendrons sur les différentes orientations empreintées dans la littérature.

3.2.1 Motivations

La musique a longtemps été matérialisée par les instruments acoustiques. L'utilisation de l'adjectif "acoustique" signifie que l'instrument est assujetti aux lois de la mécanique vibratoire. Cette longue période de pratique instrumentale a fait naître une grande technicité chez les interprètes, de même qu'une grande diversité de techniques instrumentales suivant les cultures. Un dénominateur commun à ce panel de pratiques instrumentales existe et peut s'exprimer ainsi : l'utilisateur amène une énergie qui sera source de création du son. En amont, la transmission de cette énergie est indissociable de l'articulation gestuelle, des nuances et de certains contrôles expressifs propres à l'instrumentiste.

Il existe une multitude d'instruments et une étendue timbrale très variée, d'autant plus qu'un geste instrumental fin d'un musicien sur son instrument peut avoir l'effet d'une variation fine de timbre et donc une expressivité plus importante. En outre, l'énergie transférée à l'instrument peut se faire de manière directe ou indirecte. En effet, un flûtiste aura un accès direct à la colonne d'air vibrante dans son instrument, entraînant un très bon contrôle du son, contrairement au pianiste qui n'a accès qu'à l'interface (les touches) désolidarisée du mécanisme vibratoire au moment du contact marteau/cordes.

Pourtant, l'énergie apportée par l'instrumentiste est toujours la cause de l'excitation du corps, et donc du rayonnement acoustique. Comme rappelé dans (Hunt, Andy and Wandereley, Marcelo M. 2002), le piano se trouve être un instrument de timbre très peu variable en comparaison avec un saxophone. Cependant le piano excelle dans la polyphonie et dans l'accessibilité, ce qui en fait un instrument universel. Les instruments acoustiques, ormis ce point commun inéluctable, ont leur spécificité qui leur confère des propriétés propres : timbre, polyphonie, amplitude, tessiture, ... Ces instruments sont le fruit de siècles d'évolutions et d'adaptations. L'instrument constitue une projection de l'idée créatrice musicale.

De même que pour la lutherie classique, lors de la conception d'un instrument virtuel on se pose la question des possibilités de contrôle du phénomène sonore de même que du pouvoir expressif de l'instrument. Il en résulte que le contrôle du son peut atteindre une très grande précision, et une diversification non égalée dans les instruments acoustiques, alors que le pouvoir expressif s'en trouve significativement réduit. Là où des contraintes mécaniques définissaient le comportement des instruments acoustiques, dans le cas des instruments numériques, un même ensemble de geste pour un même moteur de synthèse peut avoir des comportements tout à fait différents. L'intérêt du travail sur la correspondance geste/son est bien d'équilibrer ces notions et de définir des mapping expressifs pour la performance musicale. On propose une révision essentielle des travaux effectués dans ce champ de recherche.

3.2.2 Retour sur les différentes stratégies de mapping

Beaucoup de précédents travaux portent sur le rôle du mapping dans la musique assistée par ordinateur. On essaiera de donner une vue globale de ces travaux, afin de mieux situer cette étude et donc de lui donner un sens.

On remarque dans la littérature, que le mapping, considéré comme faisant partie de l'instrument, a été étudié suivant deux principales directions.

1. le mapping *explicite*, où la relation entre les paramètres gestuels et les paramètres sonores est connue

2. le mapping *implicite*, où cette relation est un mécanisme génératif

Ces deux stratégies globales ont leurs avantages et leurs inconvénients. De manière succinte, la forme explicite permet un contrôle total du mapping et donc de l'instrument, amenant à une meilleure connaissance de l'efficacité du mapping. En revanche la forme implicite permet la création de nouveaux paramètres, adaptatifs, et souvent sources d'une plus grande expressivité dans le jeu instrumental.

Mapping explicites

La première forme de mapping fut l'explicite, car plus facilement implémentable. Très vite une classification fut créée afin de cibler les différentes stratégies. Une classification conventionnelle peut être trouvée dans (Rovan, Wanderley, Dubnov, and Depalle Rovan et al.) et être exprimée suivant trois classes disjointes :

1. *un-vers-un* : chaque paramètre indépendant du geste est connecté à un paramètre musical, par exemple la vitesse assignée à la fréquence d'un oscillateur.
2. *divergent* : un paramètre gestuel est utilisé pour contrôler simultanément plusieurs paramètres musicaux. Celui-ci permet une approche plus global du son et ne permet pas un micro-contrôle de l'objet sonore, étant donné que plusieurs de ces paramètres évoluent en même temps.
3. *convergent* : plusieurs paramètres gestuels sont couplés pour contrôler un paramètre musical.

L'expressivité de l'instrument est conditionnée par le choix du mapping, et dans le cas de la forme explicite, elle sera conditionnée par le choix d'une de ces classes, ou d'une combinaison de celles-ci. (Van Nort, Doug et al. 2004) analyse explicitement l'expressivité de ces différents mapping, et il est montré que le simple mapping *un-vers-un* limite le potentiel expressif, pour un modèle d'instrument de type clarinette, comparé à un modèle plus complexe combinant *convergent* et *divergent*. Un étude complémentaire dans (Hunt et al. 2002) présente des expériences où les stratégies de mapping changent et remarque leur pouvoir émotionnel sur l'instrumentiste. D'autre part (Van Nort, Doug and Wanderley, Marcelo 2006) met l'accent sur le fait qu'un mapping est à la fois *qu'est ce qu*'on relie et *comment* le fait-on. Ainsi le choix des paramètres est tout aussi important que leur mis en relation. La quête de l'expressivité des instruments virtuels a amené à abstraire ou à généraliser certaines stratégies comme dans (Hunt, Andy and Wandereley, Marcelo M. 2002), (Hunt, Andy and Kirk, Ross 2000) ou encore (Wanderley, Marcelo et al. 1998) où les auteurs présentent un mapping multi-couches c'est à dire qu'ils opèrent une séparation entre deux espaces : un relatif au contrôleur gestuel et l'autre au moteur de synthèse. Ils insèrent alors

un espace intermédiaire, ou couche abstraite, permettant une plus grande flexibilité dans la manipulation des paramètres.

Mapping implicites

Cette forme de mapping utilise des mécanismes génératifs. Le modèle principalement utilisé dans un mapping implicite pour la performance musicale est le réseau de neurones. (Lee, Matthew and Wessel, David 1992) proposa un modèle de réseau de neurones multi-couches afin de contrôler les paramètres de synthèse. Puis ce modèle fut utilisé pour la synthèse de la parole (Fels and Hinton 1995) ou encore pour la performance musicale, dans (Modler, Paul et al. 2003). Cet article présente l'implémenation d'un réseau de neurones (TDNN) pour la reconnaissance du geste dans une séquence filmée d'une main gantée. Un ensemble de gestes de la main et appris par le réseau de neurones permettant de créer un processus de reconnaissance ou encore un contrôle musical expressif.

Une autre approche est l'utilisation de Chaînes de Markov Cachées (HMM) dans l'étude du geste musical. Cette méthode a été implémentée pour la reconnaissance et le suivi de geste (Bévilacqua 2005).

Ainsi, ces mécanismes génératifs se basent sur un apprentissage, c'est à dire la définition d'un ensemble d'entraînement, puis une validation par un ensemble de tests. Les réseaux de neurones et les chaînes de markov restent liés à des phénomènes temporels et l'apprentissage permet de définir la probabilité de se trouver dans un certain état à l'instant t.

L'originalité de notre étude réside en l'atemporalité des correspondances entre geste et son. Le mécanisme mathématique se base sur le principe de **corrélation** entre variables, et définit génératívement des variables de plus haut niveau afin d'améliorer cette corrélation. La méthode de mise en correspondance de variables corrélées, appelée analyse canonique (ou *canonical correlation analysis*), donnera une définition formelle de la problématique de l'étude : la gestification du son.

4

Principes de la "Gestification"

Dans ce chapitre, on présentera la méthode mathématique formelle permettant la mise en relation de deux domaines distincts, à savoir le gestuel, et le sonore. La consistance de l'étude passe par un changement de représentation des données par le biais d'une méthode statistique connue dans le domaine sociologique mais encore très peu utilisée dans le domaine du son : la *corrélation canonique* (ou CCA *canonical correlation analysis*). Comme pour toute méthode statistique, il est aisé d'obtenir des résultats et c'est pourquoi ils ne peuvent pas être dissociés de leur interprétation. La partie suivante formalise l'interprétation de l'analyse canonique sous certaines hypothèses. Enfin, il sera proposé une optimisation de l'analyse canonique par une méthode intuitive d'alignement temporel.

4.1 Formalisation : l'analyse canonique

On adoptera les conventions de notation suivantes. Une majuscule marquée, e.g. $\mathbf{M}$, désignera une matrice, et une minuscule marquée, e.g. $\mathbf{m}_j$, désignera un vecteur. Un scalaire, tel que la valeur d'une variable à l'instant t, sera noté x (ou $m_{i,j}$ pour l'élément (i, j) d'une matrice $\mathbf{M}$).

4.1.1 Introduction

Le principe de corrélation énoncé en fin de chapître 3 porte sur les variables gestuelles et sonores. Une corrélation se calcule entre deux variables aléatoires lorsqu'on a à disposition un ensemble d'observations. Dans le cas multidimensionnel, on crée deux tableaux de données (appelés aussi *groupes*) sur lesquels on effectuera les calculs. Dans notre cas, ces tableaux se construisent avec des echantillons des paramètres gestuels, d'une part, et par le calcul de descripteurs audio instantannés sur le signal, d'autre part. Si on qualifie ces ensembles de variables d'*aléatoires*, c'est pour spécifier qu'on se place dans le cadre d'expériences dont on ne connaît pas le résultat a priori mais seulement l'univers des possibles. On sera en droit d'appliquer les résultats des mathématiques probabilistes, notamment les méthodes de regression mutliple.

Dans ces méthodes, il existe un grand nombre d'outils d'analyse de la correspondance entre deux tableaux de données, par exemple l'analyse canonique, l'analyse factorielle inter-batteries ou l'analyse des redondances.

L'analyse canonique consiste à comprendre les combinaisons linéaires qui existent entre un groupe de variables à expliquer et un autre groupe de variables explicatives ; il s'agit donc de déterminer les corrélations linéaires entre ces deux groupes. Pour ce faire on projète les deux ensembles de données dans deux espaces puis on étudie la position de l'un par rapport à l'autre. Les variables explicatives sont les paramètres du geste tandis que les variables à expliquer sont les descripteurs audio. L'objectif se résume à obtenir deux espaces confondus tout en sachant que les hypothèses sous-jacentes concernant la correspondance paramètres gestuels/descripteurs audio sont :

1. leur relation est linéairement corrélée
2. leur relation est instantannée

L'analyse canonique a été préférée aux autres méthodes précédemment citées car la maximisation de la corrélation assure une stratégie de mapping fiable entre les paramètres gestuels (projetés dans l'espace) et les descripteurs audio (projetés). De plus, elle offre le grand avantage d'être invariante par transformation affine sur les données.

4.1.2 Formalisation

Considérons deux tableaux de données $\mathbf{X}$ et $\mathbf{Y}$ représentant respectivement p variables $\mathbf{x}_j$ et q variables $\mathbf{y}_k$ observées sur les mêmes n individus. Dans la suite, nous supposerons que le rang de $\mathbf{X}$ est égal à p et celui de $\mathbf{Y}$ à q. De même, sans perte de généralitié, on suppose les variables x_j et y_k centrées.

On cherche alors une autre base de représentation pour $\mathbf{X}$ en choisissant une direction $\mathbf{A}$ et en projetant $\mathbf{X}$ dans cette direction :

$$\mathbf{X} \longrightarrow \langle \mathbf{A}, \mathbf{X} \rangle$$

En faisant de même pour le tableau de données $\mathbf{Y}$ en choisissant une direction $\mathbf{B}$, on obtient deux nouvelles représentations de nos données initiales

$$\mathbf{T} = \left(\langle \mathbf{A}, \mathbf{x}_1 \rangle, \langle \mathbf{A}, \mathbf{x}_2 \rangle, ..., \langle \mathbf{A}, \mathbf{x}_p \rangle \right)$$
$$\mathbf{U} = \left(\langle \mathbf{B}, \mathbf{y}_1 \rangle, \langle \mathbf{B}, \mathbf{y}_2 \rangle, ..., \left\langle \mathbf{B}, \mathbf{y}_q \right\rangle \right)$$

La première étape de la corrélation canonique consiste à choisir les directions $\mathbf{A}$ et $\mathbf{B}$ de manière à maximiser la corrélation entre les deux nouveaux tableaux de données $\mathbf{T}$ et $\mathbf{U}$.

Ainsi, il s'agit de maximiser la fonction suivante :

$$\begin{aligned}\mathbf{R} &= \max_{\mathbf{A},\mathbf{B}} corr\left(\mathbf{T},\mathbf{U}\right) \\ &= \max_{\mathbf{A},\mathbf{B}} \frac{cov(\mathbf{T},\mathbf{U})}{\sqrt{var(\mathbf{T})}\sqrt{var(\mathbf{U})}}\end{aligned}$$

On se rappelle que les variables $\mathbf{x}_i$ et $\mathbf{y}_j$ sont centrées, alors $\mathbf{t}_i$ et $\mathbf{u}_j$ le sont aussi, ce qui nous permet d'écrire,

$$\mathbf{R} = \max_{\mathbf{A},\mathbf{B}} \frac{\mathbb{E}\left[\mathbf{T}.\mathbf{U}^T\right]}{\sqrt{\mathbb{E}\left[\mathbf{T}.\mathbf{T}^T\right]\mathbb{E}\left[\mathbf{U}.\mathbf{U}^T\right]}} \tag{4.1}$$

Où $\mathbb{E}[X]$ dénote l'espérance de la variable aléatoire X. Afin de calculer la corrélation, cette espérance doit être estimée. L'utilisation d'un estimateur biaisé ou non biaisé n'a ici que peu d'importance, car la corrélation est normalisée. On note $\hat{\mathbb{E}}\left[f(x,y)\right]$, l'espérance empirique d'une fonction f qui se définit par :

$$\hat{\mathbb{E}}\left[f(x,y)\right] = \frac{1}{m}\sum_{i=1}^{m} f(x_i,y_i)$$

On peut réécrire l'expression de la corrélation estimée comme

$$\begin{aligned}\hat{\mathbf{R}} &= \max_{\mathbf{A},\mathbf{B}} \frac{\hat{\mathbb{E}}\left[\langle\mathbf{A},\mathbf{X}\rangle\langle\mathbf{B},\mathbf{Y}\rangle\right]}{\sqrt{\hat{\mathbb{E}}\left[\langle\mathbf{A},\mathbf{X}\rangle^2\right]\hat{\mathbb{E}}\left[\langle\mathbf{B},\mathbf{Y}\rangle^2\right]}} \\ &= \max_{\mathbf{A},\mathbf{B}} \frac{\hat{\mathbb{E}}\left[\mathbf{A}^T\mathbf{X}\mathbf{Y}^T\mathbf{B}\right]}{\sqrt{\hat{\mathbb{E}}\left[\mathbf{A}^T\mathbf{X}\mathbf{X}^T\mathbf{A}\right]\hat{\mathbb{E}}\left[\mathbf{B}^T\mathbf{Y}\mathbf{Y}^T\mathbf{B}\right]}}\end{aligned} \tag{4.2}$$

Ce qui conduit à la formulation,

$$\hat{\mathbf{R}} = \max_{\mathbf{A},\mathbf{B}} \frac{\mathbf{A}^T\hat{\mathbb{E}}\left[\mathbf{X}\mathbf{Y}^T\right]\mathbf{B}}{\sqrt{\mathbf{A}^T\hat{\mathbb{E}}\left[\mathbf{X}\mathbf{X}^T\right]\mathbf{A}\mathbf{B}^T\hat{\mathbb{E}}\left[\mathbf{Y}\mathbf{Y}^T\right]\mathbf{B}}} \tag{4.3}$$

On adoptera conventionnellement les notations suivantes pour la covariance estimée entre les ensembles de données $\mathbf{X},\mathbf{Y}$,

$$\hat{cov}(\mathbf{X},\mathbf{Y}) = \hat{\mathbb{E}}\left[\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix}\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix}^T\right] = \begin{bmatrix}\mathbf{C}_{xx} & \mathbf{C}_{xy}\\ \mathbf{C}_{yx} & \mathbf{C}_{yy}\end{bmatrix}$$

On appelle les matrices $\mathbf{C}_{xx}$ et $\mathbf{C}_{yy}$ les matrices de covariance intra-groupe et $\mathbf{C}_{xy},\mathbf{C}_{yx}$ (où $\mathbf{C}_{xy} = \mathbf{C}_{yx}^T$) les matrices de covariance inter-groupe. On remarque que l'analyse en composantes principales permet de décrire la

matrice de corrélation $\mathbf{C}_{xx}$ ou $\mathbf{C}_{yy}$ alors que l'analyse canonique décrit la matrice inter-groupe $\mathbf{C}_{xy}$.

On formulera la matrice $\hat{\mathbf{R}}$ de préférence en utilisant les matrices de covariance, soit

$$\hat{\mathbf{R}} = \max_{\mathbf{A},\mathbf{B}} \frac{\mathbf{A}^T \mathbf{C}_{xy} \mathbf{B}}{\sqrt{\mathbf{A}^T \mathbf{C}_{xx} \mathbf{A}\, \mathbf{B}^T \mathbf{C}_{yy} \mathbf{B}}} \tag{4.4}$$

Ainsi, en analyse canonique on recherche successivement des composantes, dites **canoniques**, orthogonales $\mathbf{t}_h = \mathbf{X}\mathbf{a}_h$ et $\mathbf{u}_h = \mathbf{Y}\mathbf{b}_h$ de corrélation maximale, où h évolue dans un ensemble fini discret de cardinal le minimum des rangs des matrices $\mathbf{X}$ et $\mathbf{Y}$. La résolution de ce problème constitue l'analyse canonique.

4.1.3 Résolution

On se réfère à l'équation 4.4. Dans ce paragraphe on ne manipulera que des vecteurs et des scalaires pour plus de simplicité. L'équation 4.4 se réécrit comme

$$\hat{\mathbf{r}}_h = \max_{\mathbf{a}_h,\mathbf{b}_h} \frac{\mathbf{a}_h^T \mathbf{C}_{xy} \mathbf{b}_h}{\sqrt{\mathbf{a}_h^T \mathbf{C}_{xx} \mathbf{a}_h\, \mathbf{b}_h^T \mathbf{C}_{yy} \mathbf{b}_h}} \tag{4.5}$$

En premier lieu, on peut remarquer qu'une multiplication des vecteurs $\mathbf{a}_h$ et $\mathbf{b}_h$ par un scalaire n'affecte pas la corrélation, ainsi l'optimisation de $\hat{\mathbf{r}}_h$ se fait sous les contraintes $\mathbf{a}_h^T \mathbf{C}_{xx} \mathbf{a}_h = 1$ et $\mathbf{b}_h^T \mathbf{C}_{yy} \mathbf{b}_h = 1$. On remarque qu'on pourrait aussi bien imposer aux vecteurs $\mathbf{a}_h$ et $\mathbf{b}_h$ d'être unitaires et ceci élaire une des difficultés pratiques de l'analyse canonique. Lorsqu'on cherche des vecteurs $\mathbf{a}_h, \mathbf{b}_h$ afin de maximiser $\hat{\mathbf{r}}_h$, on cherche à la fois à maximiser la covariance $cov(\mathbf{X}\mathbf{a}_h, \mathbf{Y}\mathbf{b}_h)$ et à minimiser les variances $var(\mathbf{X}\mathbf{a}_h)$ et $var(\mathbf{Y}\mathbf{b}_h)$. Cela amène à trouver des composantes canoniques bien corrélées mais peu explicatives de leur groupe respectif.

Le problème qui en résulte est la maximisation d'une forme quadratique sous contraintes linéaires. De manière équivalente, le problème peut être posé en utilisant les multiplicateurs de Lagrange et l'opérateur L :

$$L(\lambda, \mathbf{a}_h, \mathbf{b}_h) = \mathbf{a}_h^T \mathbf{C}_{xy} \mathbf{b}_h - \frac{\lambda_x}{2}\left(\mathbf{a}_h^T \mathbf{C}_{xx} \mathbf{a}_h - 1\right) - \frac{\lambda_y}{2}\left(\mathbf{b}_h^T \mathbf{C}_{yy} \mathbf{b}_h - 1\right)$$

La maximisation de cet opérateur est un problème bien connu de l'optimisation. Le maximum de la fonction est atteint et correspond à un point selle. Pour ce faire, il faut annuler les dérivées de l'opérateur L suivant les directions $\mathbf{a}_h$ et $\mathbf{b}_h$. Il en résulte les équations suivantes,

$$\frac{\partial L}{\partial \mathbf{a}_h} = \mathbf{C}_{xy} \mathbf{b}_h - \lambda_x \mathbf{C}_{xx} \mathbf{a}_h = 0 \tag{4.6}$$

$$\frac{\partial L}{\partial \mathbf{b}_h} = \mathbf{C}_{yx} \mathbf{a}_h - \lambda_y \mathbf{C}_{yy} \mathbf{b}_h = 0 \tag{4.7}$$

Ces équations, ajoutées aux contraintes linéaires, induisent que les scalaires λ_x et λ_y sont égaux. On notera $\lambda_x = \lambda_y = \lambda$. On en déduit aisément,

$$\mathbf{b}_h = \frac{\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{a}_h}{\lambda} \tag{4.8}$$

Qui en substituant dans l'équation 4.6 donne

$$\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{a}_h = \lambda^2\mathbf{a}_h \tag{4.9}$$

On pourra en déduire $\mathbf{b}_h$ facilement grâce aux équations 4.8 et 4.9. Pourtant les rôles de $\mathbf{a}_h$ et $\mathbf{b}_h$ dans les calculs précédents sont équivalents, et un déroulement similaire conduirait à l'équation,

$$\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{b}_h = \lambda^2\mathbf{b}_h \tag{4.10}$$

Remarque 4.1.3.1 *Les matrices symétriques et positives $\boldsymbol{C}_{xx}$ et $\boldsymbol{C}_{yy}$ sont inversées dans le calcul des vecteurs propres. L'algorithme prévoit une réduction des matrices $\boldsymbol{X}$ et $\boldsymbol{Y}$ de manière à ce qu'elles soient de rang maximal conférant ainsi aux matrices $\boldsymbol{C}_{xx}$ et $\boldsymbol{C}_{yy}$ la propriété "définies positives" et donc inversibles.*

L'analyse canonique revient à un problème de calcul de valeurs propres. C'est un problème bien connu et une solution peut-être calculée avec les logiciels classiques de calcul numérique tels que `Matlab`.

L'orthogonalité des composantes canoniques u_j et t_j est immédiate. En effet, il suffit de nous rappeler que les vecteurs propres d'une matrice symétrique sont orthogonaux.

Finalement l'analyse canonique nous procure :

1. Une matrice $\mathbf{R}$ des coefficients de corrélation (les valeurs propres)
2. Deux matrices $\mathbf{A}$ et $\mathbf{B}$ de projection (les vecteurs propres)
3. Les composantes canoniques (variables corrélées)

4.1.4 Interprétation géometrique

Les propriétés de la régression multiple induisent une propriété géometrique simple entre les composantes canoniques et les variables initiales. Cette propriété s'illustre graphiquement par la figure [4.1], et s'exprime par

1. La composante $\mathbf{X}a_j$ est colinéaire à la projection de $\mathbf{Y}b_j$ sur l'espace engendré par les colonnes de X
2. La composante $\mathbf{Y}b_j$ est colinéaire à la projection de $\mathbf{X}a_j$ sur l'espace engendré par les colonnes de Y

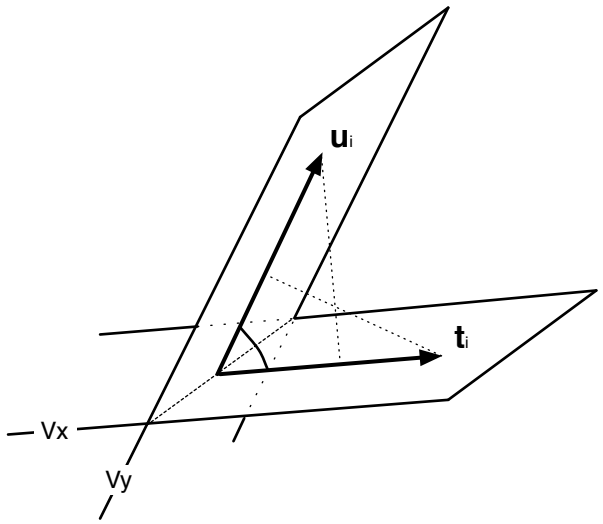


FIG. 4.1 – Interprétation géometrique de l'analyse canonique. On veut réduire l'angle entre les vecteurs de V_x et V_y c'est à dire faire coïncider les espaces. Le cosinus de l'angle est précisément le coefficient de corrélation canonique des deux composantes elles-même canoniques.

Pour une démonstration de ces propriétés, on réfère le lecteur à l'annexe A. La corrélation entre les couples de variables projetées (ou composantes canoniques) correspond à l'angle formé entre les deux espaces. Si les espaces sont confondus, la corrélation est maximale, et si les espaces sont perpendiculaires, les variables d'un ensemble sont totalement indépendantes des variables de l'autre ensemble.

De simples relations peuvent en être extraites, et sont rapportées ci-dessous.

$$\mathbf{X}^T\mathbf{u}_h = \lambda\mathbf{X}^T\mathbf{t}_h \tag{4.11}$$

$$\mathbf{Y}^T\mathbf{t}_h = \lambda\mathbf{Y}^T\mathbf{u}_h \tag{4.12}$$

Autrement dit, en divisant ces relations par le nombre d'observations et par la racine carré du produit des variances, on obtient

$$corr(\mathbf{x}_j, \mathbf{u}_h) = \lambda_h corr(\mathbf{x}_j, \mathbf{t}_h), j = 1, ..., p \tag{4.13}$$

$$corr(\mathbf{y}_j, \mathbf{t}_h) = \lambda_h corr(\mathbf{y}_j, \mathbf{u}_h), j = 1, ..., q \tag{4.14}$$

Ainsi, le vecteur des corrélations entre une variable $\mathbf{x}_j$ initiale et une composante canonique $\mathbf{Xa}_h$ est colinéaire au vecteur des corrélations entre cette même variable $\mathbf{x}_j$ et la composante canonique $\mathbf{Yb}_h$ (Il en va de même pour les variables $\mathbf{y}_j$). En d'autres termes, on retrouve par une interprétation vectorielle que les composantes canoniques peuvent servir d'interprétation de l'ensemble des variables initiales.

4.2 Interprétation des résultats

Il n'est pas immédiat que les composantes canoniques obtenues soient satisfaisantes, significatives ou encore non redondantes. Nécessairement, une interprétation des résultats doit être faite. Dans le cadre de l'analyse canonique, l'interprétation se divise en une partie quantitative où on calculera certains indices interprétables, et une partie qualitative qui met en exergue des objets mis en jeu par ces indices.

4.2.1 Examen des corrélations simples

Avant d'effectuer une analyse canonique, on se doit de mieux connaître les données initiales auxquelles nous avons accès. Dans ce paragraphe on utilisera l'exemple donné en annexe A.

L'exemple utilisé comporte trois variables X_i et six variables Y_j, représentant, respectivement, les observations du geste et du son. Cette première étape nous permet de repérer les relations linéaires entre les variables dans chacun des deux groupes par un calcul de la corrélation simple *intra-groupe* (cf. A.1 et A.2, figures [A.1] et [A.2]). Ainsi on observe qu'il existe une relation forte entre Y_6 et Y_3 car leur coefficient de corrélation est 0.9938.

Ensuite, on calcule la corrélation simple *inter-groupe*, ce qui nous informe des relations linéaires entre les variables d'un groupe et les variables d'un autre groupe (cf. A.3). Par exemple, la variable X_3 est corrélée à presque 60% avec la variable Y_2. Cette valeur du coefficient de corrélation est le maximum de la matrice. En outre, la valeur du premier coefficient de corrélation canonique est au moins égale au maximum des coefficients de corrélation simple donc on est assuré d'obtenir deux composantes canoniques corrélées au moins à 60%. D'ailleurs, la figure [A.3] montre une corrélation canonique de 0.8714 pour la première composante.

4.2.2 Tests d'hypothèse, de signification

Dans le domaine statistique, il est nécessaire de savoir dans quelle mesure les données sont interprétables. Par exemple, lorsqu'on estime une moyenne ou une variance, le seul résultat numérique est insuffisant. Conventionnellement, ce résultat est accompagné d'un intervalle de confiance dans lequel on sait que notre estimateur donne une valeur juste à un certain seuil près (souvent 95%). De la même manière, il est utile de savoir si un résultat donné sous forme de tableau de valeurs a une espérance en accord avec celle d'un modèle, si sa variance correspond bien à ce qu'on attendait ou encore si les variables mises en jeu sont indépendantes et suivent bien une loi de probabilité donnée.

Introduction

Les tests d'hypothèse constituent un outil de l'estimation de la pertinence de résultats statistiques. Ils ne sont pas un moyen d'obtention d'informations sur un ensemble de données, mais un moyen de vérification d'intuitions sur ces données.

Malgré une formalisation peu intuitive, le test d'hypothèse est un processus qui peut se comprendre aisément de manière informelle. Par exemple, on dispose un ensemble de cartes sur une table, faces cachées, tout en ayant pris soin de laisser seulement les cartes de couleur noire parmi cet ensemble. Puis on demande à une personne de retourner les cartes une à une. La première est noire, puis la seconde, puis la troisième, etc ... La question est : à partir de quel nombre de cartes retournées va t-elle se rendre compte que toutes les cartes sont de couleur noire ? L'expérience refaite sur un grand nombre de personnes nous permet de définir un seuil (un nombre de cartes) et une probabilité P. Cela signifie que la probabilité qu'une personne devine que toutes les cartes de la table sont noires avant d'atteindre le seuil de nombre de cartes retournées, est égale à P. L'hypothèse vérifiée est : "toutes les cartes sur la table sont noires".

Dans notre étude, nous nous intéressons à réfuter l'hypothèse "le groupe de variables gestuelles et le groupe de descripteurs sonores ne sont pas corrélés" car l'hypothèse intiale est que l'ensemble des paramètres gestuels et l'ensemble des descripteurs audio sont corrélés linéairement. De fait, on va étudier à partir de quel nombre de coefficients de corrélation canonique, on aura expliqué la covariance inter-groupe.

Le méthodologie des tests d'hypothèse fait apparaître différentes étapes essentielles pour la validité du processus.

- ▷ Test d'Hypothèse : on émet une hypothèse à vérifier
- ▷ Statistique utilisée T : on définit une statistique sur les données numériques à disposition
- ▷ Loi de probabilité de la statistique T : on définit une loi de probabilité suivie par la statistique précédemment définie
- ▷ Règle de décision : seuil d'acceptation, ou de refutation, de l'hypothèse
- ▷ Niveau de signification de la donnée numérique d'une observation de la statistique

Les tests d'hypothèse pour l'analyse canonique

La démarche suivra un schéma qui testera la signification des coefficients de corrélation canonique itérativement. Au premier coefficient non-significatif trouvé, on est assuré que les prochains coefficients seront tous non-significatifs.

Précédemment, on avait noté $\mathbf{C}_{xy}$ la matrice de covariance entre le groupe de variables $\mathbf{X}$ et le groupe de variables $\mathbf{Y}$. La formalisation de l'hypothèse

citée dans le paragraphe précédent, pour tester le premier coefficient, sera

$$H_0 : \mathbf{C}_{xy} = 0$$

Si on appelle ρ_i, où $i = 1..k$, les coefficients de corrélation simple entre $\mathbf{X}$ et $\mathbf{Y}$, l'hypothèse précédente peut être mise sous la forme

$$H_0 : \rho_0 = 0, \rho_1 = 0, ..., \rho_k = 0$$

La prochaine étape qui consiste à définir une statistique sur les observations est difficile. Dans le cas du jeu de cartes, une variable aléatoire modélise le résultat d'un tirage et suit une loi de Bernoulli ($\frac{p}{2} + \frac{1-p}{2}$). Si on voulait écrire une variable aléatoire donnant le nombre de cartes noires retournées ou le nombre de cartes rouges, sa statistique serait la somme des tirages, et cette somme respecterait une loi binomiale.

Lorsqu'on ne connaît pas, *a priori*, la statistique des observations on utilise des méthodes pour trouver la meilleure statistique, relative à une hypothèse. Ces méthodes sont **le test du rapport de vraissemblance** et **le test d'union-intersection**. Dans notre étude, on utilise la permière méthode et on réfère le lecteur à l'annexe A, paragraphe A.1.1.

La statistique pour tester l'hypothèse H_0 est donnée par

$$|\mathbf{Id} - \mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}| = \prod_{i=1}^{k}\left(1 - r_i^2\right) \tag{4.15}$$

La loi de probabilité assignée à la statistique 4.15 s'obtient en utilisant l'approximation de Bartlett (cf. (Tenenhaus, Michel 1998) ou (Mardia, K.V. et al. 1979)) qui s'écrit

$$-\left[n - \frac{1}{2}(p + q + 3)\right]\log\prod_{i=1}^{k}\left(1 - r_i^2\right) \sim \chi_{pq}^2 \tag{4.16}$$

Celle-ci est valable lorsque n est "grand". Comme on travaille avec des observations dans le temps, et que le nombre de nos observations n'est jamais inférieur à quelques centaines, on considèrera l'approximation 4.16 comme valable. En outre, p et q représentent respectivement le nombre de variables dans chacun des deux groupes : p paramètres gestuels et q descripteurs audio.

Ainsi, on obtient une statistique décrite par une loi bien connue : la loi du χ^2. Les valeurs numériques de la loi χ^2 s'obtiennent en fixant le nombre de degrés de liberté pq et le risque α (valeur numérique comprise entre 0 et 1).

Ensuite, afin de valider la statistique et connaître le niveau de signification des coefficients de corrélation canonique, il faut définir une règle de décision.

Au vu de 4.16, plus les r_k^2 seront proches de 1 et plus le membre de gauche sera positivement grand. D'autre part, la loi du χ^2 pour pq donné, augmentera si on fait diminuer le risque. Communément, le risque est $\alpha = 0.5$ et correspond au pourcentage d'erreur admissible : 5%. De fait, plus les coefficients de corrélation canonique sont proches de 1 et plus on peut faire tendre α vers 0 donc maximiser la signification de la corrélation. En d'autres termes on minimise l'effet des singularités du geste prises dans le temps.

Ainsi la règle de décision sera « On accepte H_0 si la statistique T définie par 4.15 est supérieure ou égale à c ». Le risque, introduit précédemment, réside dans la probabilité d'effectuer l'erreur (dite de première espèce) : accepter H_0 alors que H_0 est fausse. Si on appelle α le risque,

$$\alpha = P(T \geq c \,|\, H_0 \quad \text{fausse}) \tag{4.17}$$

Le seuil critique c sera calculé à partir de 4.17 en fixant $\alpha = 0.05$ et en ayant une table de la loi concernée.

L'étape finale est le calcul à proprement parlé.

Et ceci nous permet de conclure sur la signification du premier coefficient de corrélation canonique. Itérant sur les coefficients, on aimerait formuler une hypothèse générale. Pour le coefficient de corrélation canonique r_{s+1} où $s > 0$, celle-ci peut s'écrire ainsi

$$H_s : \rho_1 \neq 0, ..., \rho_s \neq 0, \rho_{s+1} = 0, ..., \rho_k = 0$$

On peut dès lors utiliser la statistique suivante,

$$\prod_{i=s+1}^{k} \left(1 - r_i^2\right) \tag{4.18}$$

Et donc utiliser l'approximation de Bartlett de cette statistique par la loi du χ^2,

$$-\left[n - \frac{1}{2}(p + q + 3)\right] \log \prod_{i=s+1}^{k} \left(1 - r_i^2\right) \sim \chi^2_{(p-s)(q-s)} \tag{4.19}$$

La règle de décision est la même que précédemment.

De cette manière, il nous est possible de tester chacun des coefficients de corrélation canonique et d'établir un sous-ensemble de coefficients significatifs.

Remarque 4.2.2.1 *Le test de signification des coefficients dépend des coefficients suivants. En effet, la statistique induit un produit sur tous les coefficients de corrélation canonique. Ainsi, cela ne suffit pas d'observer qu'un coefficient de corrélation est proche de* 1 *si on n'a pas effectué les tests de signification présentés.*

4.2.3 Redondance

Comme nous l'avons vu, les composantes canoniques ne peuvent expliquer qu'une petite partie de la variance de leur groupe et de l'autre groupe. Cette partie s'attache à évaluer la part de la variance expliquée par les variables canoniques. En d'autres termes, on s'efforce de calculer la proximité d'une composante canonique avec les variables de son groupe ou de l'autre groupe. On utilise classiquement l'appellation *redondance*, notée *Rd*, au lieu de proximité. Les indices sont calculés de la manière suivante,

1. Part de la variance de $\mathbf{X}$ expliquée par $\mathbf{t}_h$,

$$Rd(\mathbf{X}, \mathbf{t}_h) = \frac{1}{p} \sum_{j=1}^{p} \left[corr(\mathbf{x}_j, \mathbf{t}_h)\right]^2$$

2. Part de la variance de $\mathbf{X}$ expliquée par $\mathbf{u}_h$,

$$Rd(\mathbf{X}, \mathbf{u}_h) = \frac{1}{p} \sum_{j=1}^{q} \left[corr(\mathbf{x}_j, \mathbf{u}_h)\right]^2$$

De même pour les variables de $\mathbf{Y}$. Il s'ensuit que le calcul de la redondance de $\mathbf{X}$ par rapport à un ensemble de variables $\mathbf{u}_1, ..., \mathbf{u}_p$ sera la somme des redondances de $\mathbf{X}$ par rapport à chacune des variables précédemment citées.

4.3 Optimisation de la corrélation

Une originalité de la méthode tient dans une amélioration cruciale du calcul de la corrélation. On commencera par valider informellement son utilisation pour justifier sa présentation. On développera ensuite formellement la mise en place d'une telle optimisation.

4.3.1 Préambule

Le lecteur pourra vérifier dans le prochain chapître que la méthode de corrélation obtient de très bons résultats sous couvert d'une hypothèse. La projection des variables dans deux nouveaux espaces a vocation à améliorer la corrélation entre les composantes canoniques. Il ne faut pas oublier qu'une corrélation illustre le comportement commun de deux variables (ou plus) dans le temps. C'est à dire que c'est un calcul sur les observations, qui sont ici des échantillons. Dès lors, si deux variables sont égales mais décalées dans le temps, l'évaluation de la corrélation ne reflètera pas la relation entre ces variables. L'hypothèse est donc que le geste soit parfaitement en phase avec le signal audio.

Dans notre travail, le geste est effectué et saisi de manière synchrone avec l'écoute d'une séquence sonore. Inévitablement, le geste sera en retard par

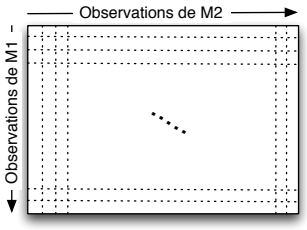


FIG. 4.2 – Structure globale de la matrice de coût : le nombre de lignes correspond au nombre d'observations de $\mathbf{M}^1$ et le nombre de colonnes aux observations de $\mathbf{M}^2$.

rapport au signal sonore. En outre, il peut parfois être en avance si l'utilisateur a voulu anticiper le son, ce qui peut se produire dans le cas de plusieurs écoutes préalables du même son avant la saisie réelle du geste.

4.3.2 Alignement temporel

Soient $\mathbf{M}^1$ et $\mathbf{M}^2$ deux matrices contenant le même nombre p de variables. Notre but est l'alignement temporel des variables contenues dans ces deux matrices. Pour ce faire, nous allons séparer la méthode en deux parties : le calcul d'une matrice de coût et la recherche du chemin de moindre coût dans cette matrice. On appelera *observations* les lignes de ces matrices afin d'établir un lien avec les matrices mises en jeu dans notre travail. Soient N_1 (resp. N_2) le nombre d'observations de $\mathbf{M}^1$ (resp. $\mathbf{M}^2$).

Calcul d'une matrice de coût

Une matrice de coût représente le coût de tous les chemins possibles lorsqu'on parcourt toute paire d'observations des matrices $\mathbf{M}^1$ et $\mathbf{M}^2$, ce qui est illustré de manière globale sur la figure [4.2].

On appelle $\mathbf{C}$ cette matrice. Les éléments de celle-ci représente le coût du passage par l'observation (i, j), soit l'observation i de $\mathbf{M}^1$ et j de $\mathbf{M}^2$. Ce coût induit une notion de distance entre ces deux observations. Une solution simple consiste à utiliser une distance euclidienne entre le vecteur $(m^1_{i,1}, ..., m^1_{i,p})$ et le vecteur $(m^2_{j,1}, ..., m^2_{j,p})$.

Concrètement, il s'agit de calculer la distance à un instant t entre la valeur des paramètres gestuels et les valeurs des descripteurs audio prises à chaque instant du signal.

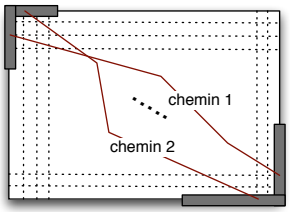
On est amené à la formulation suivante

$$\mathbf{C} = (c_{i,j})_{i=1..N_1, j=1..N_2} \quad \text{tel que} \quad c_{i,j} = \sum_{k=1}^{p} \left(m^1_{i,k} - m^2_{j,k} \right)^2 \tag{4.20}$$

Recheche du chemin de moindre coût

La seconde partie de la méthode consiste à parcourir la matrice **C** et à trouver le chemin de moindre coût. Pour ce faire, on utilise un algorithme de programmation dynamique qui consiste à trouver la solution d'un problème d'optimisation dont la fonction objectif se décrit comme la somme de fonctions monotones non-décroissantes des distances.

Ce type de résolution permet de définir des contraintes ou en enlever. En programmation dynamique, la méthode est de considérer qu'à la fin nous avons atteint notre but avec un coût minimal, et de revenir pas à pas, par résolutions successives de sous-problèmes, jusqu'au début. Alors, il est possible de définir où l'algorithme pourra commencer et où il pourra se terminer (on réfère le lecteur à la figure ci-contre où les parties grisées sont les degrés de liberté en début et fin de chemin).



Cependant, la relaxation des contraintes induit une coupure du signal initial. Ainsi, il est possible que le geste, ou le signal audio, soit tronqué. Cela peut ne pas être crucial dans la mesure où on peut expliquer pourquoi cette partie du geste ou du son n'était pas pertinente. Ici, le geste et le son sont synchronisés dans leur commencement et dans leur fin, il n'est donc pas utile de laisser ces degrés de liberté à l'algorithme.

5

Réalisation du Mapping et Validation

Des discussions formelles précédentes, où le modèle est justifié et les résultats théoriques analysés, se dégage une méthode concrète applicative pour le domaine musical. Cette méthode constitue le but de l'étude, à savoir un mapping geste/son adapté à la problématique. La CCA définit ce mapping et l'alignement temporel rectifie ce lien tressé par l'analyse canonique. Leur alliance est le processus de gestification du son. La définition formelle du mapping posée, et l'implémentation sous forme de processus calculable effectuée, ce chapître s'achèvera sur la validation expérimentale de la méthode.

5.1 Algorithme de mapping

Mathématiquement, un mapping est une application. Celle-ci s'avère être une fonction calculable. En effet, sa calculabilité (i.e. la possibilité d'avoir un programme informatique calculant le mapping en un temps fini) est assurée par la formalisation du mapping sous forme explicite.

5.1.1 Fonction de Mapping

On utilise les notations introduites dans le paragraphe précédent. De la phase d'analyse canonique alignée, on retient les données suivantes,

1. La matrice des composantes canoniques du son **U** (i.e. **Y**.**B**)
2. La matrice de projection **A** des données gestuelles

Soient N le nombre de paramètres gestuels considérés, et M le nombre de descripteurs audio calculés. On définit l'espace X par

$$X = \{(x_{i,1}, ..., x_{i,N})\}$$

C'est un espace vectoriel de dimension N isomorphe à $\mathbb{R}^N$.

Ainsi on définit la fonction $F_1 : X \longrightarrow X$, correspondant à la projection des données gestuelles dans la direction de **A**, par

$$F_1 : (x_{i,1}, ..., x_{i,N}) \mapsto (t_{i,1}, ..., t_{i,N}) = (x_{i,1}, ..., x_{i,N}) .\mathbf{A} \tag{5.1}$$

Tant que le nombre de paramètres gestuels est inférieur au nombre de descripteurs audio, soit $N < M$, l'espace image de la fonction F_1 est inclu dans X. D'autre part, la fonction $F_2 : X \longrightarrow X$, assurant l'alignement temporel, se définit par

$$F_2 : (t_{i,1}, ..., t_{i,N}) \mapsto \text{argmin}_{(u_{i,1},...,u_{i,N})} \sqrt{\sum_{k=1}^{N} (t_{i,k} - u_{i,k})^2} \tag{5.2}$$

Cette fonction permet, à un vecteur $(t_{i,1}, ..., t_{i,N})$ donné, de trouver le vecteur le plus proche dans X, en terme de distance euclidienne.

Alors la fonction de mapping est la composition d'une projection et d'un alignement. On note F_M cette fonction, elle se définit comme

$$F_M = F2 \circ F1 \tag{5.3}$$

Ou encore,

$$F_M\,(x_{i,1}, ..., x_{i,N}) = \text{argmin}_{(u_{i,1},...,u_{i,N})} \sqrt{\sum_{k=1}^{N} (\sum_{l=1}^{N} (x_{i,l}.a_{l,k}) - u_{i,k})^2} \tag{5.4}$$

L'implémentation consistera à la concrétisation de cette application.

5.1.2 Implémentation

L'algorithme général implémenté dans le cadre du stage se décompose en deux phases principales : une phase d'analyse et une phase de synthèse.

La phase d'analyse est schématiquement représentée sur la figure [5.1] et concerne **la représentation du son par le geste**. Elle suit le déroulement ci-dessous

1. Les descripteurs sonores sont extraits du signal audio
2. Le processus de gestification calcule des descripteurs "gestifiés" à partir des descripteurs sonores
3. Les descripteurs "gestifiés" décrivent un corpus sonore

On va expliciter l'étape fondamentale qui consiste au calcul des descripteurs "gestifiés". L'analyse permet la construction des espaces de projection, et on optimise la corrélation grâce à l'alignement temporel des composantes canoniques. Une première analyse canonique calcule les matrices de projection **A** et **B**, la matrice des coefficients de corrélation canonique **R** et enfin les variables centrées projetées **T** et **U**. Celles-ci sont ensuite alignées dans le temps. Il est important de noter que la convergence du processus d'itération n'est pas prouvée et par là certaines précautions d'implémentation sont à prendre en compte.

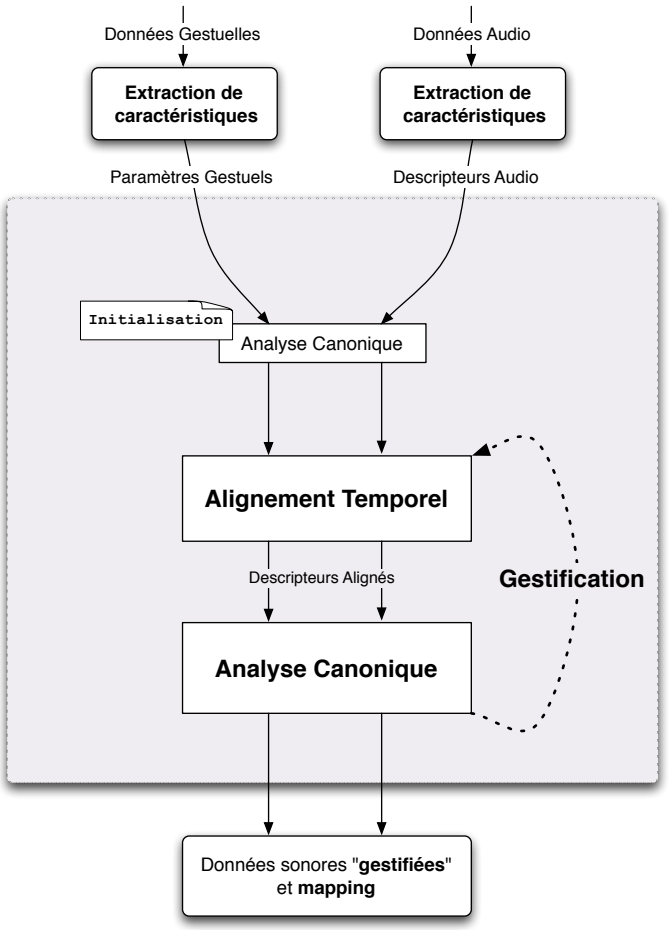


FIG. 5.1 – Schéma de l'algorithme d'analyse avec en entrée les paramètres gestuels et les descripteurs audio et en sortie les données de l'analyse.

La seconde partie concerne la synthèse, c'est à dire l'implémentation de la fonction de mapping. Le son est décrit par ses composantes canoniques, où chaque observation correspond à un grain de T millisecondes. Ce résultat de l'analyse est une donnée de l'algorithme de synthèse. Les paramètres gestuels entrent dans l'algorithme à une certaine fréquence F. Ce sont les éléments de G. Leur projection sera comparée à l'ensemble des observations des composantes canoniques du son (les éléments de D, en nombre fini pour une séquence), et il sera sélectionné le minimum en terme de distance euclidienne. Ce minimum correspond à un grain de la séquence sonore à lire. C'est ainsi que le moteur de synthèse granulaire est contrôlé par des descripteurs sonores de haut-niveau : les composantes canoniques sonores alignées (cf. figure [5.2]).

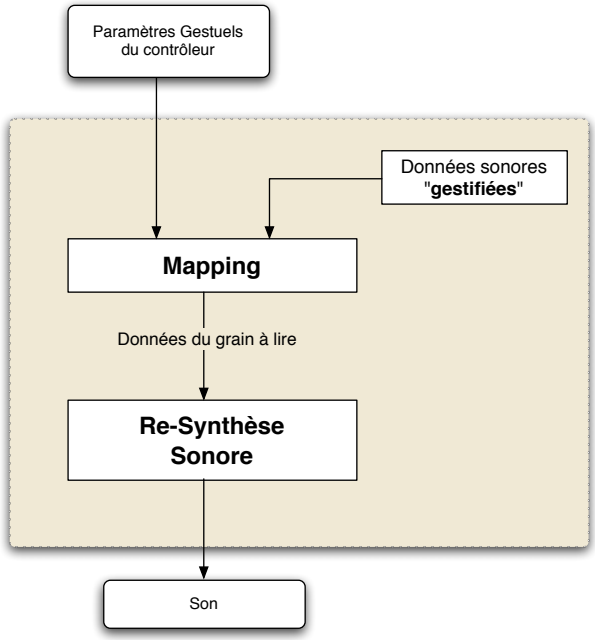


FIG. 5.2 – Schéma de l'algorithme de synthèse avec en entrée les paramètres gestuels. En temps réel le vecteur des paramètres entre dans l'algorithme avec une certaine fréquence. Le grain le plus proche de ce vecteur est sélectionné et entre dans un moteur de synthèse granulaire.

5.2 Validation de la méthode

Il s'agit ici de valider expérimentalement la méthode développée pendant la période de stage. On procédera par incrément. Dans une première partie, le mapping est imposé. On effectuera trois validations,

1. Le geste est synchrone avec le son : test de l'analyse canonique
2. Le son est dilaté par morceaux : test de l'alignement temporel
3. La temporalité est libre : test des deux incréments précédents réunis

Enfin, la seconde partie est un exemple où le son est quelconque et où la gestification est libre. Le mapping n'est plus imposé.

Tous les tests ont été effectués à partir de `Max/MSP` pour la saisie de geste et la synthèse sonore, et `Matlab` pour l'analyse numérique.

5.2.1 Mapping est imposé : sans alignement

Le test de validation met en jeu l'analyse canonique indépendamment de l'alignement temporel. Le mapping est imposé et c'est le geste qui synthétise

le son. Il s'agit ici d'établir le lien conforme à ce qu'on attendait entre des variables les mieux corrélées.

Principe de l'expérience

Dans l'expérience, le paramètre x des abscisses est relié linéairement à un oscillateur dont la fréquence évolue dans l'intervalle $[220, 660]$ Hertz.

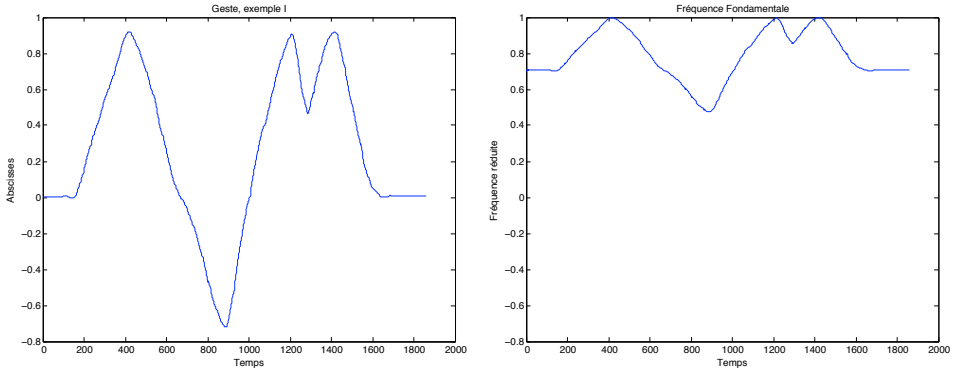


FIG. 5.3 – Geste pour un premier exemple de validation. A gauche, la ligne des temps est représentée en abscisses et les abscisses de la tablette graphique se placent en ordonnées. Le mapping imposé relie linéairement ce paramètre à la fréquence d'un oscillateur. À droite est l'évolution de la fréquence fondamentale en temps, non centrée et non normalisée.

Le geste effectué pour cette validation est rapporté par la figure [5.3]. A gauche, on a illustré le geste, et à droite l'évolution de la fréquence fondamentale. Le lien entre ces deux variables transparaît sur cette figure, et c'est ce qu'on cherche à valider par notre algorithme. Les données numériques utiles pour mieux comprendre l'expérience sont rapportées dans le tableau suivant.

	Paramètres Gestuels	Paramètres Sonores	Matrice **A**	Matrice **B**	Matrice **R**
Taille	1859×1	1859×155	1×1	155×1	1×1

L'unique paramètre gestuel est l'abscisse x de la tablette graphique, il n'y a donc qu'un descripteur "gestifié".

Résultats

L'analyse canonique fournit deux ensembles de composantes canoniques et une matrice de corrélation. Dans cet exemple, chaque groupe est constitué d'une unique variable corrélée par un unique coefficient.

	Max(corrélations simples)	Corrélation canonique
inter-groupe	`0.9983`	`0.9991`

Le paramètre gestuel x est initialement corrélé avec les descripteurs audio. Particulièrement, le maximum est atteint pour le descripteur `Fundamental Frequency`, en accord avec la figure [5.3]. L'analyse canonique nous garantie de trouver un coefficient de corrélation au moins égal, en valeur absolue, au maximum des coefficients de corrélation simple. Ce qui se retrouve dans le tableau.

Cela signifie que la composante canonique gestuelle se confond quasiment avec la composante canonique des descripteurs audio (l'angle formé entre les deux espaces de projection, en l'occurrence des droites, est arcos(0.9991) i.e environ $2,4^{\circ}$). Graphiquement, on visualise clairement cette propriété sur la figure [5.4].

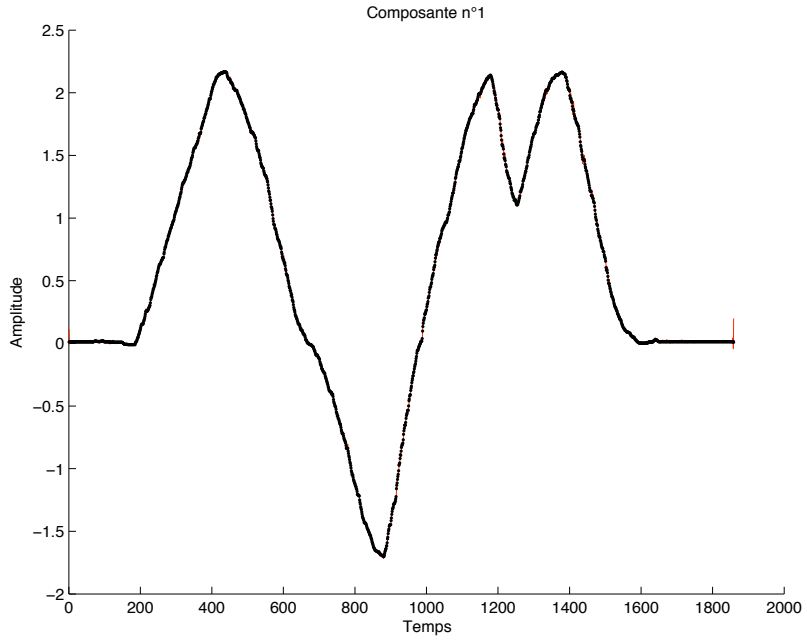


FIG. 5.4 – La composante canonique audio (ligne continue) décrit le son. La composante canonique gestuelle (ligne discontinue) est construite à partir du geste. La corrélation est excellente.

Interprétation

Il s'agit d'étudier la signification du coefficient de corrélation canonique obtenu. Comme c'est l'unique coefficient, sa signification est donnée par sa

valeur. Celle-ci étant très proche de 1, c'est un coefficient significatif dans l'explication de la covariance entre geste et son. Vérifions-le par le calcul.

La formule 4.16 induit d'une part,

$$-\left[n-\frac{1}{2}\left(p+q+3\right)\right]\log\prod_{i=s+1}^{k}\left(1-r_i^2\right)=5692.2$$

avec $n = 997, p = 1, q = 195$. D'autre part, pour un risque $\alpha = 5\%$, et un nombre de degrés de liberté $pq = 195$, le fractile [1] de $\chi^2_{195}(1-0.05)$ est 228.5799. Le premier coefficient est bel et bien significatif.

Ensuite, on aimerait savoir si le mapping a bien su retrouver que le paramètre x était lié à la fréquence fondamentale. Pour ce faire, on va inspecter la corrélation entre la composante canonique du son et le tableau de données des descripteurs audio. On cherche le maximum afin de savoir quel descripteur la composante canonique représente le plus, soit

$$\text{argmax}_{i=1\ldots M} corr(\mathbf{y}_i, \mathbf{u})$$

Celui-ci est atteint pour le descripteur `Fundamental Frequency` avec un coefficient de corrélation de 0.9994.

D'autre part, la composante canonique du son explique parfaitement la variance du paramètre gestuel x car le calcul de la redondance donne

$$Rd(\mathbf{X}, \mathbf{u}) = 0.9999$$

Alors qu'elle n'explique que très peu la covariance de son propre groupe,

$$Rd(\mathbf{Y}, \mathbf{u}) = 0.3458$$

La gestification du son a "filtré" les données afin de garder ce que le geste a trouvé comme significatif dans le son, quitte à perdre des niveaux de représentation.

Dans un cas de mapping imposé, l'analyse canonique nous permet bien de retrouver de manière quasi-parfaite le mapping pré-défini. C'est une première validation.

5.2.2 Mapping est imposé : avec alignement

En se basant sur une analyse canonique valide, on va tester l'alignement temporel en utilisant à la fois un mapping imposé, et un étirement temporel arbitraire.

[1] le fractile d'ordre p $(0 \leq p \leq 1)$ associé à une variable aléatoire X dont la fonction de répartition est $F(x)$ est une valeur x_p telles que :

$$F(x_p) = P(X \leq x_p) = p$$

Principe de l'expérience

On utilise le mapping de la partie précédente, c'est à dire qui à x associe la fréquence de l'oscillateur. Cependant, les descripteurs sonores vont être sous-échantillonnés ou sur-échantillonnés par morceaux.

Ainsi, le premier quart va voir sa fréquence d'échantillonnage divisée par 5. Ensuite, de ce premier quart jusqu'au deux-tiers du signal (soit 5/12 du signal), on ré-échantillonnage à 5/4. Puis la fréquence d'échantillonnage du dernier tronçon est multipliée par 7/2. Le nombre de descripteurs est inchangé en revanche, le nombre d'observations l'est.

Le geste reste le même que pour la première phase de validation (cf. figure [5.3]). Le but de l'expérience est que le geste, non modifié, s'aligne temporellement avec la matrice de descripteurs modifiée. On résume les caractéristiques numériques de l'expérience par le tableau suivant.

	Paramètres Gestuels	Paramètres Sonores	Matrice **A**	Matrice **B**	Matrice **R**
Taille	1859×1	3231×155	1×1	155×1	1×1

Résultats

La figure [5.5] montre deux figures : le chemin d'alignement dans la matrice coût (à gauche) et les représentations des composantes canoniques gestuelles et sonores.

Nous pouvons observer trois phases distinctes dans le chemin de moindre coût. Ces phases correspondent à l'alignement du geste sur un signal dilaté par morceaux.

- Première phase, le chemin parcourt un grand nombre d'observations du geste par rapport aux observations de descripteurs.
  Approximativement, les observations 1 à 450 du geste, soit environ 1/4 du signal, correspondent aux observations 1 à 100 des descripteurs, soit 1/5 du quart du signal.

- Deuxième phase, le chemin parcourt un nombre d'observations du geste à peu près équivalent aux observations de descripteurs.
  Approximativement, les observations 450 à 1200 du geste, soit environ 5/12 du signal, correspondent aux observations 100 à 1000 des descripteurs, soit 5/4 des 5/12 du signal.

- Troisième phase, le chemin parcourt un nombre d'observations du geste plus faible que les observations de descripteurs.
  Approximativement, les observations 1200 à 1800 du geste, soit environ

1/3 du signal, correspondent aux observations 1000 à 3200 des descripteurs, soit 7/2 des 2/3 du signal.

L'alignement temporel a parfaitement corrigé les dilatations effectuées sur les observations des vecteurs de descripteurs.

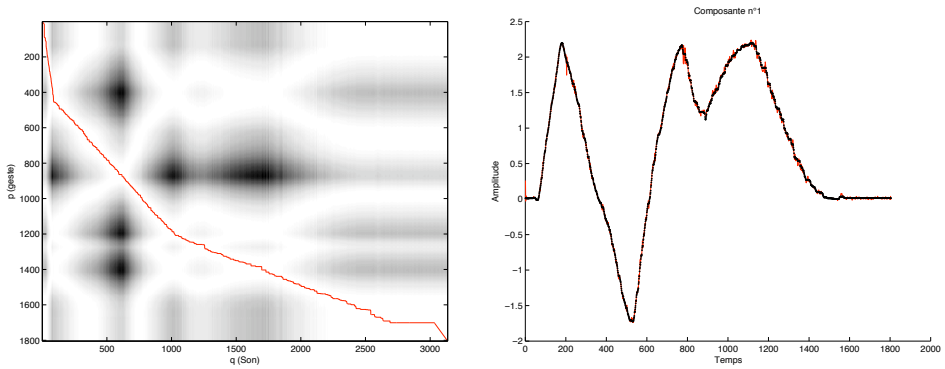


FIG. 5.5 – Test de l'alignement temporel : à gauche, le chemin de moindre coût dans la matrice comporte trois phases distinctes corrigeant les modifications sur le signal. À droite, les composantes canoniques gestuelles et sonores concordent.

On reporte ci-dessous le tableaux de valeurs du maximum des corrélations simples entre ces signaux, puis le coefficient de corrélation canonique avant l'alignement et enfin le coefficient après alignement.

Max(corrélations simples)	Corrélation canonique avant alignement	Corrélation canonique après alignement
`0.6396`	`0.8517`	`0.99966`

Interprétation

Les trois coefficients précédemment trouvés expliquent significativement la covariance inter-groupe. Seulement, leurs valeurs diffèrent sensiblement. La corrélation entre les composantes canoniques se trouve très améliorée par l'alignement temporel induisant un mapping plus précis.

A l'image du paragraphe précédent, on va calculer la corrélation entre la composante canonique du son générée et l'ensemble des descripteurs audio initiaux. Le maximum est atteint pour le descripteur `Fundamental Frequency` avec une corrélation de 0.9970. C'est encore un très bon résultat.

Le calcul des redondances donne,

$$Rd(\mathbf{X}, \mathbf{u}) = 0.9995$$

$$Rd(\mathbf{Y}, \mathbf{u}) = 0.3388$$

L'expérience obtient des résultats de l'ordre de l'exemple précédent. Cela implique que l'algorithme a pu aligner les variables très convenablement, et le mapping a été décelé avec un coefficient de corrélation proche de 1 ce qui valide la seconde phase de la méthode.

5.2.3 Mapping est imposé : temporalité libre

Il s'agit ici de réunir les deux premiers incréments, formant le processus de gestification illustré par la figure [5.1], afin de les tester conjointement.

Principe de l'expérience

Précédemment, on a pu valider la méthode d'analyse canonique et la méthode d'alignement temporel grâce à des données artificielles qu'on a reconstruit. Le protocole expérimental sera très similaire dans cette partie.

Le mapping est imposé. L'abscisse x est liée linéairement au centroïde spectral d'un signal synthétisé ; l'ordonnée y est liée linéairement à l'amplitude du signal de sortie. On commence par effectuer un geste prédeterminé et noté (on réfère le lecteur à la figure [5.6] afin de visualiser le schéma gestuel utilisé pour ce test). Celui-ci synthétisera un son suivant le mapping précédemment cité. On enregistre la sortie et ensuite on essaie de refaire le geste prédéterminé dans les conditions normales de captation du geste sur un son enregistré (cf. annexe B, patch B.1). Le geste doit être le plus fidèle possible à l'initial, ce qui ne pourra pas empêcher pas les éventuels retards et anticipations.

La figure [5.6] permet de visualiser les deux gestes. Le second étant très proche du premier, et étant effectué sur une séquence sonore enregistrée, synthétisée artificeillement par le premier geste, on devrait retrouver le mapping initial.
On reporte dans le tableau suivant quelques données numériques utiles à la compréhension de l'expérience.

	Paramètres Gestuels	Paramètres Sonores	Matrice **A**	Matrice **B**	Matrice **R**
Taille	2799×2	2799×155	2×2	155×2	2×2

Les paramètres gestuels sont : x et y.

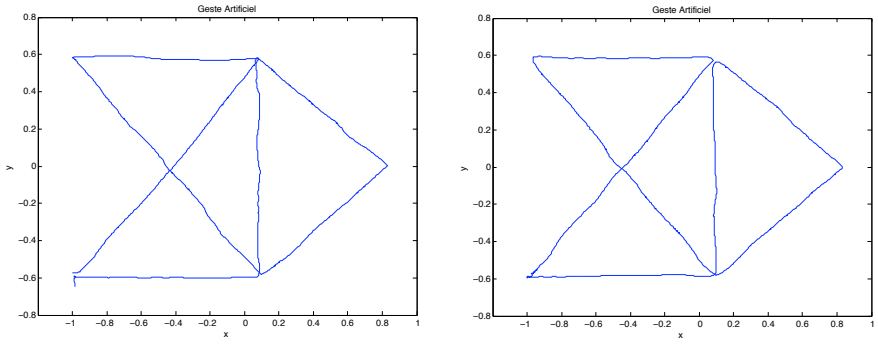


FIG. 5.6 – Gestes de validation. À gauche, la figure montre le geste lié au mapping imposé. À droite, le graphique représente le geste refait sur la séquence sonore, précédemment synthétisé, maintenant enregistré.

Résultats

Au vu du tableau de données précédent, sont induites 2 composantes canoniques relatives au groupe geste et 2 composantes canoniques relatives au groupe son.

Max(corrélations simples)	Corrélation canonique avant alignement	Corrélation canonique après alignement
0.9611	0.99612	0.99917
0.3136	0.98025	0.99574

La donnée des corrélations illustre bien la bonne approximation des composantes canoniques : les premières composantes canoniques gestuelle et sonore se corrèlent à $99,917\%$ alors que les secondes à $99,574\%$. En outre, l'alignement temporel améliore la corrélation. La figure [5.9] illustre l'alignement, en montrant que le chemin de moindre coût n'est pas tout a fait la diagonale, et la bonne approximation de la première composante canonique.

Interprétation

On peut se rendre tout de suite compte que les coefficients canoniques sont significatifs. En effet, on a les calculs suivant,

$$-\left[n-\frac{1}{2}\left(p+q+3\right)\right]\log\prod_{i=1}^{2}\left(1-r_i^2\right)=30368$$
$$-\left[n-\frac{1}{2}\left(p+q+3\right)\right]\log\prod_{i=2}^{2}\left(1-r_i^2\right)=12963$$

avec $n=3231, p=2, q=155$. D'autre part, pour un risque $\alpha=5\%$, et un nombre de degrés de liberté $pq=390$, le fractile de $\chi^2_{390}(1-0.05)$ est 437.05. Le premier et le deuxième coefficient sont bel et bien significatifs.

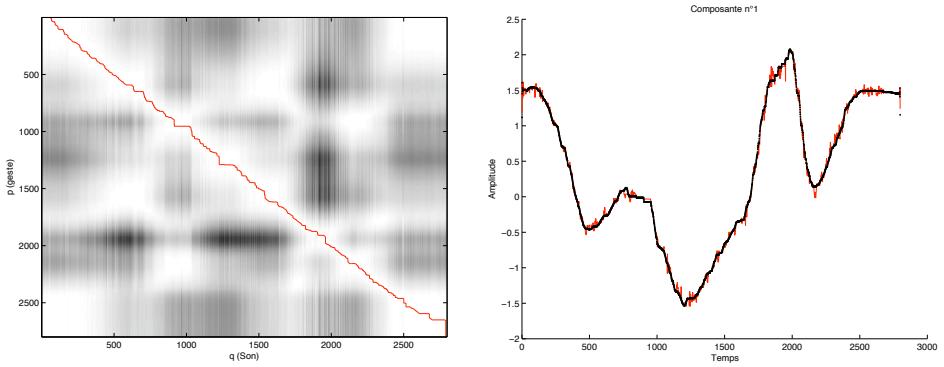


FIG. 5.7 – Schémas de correspondance entre les composanes canoniques. À gauche, la première composante canonique gestuelle se confond avec la première composante canonique relative au son. À droite, on a une très bonne corrélation des deuxièmes composantes canoniques, montrant que le mapping cherché a été bien approximé.

L'analyse canonique crée un espace dans lequel les vecteurs colonnes forment une base orthogonale. Dans cet exemple, les variables initiales x, y forment déjà une base orthogonale. De plus le mapping les découple en liant une au centroïde et l'autre à l'amplitude du signal. Ainsi, la projection en composantes canoniques devrait respecter cette base orthogonale. La matrice de projection $\mathbf{A}$ s'écrit,

$$\begin{pmatrix} -1.8411 & -0.1501 \\ 0.5068 & 2.2951 \end{pmatrix}$$

Et montre que la première composante canonique va plutôt solliciter la première variable x alors que la seconde composante canonique va solliciter la seconde variable y.

On calcule les corrélations entre composantes canoniques du son et tableaux de descritpeurs. Pour la première composante, on trouve une corrélation en valeur absolue égale à 0.9444 avec le descripteur `Spectral Centroid`. Pour la seconde composante, on obtient une corrélation égale à 0.8103 avec le descripteur `Noise Energy`.

En outre le calcul des redondances montre une baisse de l'explication des variances intra-groupes.

$$Rd(\mathbf{X}, \mathbf{u}) = 0.4809$$

$$Rd(\mathbf{Y}, \mathbf{u}) = 0.5140$$

On explique d'autant moins la variance du groupe que la corrélation est importante.

Ainsi, l'algorithme a retrouvé le mapping imposé initialement par l'intermédiaire d'un geste normal, se rapprochant du geste qui avait synthétisé le

son. Ceci valide la méthode globale de génération de mapping. La prochaine étape est l'application de l'algorithme pour un son quelconque.

5.2.4 Gestification libre

La phase finale de la validation de l'algorithme consiste à utiliser un son quelconque et un geste libre créé par l'utilisateur, celui-ci devant représenter au mieux le son par son intention gestuelle. Le mapping n'est pas imposé et donc méconnu *a priori*. L'algorithme nous retournera en sortie un mapping adapté qu'il sera important de tester auditivement.

Principe de l'expérience

Dans cette expérience, on utilise un fichier source de bruits d'eau : `water.wav`. C'est un fichier audio d'environ 2.5 secondes. On utilise le patch `Max/MSP` illustré en annexe par la figure [B.1]. Le fichier est importé et la saisie du geste se synchronise avec sa lecture. En fin de séquence, la saisie du geste s'arrête.

L'algorithme d'analyse récupère ces données gestuelles et les descripteurs audio du son `water.wav`. On présente ici les résultats de l'analyse canonique effectuée sur ces données. La séquence audio a trois parties récurrentes d'un bruit d'eau remuée par un objet (cf. figure [5.8], à gauche). Le geste a essayé de repérer ces bruits par trois gestes similaires vers le haut (cf. figure [5.8], à gauche).

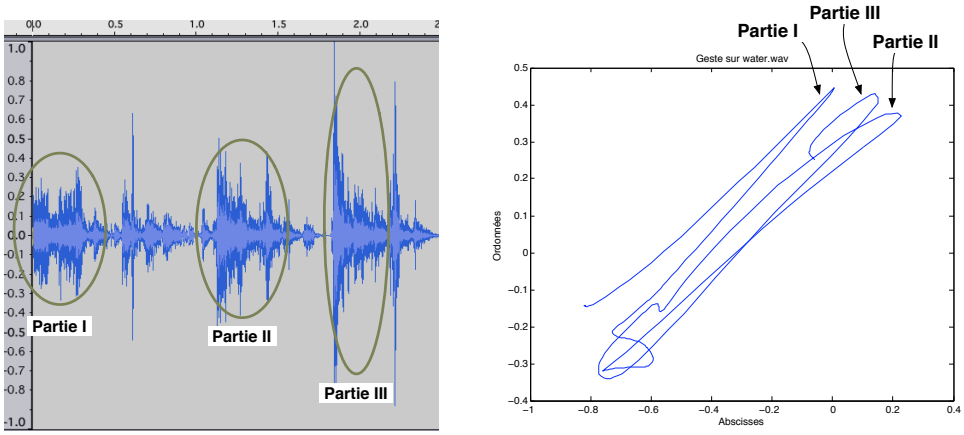


FIG. 5.8 – À gauche, le signal `water.wav` où on a entouré les parties récurrentes et significatives du signal. Ces parties ont été gestualisées (à droite) par le stylet et la tablette graphique, représentée par le plan 2-D.

Les données numériques relatives à cette expérience sont rapportées dans le tableau suivant,

	Paramètres Gestuels	Paramètres Sonores	Matrice **A**	Matrice **B**	Matrice **R**
Taille	250×8	250×155	8×8	250×8	8×8

Les huit paramètres du geste utilisés pour l'expérience sont : *vitesse*$_x$, *vitesse*$_y$, *accélération*, *norme de la position*, *vitesse angulaire*, *accélération angulaire*, *pression*, *dérivée première de la pression*. Il y a donc huit descripteurs "gestifiés".

Résultats

L'analyse canonique nous fournit deux ensembles de huit composantes canoniques, ayant des corrélations inter-groupe décroissantes. Pour l'expérience avec le fichier audio `water.wav`, le tableau suivant résume les données numériques des corrélations calculées.

Max(corrélations simples)	Corrélation canonique avant alignement	Corrélation canonique après alignement
0.4789	0.98187	0.98385
0.2067	0.97787	0.98080
0.0178	0.95463	0.95917
-0.0294	0.94106	0.94540
-0.0543	0.90092	0.91361
-0.1091	0.87744	0.89780
-0.1920	0.85378	0.86426
-0.2222	0.77958	0.78729

Ainsi, la première composante canonique du geste et celle du son sont très bien corrélées (à 98.39%). En revanche, la huitième composante canonique d'un groupe n'obtient pas le même score avec la huitième composante canonique de l'autre groupe : la corrélation subit un abattement d'environ 20%. On vérifie ceci graphiquement par la figure [5.9].

De plus on peut remarquer qu'il n'y pas, dans ce cas précis, de changements très importants entre l'analyse canonique et l'analyse canoniqe alignée temporellement. Ceci s'explique par une bonne synchronisation entre geste et son, qui se retrouve dans l'alignement, et donc dans la recherche d'un chemin optimal dans la matrice de coût (cf. figure [5.10]).

Interprétation

On a vu que les coefficients canoniques diminuaient significativement. Il est donc légitime de chercher à savoir s'ils restent significatifs dans l'explication

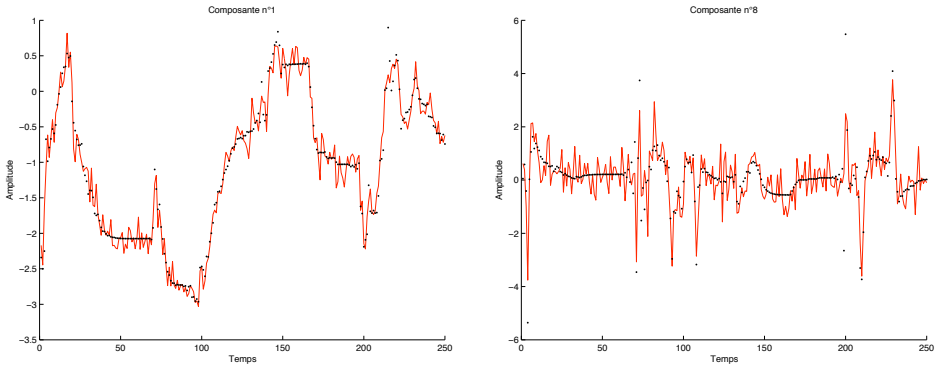


FIG. 5.9 – Schémas de correspondance entre les composanes canoniques. À gauche, la première composante canonique gestuelle (discontinue) s'approxime bien avec la première composante canonique relative au son (continue). À droite, la huitième composante canonique gestuelle suit en moyenne son homologue relatif au son, et ne suit pas un bon nombre de ses variations.

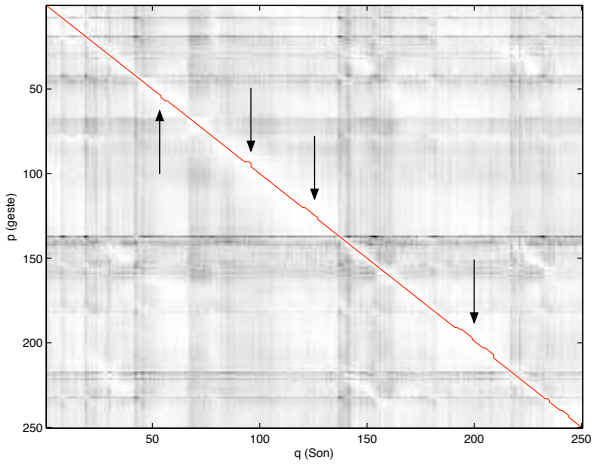


FIG. 5.10 – Chemin optimal dans la matrice de coût : la ligne est quasiment diagonale (les flèches indiquent les lieux de bifurcations), ce qui signifie que le geste et le son sont bien alignés.

de la covariance inter-groupe. Les tests d'hypothèse énoncés dans la partie précédente nous avaient amené à la relation suivante,

$$-\left[n-\frac{1}{2}\left(p+q+3\right)\right]\log\prod_{i=s+1}^{k}\left(1-r_i^2\right)\sim\chi^2_{pq}$$

Relation qu'on va vérifier pour chacun des coefficients canoniques trouvés précédemment dans l'analyse statistique. Le tableau ci-dessous résume la valeur du membre de gauche lorsque s augmente.

	$R_s = -\left[n - \frac{1}{2}(p+q+3)\right] \log \prod_{i=s+1}^{8} \left(1 - r_i^2\right)$
$s = 0$ (Coef I)	2882.33751
$s = 1$ (Coef II)	2720.82762
$s = 2$ (Coef III)	2491.34461
$s = 3$ (Coef IV)	2217.44319
$s = 4$ (Coef V)	1916.86125
$s = 5$ (Coef VI)	1542.40487
$s = 6$ (Coef VII)	1120.59338
$s = 7$ (Coef VIII)	574.61290

Chacune de ces valeurs doit être comparée avec le fractile f_α de $\chi^2_{1240}(1 - 0.05)$ pour un risque classique de $\alpha = 5\%$. Comme $f_\alpha = 1323.03$, il s'ensuit que $R_s > f_\alpha$ pour $s = 0..5$, en revanche, $R_s < f_\alpha$ pour $s = 6$ ou $s = 7$, c'est à dire que ces deux derniers coefficients ne sont pas significatifs. Si on diminue le risque α à 1%, $f_\alpha = 1358.8$ ce qui ne modifie pas la conclusion précédente.

Dans ce cas précis, il est difficile d'inspecter les corrélations entre composantes canoniques du son et le tableau de données des descripteurs car le mapping n'est plus aussi clair que dans les exemples précédents. On a obtenu les descripteurs haut-niveau. Par exemple, la première composante est corrélée avec un taux de 0.7165 au descripteur `Relative Specific Loudness` tandis que la seconde composante obtient un taux de 0.2484 avec le descripteur `Signal Auto Correlation`.

D'autre part, le calcul des redondances donne des résultats très faibles. Par exemple, pour la première composante canonique,

$$Rd(\mathbf{X}, \mathbf{u}_1) = 0.1665$$

$$Rd(\mathbf{Y}, \mathbf{u}_1) = 0.1205$$

Et il en va de même pour les composantes canoniques du geste avec son propre groupe. Par exemple, pour la première composante,

$$Rd(\mathbf{X}, \mathbf{t}_1) = 0.1710$$

On retrouve ici ce qu'on avait énoncé préalablement : une des difficultés pratiques de l'analyse canonique réside dans le fait que la maximisation de la corrélation induit une minimisation des variances des variables. On sacrifie l'explication des données initiales pour la corrélation de descripteurs de haut-niveau.

En conclusion, le mapping obtenu est très satisfaisant au vu de la valeur des coefficients de corrélation, proche de 1, et du nombre de coefficients significatifs, 6 sur 8.

6

Conclusion et Perspectives Futures

6.1 Conclusion

L'étude présentée est une nouvelle approche dans la création d'une fonction de mapping entre le geste et le son. La correspondance se focalise davantage sur les aspects récurrents de la séquence sonore plutôt que sur un suivi temporel du signal audio. En somme, on a pu voir que le mapping obtenu est une représentation du son par le geste. Pour ce faire, la méthode utilisée s'inspire des statistiques en adaptant un alignement temporel sur les observations de variables aléatoires générées par une analyse canonique. Le formalisme de cette méthode permet d'établir rigoureusement le cadre d'application de la méthode et donc d'en extraire les limitations.

Le travail proposé possède deux limitations majeures. Tout d'abord, la restriction au cas linéaire de la corrélation entre deux ensembles de variables. En effet, la corrélation canonique ne permet de mettre à jour que les combinaisons linéaires existantes entre les variables d'un ensemble explicatif et les variables d'un ensemble à expliquer. Néammoins, un avantage a été de nous permettre de mieux comprendre la méthode grâce à son interprétation géometrique simple, et on a pu implémenter rapidement l'algorithme afin de pouvoir valider les résultats expérimentaux. Cependant, une permière solution simple à cette limitation peut-être trouvée en adaptant une méthode des noyaux à l'analyse canonique : la *kernel canonical correlation analysis* (KCCA).

En deuxième lieu, une autre limitation porte sur la temporalité du mapping. La relation cherchée se focalise sur des données instantanées et le mapping est en ce sens instantané. La méthode ne prend pas en compte les notions de continuité temporelle dans le processus de mapping alors que les données possèdent une telle cohérence temporelle. On discutera ce point dans le paragraphe suivant.

Ce travail a pourtant obtenu des résultats très encourageants pour des travaux futurs. La validation effectuée a permis de mettre en avant que l'analyse canonique contribuait effectivement à trouver une correspondance entre geste et son. Pourtant, la méthode d'analyse canonique n'avait jamais encore

été appliquée dans ce champ de recherche. Par ailleurs, l'alignement temporel contribue à rendre l'algorithme plus souple, et devient tout à fait adapté à retrouver des similarités temporelles entre deux ensembles de composantes canoniques. Les idées de poursuite abondent, et nous allons en présenter une partie dans le paragraphe suivant.

6.2 Perspectives Futures

La première étape d'un travail futur serait l'établissement d'une plus grande base de données de gestes alliés au son de la manière présentée dans ce document. Un plus grand panel d'expériences mettant en jeu plusieurs sujets se représentant les sons différemment, avec des séquences sonores très différentes, permetterait d'enrichir l'étude à plusieurs niveaux.

- Jusqu'à présent, on ne peut rien dire sur la validité des descripteurs gestuels et sonores de plus haut niveau que les paramètres gestuels saisis ou les descripteurs sonores classiques. Nous avons d'ailleurs très peu écrit à ce sujet dans ce document. Des études comparatives grâce à plusieurs sujets, donnant des interprétations différentes de séquences sonores, permetteraient de mettre en exergue une validité de ces descripteurs, et donc d'améliorer leur utilisation dans un contexte d'écoute et de performance.

- D'autre part, ces données expérimentales souligneraient d'éventuelles invariances, existantes pour chaque sujet, dans la description haut niveau des données. On obtiendrait de plus amples informations quant au mode d'écoute des sujets et à leur interprétation des phénomènes sonores.

- Enfin, la remarque précédente amène à prolonger l'étude sur les invariances entre plusieurs sujets. Cette étude amènerait à caractériser des descripteurs sonores initiaux par leur sollicitation pour ensuite créer de nouvelles classes de descripteurs "gestifiés".

Par ailleurs, tout au long du projet de stage nous nous sommes beaucoup intéressés au *geste effecteur*. Producteur du son, il ne reflète pas le geste dans sa globalité (comme indiqué dans la partie préliminaire à l'étude, cf. chapître 2). L'hypothèse de synchronicité geste/son était pertinente pour le cas d'une geste effecteur, cependant, le geste instrumentale met aussi en jeu des gestes d'anticipation qui réfutent totalement cette hypothèse. L'anticipation du geste est par définition avant le geste (effecteur) lui-même, donc la prise en compte de ce type de geste nécessiterait une modification du paradigme de synchronicité.

Une perspective serait la fusion (ou l'inspiration) de la méthode de suivi de geste avec la méthode de gestification. En effet, là où la première s'intéresse au suivi temporel d'un geste par rapport à un son enregistré, la seconde

s'intéresse aux données significatives du son et mise en évidence par le geste. Il se dessine ici une complémentarité très enrichissante qu'il serait intéressant de mettre en oeuvre.

Bibliographie

Bevilacqua, Frédéric, Guédy, Fabrice, Schnell, Norbert, Fléty, Emmanuel, and Leroy, Nicolas (2007). Wireless sensor interface and gesture-follower for music pedagogy. New York, USA. NIME07 : Proceedings of the 2007 Conference on New Interfaces for Musical Expression.

Bevilacqua, Frédéric, Müller, Rémy, and Schnell, Norbert (2005). Mnm : a max/msp mapping toolbox. Vancouver, Canada. NIME05 : Proceedings of the 2005 Conference on New Interfaces for Musical Expression.

Bilmes, Jeff (2002, January). What hmms can do. Technical report, University of Washington, Department of EE, Seattle WA, 98195-2500.

Borga, M., Knutsson, H., and Landelius, T. (1997). Learning canonical correlations. *SCIA'97*.

Bévilacqua, F. (2005). A gesture follower for performing arts. Gesture Workshop.

Cadoz, Claude (1988). Instrumental gesture and musical composition. San Francisco, USA, pp. 1–12. Proceedings of the 1988 International Computer Music Conference.

Cadoz, Claude (1999). Musique, geste, technologie. *Les nouveaux gestes de la musique*, 47–92.

Cadoz, Claude and Wanderley, Marcelo M. (2000). Gesture - music. *Trends in Gestural Control of Music*, 71–94.

Cont, Arshia, Coduys, Thierry, and Henry, Cyrille (2004). Augmented mapping : Towards an intelligent user-defined gesture mapping. Sound and Music Conference.

Delalande, François (1988). La gestique de gould : éléments pour une sémiologie du geste musical. *G.Guertin, ed. Glenn Gould, Pluriel*, 83–111.

Dietterich, Thomas G. (2002). Machine learning for sequential data : a review. *Lectures Notes in Computer Science 2396*, 15–30.

Duda, Richard O., Hart, Peter E., and Stork, David G. (2000). *Pattern Classification (2nd Edition)*. Wiley-Interscience, ISBN 0-471-05669-3.

Fels, S. and G. E. Hinton (1995). Glovetalkii : An adaptive gesture-to-formant interface. In *CHI*, pp. 456–463.

Hardoon, David R., Szedmak, Sandor, and Shawe-Taylor, John (2003). Canonical correlation analysis : an overview with application to learning methods. Technical report, University of London, Department of Computer Science, Egham, Surrey TW20 0EX, England.

Hotelling, Harold (1936, December). Relations between two sets of variates. *Biometrika 28*(3/4), 321–377.

Hummels, Caroline, Smets, Gerda, and Overbeeke, Kees (1998). An intuitive two-handed gestural interface for computer supported product design. *Gesture and Sign Language in Human-Computer Interaction*, 197–208.

Hunt, A., M. Wanderley, and M. Paradis (2002). The importance of parameter mapping in electronic instrument design. pp. 149–154. NIME02 : Proceedings of the 2002 Conference on New Interfaces for Musical Expression.

Hunt, Andy and Kirk, Ross (2000). Mapping strategies for musical performance. *Trends in Gestural Control of Music*, 259–268.

Hunt, Andy and Wandereley, Marcelo M. (2002). Mapping performer parameters to synthesis engines. *Organised Sound 7*, 97–108.

Iazzetta, Fernando (2000). Meaning in musical gesture. *Trends in Gestural Control of Music*, 259–268.

Kidron, Einat, Schechner, Yov Y., and Elad, Michael (2005, June). Pixels that sound. *IEEE Computer Vision & Pattern Recognition (CVPR 2005) 1*, 88–95.

Kurtenbach, G. and Hulteen, E.A. (1990). Gestures in human-computer communication. *The Art of Human-Computer Interface Design*, 309–317.

Laban, Rudolf (1994). *La Maîtrise du Mouvement.* Arles : Actes Sud.

Lambert, J.-P. (2004). Projet phase jouer de la musique avec un bras haptique.

Lee, Matthew and Wessel, David (1992). Connectionist models for real-time control of synthesis and compositionnal algorithms. San Francisco, USA, pp. 277. ICMC92 : Proceedings of the 1992 International Computer Music Conference.

Louppe, Laurence, Dobbels, Daniel, Virilio, Paul, and Thom, René (1991). *Danses Tracées.* Paris : Edition Dis Voir. Livre édité à l'occasion de l'exposition "Danses tracées" réalisée par les musées de Marseille, à l'initiative du Centre National de la Danse Contemporaine d'Angers.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979). *Multivariate Analysis.* London : Academic Press.

Métois, E. (1996). *Musical Sound Information - Musical Gestures and Embedding Systems.* Ph. D. thesis, Massachusetts Institute of Technology.

Modler, Paul, Myatt, Tony, and Saup Michael (2003). An experimental set of hand gestures for expressive control of musical parameters in realtime. Montreal, Canada. NIME03 : Proceedings of the 2003 Conference on New Interfaces for Musical Expression.

Peeters, Geoffroy (2004). A large set of audio features for sound description. *CUIDADO Project*.

Peeters, Geoffroy and Rodet, Xavier (2002). Automatically selecting signal descriptors for sound classification. *CUIDADO*.

Peirce, C. S. (2006). Peirce's theory of signs.

Rovan, J., M. Wanderley, S. Dubnov, and P. Depalle. Instrumental gestural mapping strategies as expressivity determinants in computer music performance. KANSEI - The Technology of Emotion, AIMI International Workshop, Genova.

Schnell, Norbert (2005). Ftm - complex data structures for max/msp. Barcelona, Spain. ICMC-05 : Proceedings of the International Computer Music Conference.

Schnell, Norbert and al. (2005, September). Gabor, multi-representation real-time analysis/synthesis. Madrid, Spain. DAFx-05 : Proceedings of the 8th International Conference on Digital Audio Effects.

Schölkopf and Smola (2002). *Learning with Kernels*. Cambridge, Massachusetts : The MIT Press, ISBN 0-262-19475-9.

Schwarz, Diemo (2000, December). A system for data-driven concatenative sound synthesis. Verona, Italy. DAFx-00 : Proceedings of the COST G-6 Conference on Digital Audio Effects.

Schwarz, Diemo (2005, September). Current research in concatenative sound synthesis. Barcelona, Spain. ICMC-05 : Proceedings of the International Computer Music Conference.

Shlens, J. (2005). A tutorial on principal component analysis.

Tenenhaus, Michel (1998). *La régression PLS, Théorie et Pratique*. Paris : Éditions Technip, ISBN 2-7108-0735-1.

Tenenhaus, Michel (2007). *Statistique, méthodes pour décrire, expliquer et prévoir*. Paris : Éditions Dunod.

T.Hastie, R.Tibshirani, and J.Friedman (2001). *The Elements of Statistical Learning*. Springer-Verlag, New York : Springer Series in Statistics, ISBN 0-387-95284-5.

Traube, Caroline, Depalle, Philippe, and Wanderley, Marcelo M. (2003). Indirect acquisition of instrumental gesture based on signal, physical and perceptual information. Montreal, Canada. NIME03 : Proceedings of the 2003 Conference on New Interfaces for Musical Expression.

Van Nort, Doug and Wanderley, Marcelo (2006, May). Exploring the effect of mapping trajectories on musical performance. Marseille, France. SMC 06 : Poceedings in the International Conference of Sound and Music Computing (SMC06).

Van Nort, Doug, Wanderley, Marcelo, and Depalle, Philippe (2004). On the choice of mappings based on geometric properties. Hamamatsu, Japan. NIME04 : Proceedings of the 2004 Conference on New Interfaces for Musical Expression.

Wanderley, Marcelo, Schnell, Norbert, and Rovan, Joseph (1998). Escher - modeling and performing composed instruments in real-time. pp. 1040–1044. IEEE : Proceedings IEEE International Conference on Systems, man, Cybernetics.

Wanderley, Marcelo M. and Depalle, Philippe (2004, no. 4, April). Gestural control of sound synthesis. Volume 92, pp. 632. IEEE.

Wanderley, Marcelo M., Depalle, Philippe, and Rodet, Xavier (1999). Contrôle gestuel de la synthèse sonore. *Interfaces Homme-Machine et Création Musicale*, 145–163.

A

Analyse Canonique

A.1 Compléments en probabilité

A.1.1 Estimation

Dans la théorie de l'estimation, une notion fondamentale est : la fonction de vraisemblance. Elle se définit comme suit.

Fonction de Vraissemblance
Supposons que $x_1, x_2, ..., x_n$ soient des observations d'une population dont la fonction de densité est $f(\boldsymbol{x}; \boldsymbol{t})$ où $\boldsymbol{t}$ est un vecteur de paramètres. La fonction de vraisemblance est définit par

$$L(\mathbf{X}; \mathbf{t}) = \prod_{i=1}^{n} f(x_i; \mathbf{t})$$

Ce qui nous permet d'écrire la définition du test du rapport des vraisemblance utilisée dans l'obtention d'une statistique inconnue.

Test du rapport des vraisemblances
Si la distribution d'une variable aléatoire $\boldsymbol{X} = (x_1, x_2, ...x_n)$ dépend d'un vecteur de paramètre $\boldsymbol{t}$, et si $H_0 : \boldsymbol{t} \in \Omega_0$ et $H_1 : \boldsymbol{t} \in \Omega_1$ sont deux hypothèses, alors la statistique du rapport des vraisemblances pour tester H_0 contre H_1 est définie par,

$$\lambda(\mathbf{x}) = \frac{L_0*}{L_1*}$$

Où L_i* est le maximum de la fonction de vraisemblance sur l'ensemble Ω_i, $i = 0$ ou $i = 1$.

A.2 Démonstrations

Interprétation Géometrique

Énoncé :

1. La composante $\mathbf{X}a_j$ est colinéaire à la projection de $\mathbf{Y}b_j$ sur l'espace engendré par les colonnes de X
2. La composante $\mathbf{Y}b_j$ est colinéaire à la projection de $\mathbf{X}a_j$ sur l'espace engendré par les colonnes de Y

Démonstration
De l'équation 4.6 on déduit,

$$\mathbf{C}_{xy}\mathbf{b}_h = \lambda_x \mathbf{C}_{xx}\mathbf{a}_h$$

Qui peut être aussi bien écrit,

$$\mathbf{X}^T\mathbf{Y}\mathbf{b}_h = \lambda_x \mathbf{X}^T\mathbf{X}\mathbf{a}_h$$

$\mathbf{X}^T\mathbf{X}$ est symmétrique positive et $\mathbf{X}$ est de rang maximum, donc $\mathbf{X}^T\mathbf{X}$ est aussi définie. Ceci nous permet de reformuler l'équation précédente par,

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{b}_h = \lambda_x \mathbf{a}_h$$

Soit, le résultat pour la démonstration de la première relation :

$$\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{b}_h = \lambda_x \mathbf{X}\mathbf{a}_h$$

La seconde relation se démontre de manière tout à fait similaire en utilisant la deuxième équation de 4.6.

A.3 Données

X1	X2	X3
0.1411	0.2647	0.0902
0.3133	0.5480	0.2792
0.3178	0.4404	0.2904
0.3748	0.4563	0.2915
0.3674	0.4262	0.3611
0.4091	0.4250	0.3718
0.4075	0.3949	0.3755
0.4284	0.2975	0.3848
0.4355	0.2938	0.3761
0.5088	0.2360	0.3791
0.5153	0.2779	0.3938
0.4975	0.5265	0.4799
0.5138	0.5477	0.4993
0.5589	0.6452	0.5390
0.5700	0.5667	0.5120
0.6307	0.2319	0.4456

0.6417	0.2277	0.4499
0.6519	0.2232	0.4436
0.6656	0.2326	0.4550
0.6778	0.2730	0.4972
0.6785	0.2681	0.4958
0.3407	0.5268	0.5249
0.3247	0.5708	0.5139
0.7878	0.4528	0.6099
0.7348	0.3986	0.5519
0.8016	0.3312	0.5409
0.3686	0.1559	0.2659

Y1	Y2	Y3	Y4	Y5	Y6
2.8590	1.4993	12.8252	-0.0066	-0.0097	0.0041
7.6873	4.3630	28.1046	-0.0105	-0.0169	0.0095
7.1724	5.2962	24.2715	-0.0072	-0.0094	0.0080
6.4657	4.5108	26.4528	-0.0102	-0.0134	0.0089
8.2374	6.1774	25.1082	-0.0058	-0.0082	0.0084
8.2662	6.6360	25.7280	-0.0056	-0.0077	0.0085
8.2315	6.3870	25.4966	-0.0057	-0.0079	0.0086
9.2915	7.6125	25.8335	-0.0031	-0.0046	0.0087
9.7306	8.1253	25.8246	-0.0034	-0.0048	0.0088
7.7590	5.8021	25.6479	-0.0069	-0.0103	0.0086
9.0823	7.5322	25.8534	-0.0044	-0.0060	0.0088
6.9920	3.9363	25.4347	-0.0089	-0.0142	0.0085
9.1442	6.3613	26.0727	-0.0055	-0.0091	0.0090
8.1695	5.2208	25.5660	-0.0067	-0.0106	0.0086
11.3938	9.2072	26.2435	-0.0015	-0.0032	0.0091
7.8067	5.2705	25.3735	-0.0072	-0.0111	0.0085
9.0817	5.6016	26.0846	-0.0055	-0.0100	0.0090
8.0575	5.7875	25.3681	-0.0060	-0.0091	0.0085
8.4221	5.4787	26.1329	-0.0059	-0.0101	0.0090
8.3141	5.9539	25.3306	-0.0058	-0.0088	0.0084
8.3361	7.0194	26.3118	-0.0050	-0.0060	0.0090
11.0981	8.1214	25.2568	-0.0018	-0.0046	0.0085
10.2635	7.4048	26.4352	-0.0029	-0.0061	0.0091
7.4425	5.3429	24.9257	-0.0067	-0.0090	0.0083
10.9384	7.1123	26.8554	-0.0036	-0.0075	0.0093
7.5450	5.3739	24.6422	-0.0064	-0.0085	0.0083
8.8654	6.9909	28.2389	-0.0058	-0.0078	0.0096

	X_1	X_2	X_3	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
Moyenne	0.506	0.379	0.423	8.395	6.079	25.39	-0.006	-0.009	0.009
Variance	0.028	0.019	0.013	2.751	2.401	7.068	$4.683e^{-6}$	$9.263e^{-6}$	$9.475e^{-7}$

	X_1	X_2	X_3
X_1	1.0000	-0.2160	0.7609
X_2	-0.2160	1.0000	0.3013
X_3	0.7609	0.3013	1.0000

FIG. A.1 – Corrélation simple intra-groupe X

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
Y_1	1.0000	0.9133	0.6912	0.7220	0.5848	0.7383
Y_2	0.9133	1.0000	0.5909	0.7969	0.7856	0.6341
Y_3	0.6912	0.5909	1.0000	0.0578	0.0074	0.9938
Y_4	0.7220	0.7969	0.0578	1.0000	0.9381	0.1278
Y_5	0.5848	0.7856	0.0074	0.9381	1.0000	0.0653
Y_6	0.7383	0.6341	0.9938	0.1278	0.0653	1.0000

FIG. A.2 – Corrélation simple intra-groupe Y

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
X_1	0.2641	0.1697	0.3413	0.0811	0.0658	0.3548
X_2	0.2117	0.0669	0.1433	-0.0081	-0.1174	0.1437
X_3	0.5978	0.4436	0.4640	0.3726	0.2621	0.4870

FIG. A.3 – Corrélation simple inter-groupe

	Coefficient de corrélation	erreur-standard
C_1	0.8714	0.0463
C_2	0.6784	0.1039
C_3	0.5080	0.1428

FIG. A.4 – Coefficients de corrélation canonique entre les variables X_1, X_2, X_3 et les variables $Y_1, Y_2, Y_3, Y_4, Y_5, Y_6$

Ensuite on obtient les matrices de projections suivantes

$$\mathbf{A} = \begin{pmatrix} 9.2456 & 7.6520 & -5.8661 \\ 7.3199 & 7.5600 & 3.3180 \\ -18.9818 & -4.9377 & 4.1362 \end{pmatrix}$$

$$\mathbf{B} = \begin{pmatrix} 0.456 & 3.758 & 0.826 \\ 1.078 & -2.607 & 2.214 \\ -1.878 & -0.906 & 0.267 \\ -1445.144 & -2798.115 & -195.206 \\ 127.451 & 1612.027 & -953.238 \\ 3381.488 & 1292.700 & -3871.721 \end{pmatrix}$$

B

Les différents patchs

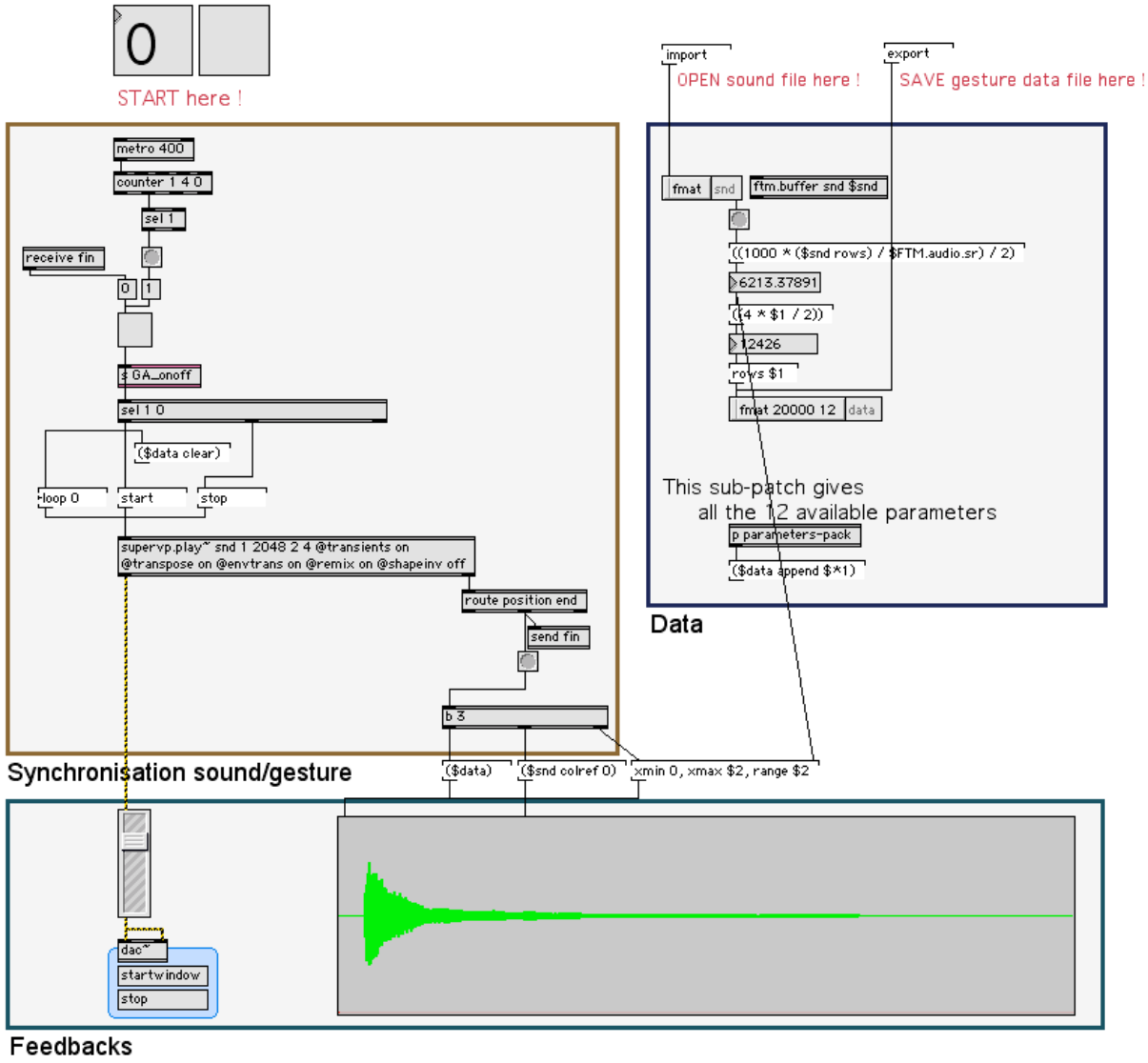


FIG. B.1 – Patch `Max/MSP`. Patch de saisie du geste sur une écoute d'un son enregistré. On assure la synchronisation des deux au début et à la fin.

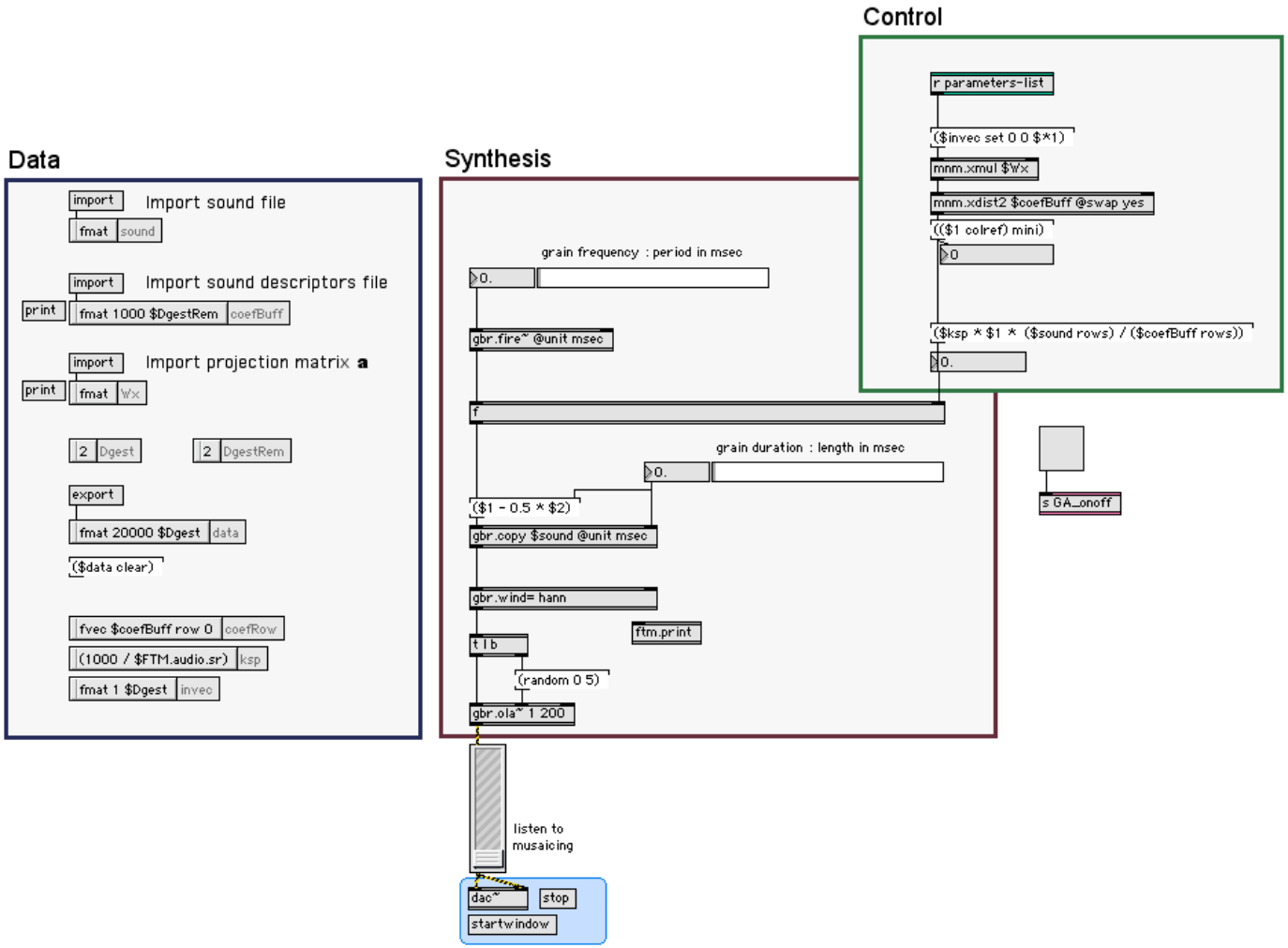


FIG. B.2 – Patch `Max/MSP`. Patch de synthèse : on importe la matrice de descripteurs audio associée à un son enregistré, qu'on importe aussi. De même on demande la matrice de projection **A**. Les données gestuelles entrent en temps réel, à la fréquence 100Hz, sont projetées et comparées à la matrice de descripteurs. L'échantillon le plus proche en terme de distance euclidienne est joué.