THÈSE DE DOCTORAT DE L'UNIVERSITÉ PIERRE ET MARIE CURIE - PARIS 6

École Doctorale EDITE

Mention

ACOUSTIQUE, TRAITEMENT DU SIGNAL ET INFORMATIQUE APPLIQUÉS À LA MUSIQUE

Présentée et soutenue par

Baptiste CARAMIAUX

ÉTUDES SUR LA RELATION GESTE–SON EN PERFORMANCE MUSICALE

Thèse dirigée par
Carlos AGON

et encadrée par
Frédéric BEVILACQUA et Norbert SCHNELL

préparée à l'Ircam – Centre Pompidou

soutenue le 14 décembre 2011

JURY

Carlos AGON	Université de Paris 6	*Directeur de thèse*
Thierry ARTIÈRES	Université de Paris 6	*Président du jury*
Frédéric BEVILACQUA	Ircam – CNRS	*Encadrant*
Sergi JORDÀ	University of Pompeu Fabra	*Examinateur*
Marc LEMAN	Ghent University	*Rapporteur*
Stephen MCADAMS	McGill University	*Examinateur*
David WESSEL	University of Berkeley	*Rapporteur*

Laboratoire d'accueil :
UMR STMS Ircam CNRS
1 place Igor Stravinsky
75004 Paris, FRANCE

Équipe *Interactions Musicales Temps Réel*

Résumé

Cette thèse présente plusieurs travaux autour de l'étude des relations entre *geste* humain et *son* enregistré qui ont pour but l'aide à la conception d'instruments numériques expressifs pour la performance musicale. L'étude de ces relations met en lien plusieurs domaines de recherche, conduisant à un travail multidisciplinaire.

Initiée par une étude exploratoire scellant les objectifs et les problématiques abordées, cette thèse s'articule autour de deux thématiques majeures : la réponse gestuelle à l'écoute d'un son (nous dirons stimulus sonore) et la modélisation du geste à des fins d'analyse et de contrôle de la synthèse.

Dans la première thématique nous menons des études expérimentales montrant les stratégies cognitives des participants dans l'exercice d'associer un geste à un son écouté. Nous montrons que ces stratégies sont liées à l'identification de la source. Lorsque la source causale n'est pas identifiable, les stratégies de correspondance entre des paramètres du geste et des paramètres du son varient.

Dans la deuxième thématique, nous abordons les problèmes de modélisation des structures temporelles dans le geste musical. Nous présentons un premier modèle permettant le suivi et la reconnaissance en temps réel des profils temporels des paramètres du geste. Motivés par les aspects structurels de la musique, nous montrons la pertinence de l'utilisation d'un modèle pour la segmentation et l'analyse syntaxique du geste. Ainsi, nous abordons l'analyse gestuelle d'un point de vue *signal* et *symbolique*.

Enfin, une implémentation concrète est présentée sous la forme d'applications des différentes contributions théoriques. Précisément, les applications sont : un système de sélection de son par le geste ; et un système de synthèse du son basée sur un suivi morphologique.

Mots clés : Instrument de musique numérique, Perception du son, Étude comportementale, Modélisation du geste, Inférence Bayésienne, Systèmes interactifs.

Abstract

This thesis presents the studies on the analysis of the relationship between gesture and sound with the aim to help with the design of digital expressive instruments for musical performance. Studies of these relationships are related to various areas of research and lead to a multidisciplinary approach.

We initiate the thesis by presenting an exploratory study sealing the objectives and issues. This thesis focuses on two main themes : the gesture response to sound stimuli and the modeling of gesture for analysis and control.

Within the first theme, we propose experimental studies showing the cognitive strategies of participants when they associate gestures to sounds they hear. First, we show that these strategies are related to the level of identification of the causal sound source. Then, when the causal source is not identifiable, relationship strategies vary in the correspondence between the parameters of both the gesture and the sound.

Within the second theme, we address the problem of modeling the musical gesture temporal structures. We present a first model for tracking and recognizing in real-time the temporal profiles of gesture parameters. Motivated by the structural aspects of music, we show the relevance of using a segmental-based Markov model for segmenting and parsing musical gesture. Thus, we approach the analysis of gesture from a signal point of view to a symbolic point of view.

Finally, applications of different theoretical contributions are presented. They are proofs of concept aiming at practically illustrating the specific research questions. Precisely, the two applications are : a system of sound selection based on gesture query ; and a system of sound re-synthesis based on morphological monitoring.

Keywords : Digital musical instruments, Sound perception, Behavioral study, Gesture modeling, Bayesian inference, Interactive systems.

Remerciements

Je voudrais tout d'abord remercier les membres de mon jury pour avoir accepté mon invitation. Tout d'abord, David Wessel et Marc Leman, merci pour votre lecture de mon manuscrit et pour vos retours très instructifs. Je voudrais remercier mes examinateurs : Thierry Artières, Sergi Jordà et Steve McAdams.

Je voudrais maintenant faire un détour par Barcelone, en 2005. Pendant cette année de césure à l'Université Polytechnique de Catalogne, au milieu de cours sur la géométrie différentielle ou la topologie algébrique, Xavier Gràcia Sabaté, physicien, proposait cette unité sur la musique et les mathématiques. Ca a été pour moi la découverte d'un univers formidable dans lequel j'ai su que je voulais évoluer. Merci encore Xavier pour cela.

Retour en France, à Paris pour intégrer ATIAM en 2007, premiers pas à l'Ircam. Je voudrais te remercier, Julien, pour tes cours de *Groove*, de bon esprit au sein de cet institut dédié à la musique contemporaine ; Marc, pour tes conseils en Hip-Hop, et plus tard pour notre belle collaboration autour de *bubulles*. Romain pour envoyer du steak quand on en a besoin. Gaetan pour ton humour... Gonçal pour m'avoir fait cette transition douce entre la vie catalane et la vie parisienne. Lise pour ne pas avoir été la seule doctorante Ircam de notre promotion ; Tifanie, Sophie, Emilien, Maxime, Sarah et Jean-Yves, vous aussi.

Je voudrais ensuite remercier Norbert pour avoir proposé un stage passionnant et pour m'avoir pris pour le faire. Merci beaucoup pour la confiance que tu m'as accordée en me le proposant. Je t'avoue avoir été un peu perdu au début, je n'y comprenais pas grand chose et je hochais la tête calmement. J'ai beaucoup appris pendant ce stage, sur des domaines très variés. Tu m'as donné l'envie de savoir et comprendre. Et c'est avec beaucoup d'enthousiasme que je voulais continuer en thèse. C'est là que je voudrais te remercier Carlos pour avoir supporté ma candidature en thèse, pour avoir bien voulu être mon directeur de thèse même si le sujet était un peu éloigné du tien. Tu as aussi su me recarder à certains moments, notamment quand tu me disais qu'il fallait arrêter d'être un doctorant "romantique". Merci pour ça. Enfin, les évolutions d'un sujet de thèse sont non-linéaires et le mien l'a été fortement. Je voudrais ainsi te remercier Frédéric pour m'avoir appris à transformer ces non-linéarités en points forts. Je voudrais te remercier de tes conseils, de ton soutien et de ton amitié. Tu as été pour moi une grande influence.

Je voudrais maintenant remercier mon équipe d'accueil, IMTR. Merci à Sarah pour beaucoup, et j'y reviendrai. Merci Tommaso pour les moments cigarettes, bières, expériences dans le studio 4, béchamel à Barcelone... Merci Fivos pour avoir fait ces belles apparitions tout au long de ma thèse ! (Mais je ne parlerai pas de Barcelone). Merci Diemo pour tes patchs minutieux qui sonnent si bien. Merci Nicolas pour toutes ces discussions durant ma thèse, pour ces brunchs aux scones, pour ces workshops qu'on devrait continuer. Merci Julien pour ton appart, tes concerts. Merci Bruno pour les happy nouilles, les moments Tunnel, qu'ils se poursuivent à Londres ! Arnaud, Arshia, merci pour vos explications sur la géométrie de l'information (même si je suis pas encore certains d'avoir compris). Merci Jules d'avoir accepté le stage et continué en thèse avec tant d'enthousiasme. Merci pour les moments partagés, Ianis, Riccardo, Eric, Emmanuel, Fabrice, Alain.

Je remercie bien évidemment mes autres collègues de l'Ircam, notamment Patrick, Olivier,

Nicolas de l'équipe Perception Design Sonore, pour le beau projet sur le geste et la perception sonore ; Sylvie, Carole, Martine pour leur soutien administratif et logistique sans faille ; Hugues pour son grand soutien au début de ma thèse ; Pauline pour les jjcaas. Et j'en oublie, désolé.

Je voudrais remercier tous mes amis qui ont partagé de près (voire de très près) ou de loin, mes années de thèse. Je vous remercie sincèrement. Nico, Chloe, Lolo, Barquette, Smithou, Anaelle, Seb, Sisi, un belle bande d'Anecdote ! Caro merci pour ces instants Café et merci par avance pour l'édition de mon manuscrit ! Orsi, Sophie, Tanguy, Moul, Caroline, Philippe, Jenni, Clément, Yann, Romain, Anne-So, Grégory, Ariadna, mes colocs tous préférés. Certains d'entre vous ont vu ce manuscrit de très près, et c'était le deuxième ! Benjamin, Aline, KéKé, Elodie, Fanny, Antonin, Alou, Antoine, mes colocataires par adoption ! Merci pour votre soutien surtout pour faire la fête après ! Antoine, Marie, Sawyer, Antoine, merci pour votre amitié durable ! Charles, Darrell, Marion, Fred, merci pour les Perières et les BDs. Clément, Grégoire, merci pour beaucoup, nous pouvons monter notre boîte maintenant ! Dude, c'est quand qu'on travaille ensemble ? Et Grenoble : Pauline, Papado, Maité, P'tit Nico, Ced, Elise, Martin, Galou, (j'en oublie, je suis désolé).

Niko, merci pour ces moments de musique pendant le début de ma thèse ça m'a permis de passer de la théorie à la pratique. T'as toujours été une influence.

Pour finir je voudrais remercier ma famille à qui je dédie cette thèse : ma mère, Michèle, une femme très courageuse et formidable, ma soeur, Maud, une soeur extraordinaire et mon beau-père, Richard. Mon dernier mot sera pour Sarah, qui me connaît mieux que personne, qui a suivi cette thèse depuis le début et qui y a contribué, pour mon plus grand bonheur.

Baptiste Caramiaux
Londres, Septembre 2012

À Michèle, Maud, et Richard.

Table des matières

Chapitre 1

Introduction

« L'organisme, justement, ne peut-être comparé à un clavier sur lequel joueraient les stimuli extérieurs et où ils dessineraient leur forme propre pour cette simple raison qu'il contribue à la constituer [...] Ce serait un clavier qui se meut lui-même, de manière à offrir – et selon des rythmes variables – telles ou telles de ses notes à l'action en elle-même monotone d'un marteau extérieur »

– Merleau-Ponty

1.1 Instrument conceptuel

Imaginons la situation d'un auditeur écoutant un morceau de musique. Il fait l'expérience sensible de phénomènes sonores qui suscitent en lui des sensations, des émotions, des réminiscences, des souvenirs, etc. Durant l'écoute, l'auditeur commence à mettre en mouvement ses bras *sur* la musique tel un chef d'orchestre, mais sans mimer pour autant les gestes de direction de l'orchestre. Les siens sont personnels, non codifiés. Ses intentions sont diverses et peuvent notamment être de communiquer son expérience sensible à l'environnement qui l'entoure. Imaginons qu'au fur et à mesure du temps, de l'écoute et des mouvements effectués durant cette écoute, la musique devienne *jouée* par les mouvements accomplis. Ces mouvements prennent le contrôle de la musique mais la musique ne s'en trouve pas pour autant simplifiée ou appauvrie. Elle devient vivante. Ce terme peut faire polémique mais entendons-le dans le sens d'une musique enregistrée qui ne serait plus passive mais liée aux mouvements de l'auditeur, lui conférant une nouvelle interprétation. Ces mouvements sont eux-mêmes toujours influencés par l'expérience sensible de l'écoute à laquelle s'ajoute la conscience d'un contrôle qu'on pourrait qualifier d'*instrumental*. Ainsi, avec les mots de Merleau-Ponty, les mouvements contribuent à constituer la forme propre de la musique et réciproquement.

Cette thèse envisage la situation précédemment observée comme un instrument de musique conceptuel. Notre objectif est l'étude des différentes notions le composant de même que la proposition de modèles pour sa réalisation. Ce manuscrit apporte des éléments de réponse à une problématique vaste, largement pluridisciplinaire et aux premiers abords « mal définie ». La première question que nous nous sommes posée est la suivante : quels sont les éléments mis en jeu dans la conception d'un tel instrument de musique auxquels peut répondre un travail de recherche en informatique ? Nous y avons répondu par l'analyse suivante.

La situation introduite commence par *l'écoute*. Ce terme est utilisé ici de manière générique et sera précisé au fil du discours. Dans l'immédiat, l'écoute est entendue comme le fait de percevoir un phénomène sonore et de l'assimiler, c'est à dire de se concentrer sur ce qui est perçu. Plus précisément ce qui apparaît comme primordial pour notre *instrument* est de comprendre les différents aspects du passage de la musique en tant que phénomène sonore à un geste en tant que phénomène physique. D'après les avancées modernes en sciences cognitives, la perception, qu'elle soit visuelle ou auditive, est sensori-motrice. Elle est liée à l'action et donc au

corps. La musique a de plus des caractéristiques qui suscitent des réactions motrices que ne provoquent pas nécessairement les stimuli visuels. Par exemple, une pièce musicale avec un rythme clair induit un mouvement périodique en phase avec le rythme perçu. L'*instrument* proposé nécessitera cependant des caractéristiques plus vastes que la métrique et le rythme. En cela, il conviendra de séparer structure temporelle (amenant au rythme par exemple) et matière sonore (qualités acoustiques). Il sera alors primordial d'interroger le lien entre ces qualités acoustiques, liées à la fois à la sémantique et l'esthétique, et une représentation gestuelle.

Ensuite, la situation introduite induit le passage d'une musique enregistrée qui n'est plus passive mais qui devient contrôlée par les mouvements de l'auditeur. Dans le contexte des instruments de musique numériques, le problème est de lier une représentation numérique du geste avec une représentation numérique de la musique. Les stratégies de liaison sont informées par les études sur la perception. Cependant, l'information véhiculée par le geste lorsqu'il est effectué à l'écoute de la musique et celle contenue dans sa représentation numérique diffèrent. Les technologies utilisées sont un enjeu majeur, mais aussi les méthodes d'extraction de l'information à partir des données échantillonnées. Le discours bascule de la performance musicale à la réunion de l'informatique, le traitement du signal et l'apprentissage automatique. Avec l'*instrument* envisagé, le musicien doit pouvoir s'exprimer au travers de modèles computationnels. Pour cela les modèles doivent prendre en compte un grand nombre de caractéristiques, gages de l'expressivité, telles que la structure temporelle, la continuité du mouvement humain ou encore certaines des contraintes biomécaniques.

1.2 Organisation

Le manuscrit présenté suit la structure de conception de l'*instrument* imaginé dans la section précédente. Notre but a été d'explorer cette thématique largement à découvrir, où peu de travaux existent et où les implications scientifiques et artistiques sont nombreuses. Pour cela nous proposons de diviser le manuscrit en quatre parties regroupant nos contributions scientifiques.

La partie I présente le contexte général de la thèse introduisant (chapitre 2) des éléments fondamentaux pour nos travaux de recherche, à savoir, la conception d'instruments de musique numériques, la notion de geste et le lien entre perception et contrôle moteur. Nous présentons ensuite une première contribution (chapitre 3) qui regroupe plusieurs notions fondamentales : réponses gestuelles à des stimuli sonores et analyse computationnelle du lien entre geste et son. Cette première étude, fondamentalement exploratoire, forme la base des travaux qui suivront : elle fait naître les questions de recherche auxquelles nous apportons des réponses dans ce manuscrit.

La partie II porte sur la perception du son en tant que stimulus à une réponse gestuelle. La première contribution de cette partie (chapitre 5) étudie les différentes stratégies de réponses gestuelles suivant un corpus de sons environnementaux liés ou non à une action humaine. Cette étude propose un protocole expérimental validant à la fois le stimulus utilisé et l'hypothèse assumée. Elle est le fruit d'une collaboration avec l'équipe Perception et Design Sonore de l'Ircam. La deuxième contribution liée aux réponses gestuelles (chapitre 6) revient sur la méthode utilisée dans l'étude exploratoire servant de base à notre propos. Cette méthode est utilisée sur un corpus de sons de synthèse contrôlés, c'est à dire synthétisés suivant l'évolution prédéfinie de descripteurs sonores. L'évolution des descripteurs sonores sera appelée *tracé sonore*. Cette étude est le fruit d'une collaboration avec l'Université d'Olso et l'Université de Chicago.

La partie III porte sur la modélisation des structures temporelles du geste. La première contribution de cette partie (chapitre 8) présente un modèle pour la reconnaissance adaptative du geste basée sur la donnée d'un exemple par référence. Le but est la modélisation de la structure temporelle fine des caractéristiques du geste : la précision temporelle est à l'échantillon.

La méthode est adaptative dans le sens où les caractéristiques sont estimées dynamiquement pendant que le geste est exécuté. Nous verrons aussi que la méthode d'inférence peut être adaptée pour une caractérisation plus haut niveau telle que la reconnaissance des qualités de mouvements (ce point sera discuté dans le chapitre 10). La deuxième contribution liée à la modélisation (chapitre 9) concerne l'utilisation d'un modèle pour la segmentation et l'analyse syntaxique du geste de l'instrumentiste. Aussi, cette analyse profite d'une représentation symbolique du geste qui permet de caractériser et de repérer les motifs récurrents dans le geste lorsqu'ils existent. Cette dernière étude est le fruit d'une collaboration avec l'Université McGill à Montréal.

La partie IV regroupe les applications des travaux présentés dans les parties précédentes et les conclusions. Nous reportons deux applications qui illustrent les travaux théoriques (chapitre 10). La première application est un système de sélection de sons issus d'une base de données par requête gestuelle. Une démonstration a été publiée et est reportée dans ce manuscrit. Une deuxième application est la synthèse morphologique de sons enregistrés contrôlée par le geste. La synthèse est basée sur un modèle d'alignement temporel du signal sonore sur le signal gestuel, l'alignement se faisant en temps réel. En outre nous présenterons un travail annexe portant sur l'extraction de qualités de mouvement dans le contexte d'une installation interactive. Ces applications seront suivies des conclusions (chapitre 11). Nous reporterons les résultats généraux portant à la fois sur les études expérimentales et la modélisation. Ensuite nous discuterons la liaison entre modèles et stratégies gestuelles. Enfin nous donnerons une série de perspectives envisagées.

1.3 Contributions

Journaux scientifiques, Chapitres d'ouvrage

B. Caramiaux, N. Montecchio, F. Bevilacqua. "Realtime Adaptive Recognition of Continuous Gestures". *Soumis au moment de la publication de la thèse (le titre peut changer).*

B. Caramiaux, P. Susini, O. Houix, F. Bevilacqua. "Study of the impact of sound causality on gesture responses". *Soumis au moment de la publication de la thèse (le titre peut changer).*

B. Caramiaux, M.M. Wanderley, F. Bevilacqua. "Segmenting and parsing instrumentalists' gestures". *Journal of New Music Research*, 41(1), pp.13-29. 2012

B. Caramiaux, F. Bevilacqua, N. Schnell. "Towards Cross-Modal Gesture–Sound Analysis". *In Embodied Communication and Human-Computer Interaction, volume 5934 of Lecture Notes in Computer Science*, 158–170. 2010

Conférences internationales à comité de lecture

J. Françoise, B. Caramiaux, F. Bevilacqua (2012) "A Hierarchical Approach for the Design of Gesture-to-Sound Mapping". *Sound and Music Computing (SMC 2012), Copenhagen, Danemark*

S. Fdili Alaoui, B. Caramiaux, M. Serrano, F. Bevilacqua. "Dance movement qualities as interaction modality". *ACM Designing Interactive Systems (DIS 2012), Newcastle, UK*
◇ Award : honorable mention

K. Nymoen, B. Caramiaux, M. Kozack, J. Tørresen. "Analyzing Sound Tracings – A Multimodal Approach to Music Information Retrieval". *In Proceedings of ACM Multimedia – MIRUM 2011*, Arizona, USA. November 2011.

B. Caramiaux, S. Fdili Alaoui, T. Bouchara, G. Parseihian, M. Rébillat. "Gestural Auditory and Visual Interactive Platform". *14th International Conference on Digital Audio Effects (DAFx-11)*, Paris, France. September 2011.

B. Caramiaux and P. Susini and T. Bianco and F. Bevilacqua and O. Houix and N. Schnell and N. Misdariis "Gestural Embodiment of Environmental Sounds : an Experimental Study". *In Proceedings of New Interfaces for Musical Expression (NIME2011)*, Oslo, Norway. June 2011.

B. Caramiaux, F. Bevilacqua, N. Schnell. "Sound Selection by Gesture". *In Proceedings of New Interfaces for Musical Expression (NIME2011)*, Oslo, Norway. June 2011.

S. Fdili Alaoui, B. Caramiaux, M. Serrano. "From dance to touch : movement qualities for interaction design". *Proceedings of ACM CHI, Extended Abstract*, Vancouver, BC, Canada. May 2011.

B. Caramiaux, F. Bevilacqua, N. Schnell. "Analyzing Gesture and Sound Similarities with a HMM-Based Divergence Measure". *In Proceedings of the Sound and Music Conference*, Barcelona, Spain. July 2010.

B. Caramiaux, F. Bevilacqua, N. Schnell. "Study on Gesture-Sound Similarity". *In Proceedings of the 3rd Music and Gesture Conference*, Montreal, Canada. March 2010.

B. Caramiaux, N. Schnell. "Towards an Analysis Tool for Gesture/Sound Mapping". *8th Gesture Workshop*, Bielefeld, Germany. February 2009.

Développement logiciel

Librairies C++

- `GfPf Lib` : filtrage particulaire pour la reconnaissance adaptative de geste en temps réel (chapitre 8).
- `DySyId Lib` : identification de systèmes dynamiques en temps réel pour le mouvement dansé

Modules pour l'environnement de programmation Max/MSP en utilisant la librairie Ftm&Co et MnM :

- `mnm.cca` : Canonical Correlation Analysis (chapitre 3 et annexe C)
- `mnm.dtw` : Dynamic Time Warping
- `mnm.lds2ndord` : Identification de systèmes dynamiques, basée sur `DySyId Lib` (cas particuliers du $2^{\text{ème}}$ ordre)

Module pour l'environnement de programmation Max/MSP indépendant des librairies Ftm&Co et MnM :

- `gfpf` : module de reconnaissance et suivi de geste basé sur `GfPf Lib` (cf. chapitre 8)

En complément du manuscrit, nous avons mis à disposition du lecteur intéressé des ressources supplémentaires à l'adresse suivante : http://baptistecaramiaux.com/blog/?page_id=14.

Première partie

Contexte et Problématiques

Chapitre 2

Contexte

L'*instrument* conceptuel présenté dans l'introduction, pris comme sujet de recherche, induit un contexte pluridisciplinaire et fascinant. Le point de départ est le phénomène sonore et les actions qu'il suscite. Nous replaçons dans un premier temps cette dépendance dans un contexte plus global (et pas forcément musical) qui est le lien Action – Perception. Dans le domaine musical, cela nous amène à la notion de *geste musical*, qui est le deuxième thème abordé. Enfin, ces notions se retrouvent dans un contexte d'application qui est le nôtre, à savoir les instruments de musique numériques.

2.1 Action–Perception

Percevoir c'est agir. L'homme perçoit l'environnement dans lequel il est par l'*expérience* de cet environnement. Ainsi agir (à entendre comme un mouvement, un déplacement intentionnel) c'est permettre à nos sens de nous renseigner sur le monde extérieur, l'action conduisant à la perception. Par exemple, lorsque nous sommes immobiles et que nous suivons du regard quelqu'un ou quelque chose qui se meut, l'action de bouger la tête permet de percevoir la cible en mouvement, celle-ci nous renseignant sur sa trajectoire qui conduira le mouvement de la tête. De même, notre perception du monde extérieur va guider nos actions, par exemple pour attraper un objet. Ainsi, l'Action–Perception est cette interdépendance entre l'environnement, le corps et l'esprit, sur lequel s'ajoute notre construction propre liée à la culture, l'histoire, le lieu, etc. Alva Noë écrit au début de son livre *Action in Perception* (Noë, 2005) :

> *The world makes itself available to the perceiver through physical movement and interaction* – Alva Noë (Chapitre I, p.1)

2.1.1 Théorie de l'*enaction*

Les écrits sur le rôle de l'action dans la perception font souvent référence, en préambule, aux travaux précurseurs des phénoménologues et notamment Husserl, Heidegger, Poincaré ou Merleau-Ponty. La phénoménologie est un courant de la philosophie qui s'attache à étudier le comportement (humain) et son lien avec la conscience (ses contenus). Parmi les philosophes cités, Merleau-Ponty s'est tout particulièrement intéressé à la phénoménologie appliquée à la perception. L'ambition était déjà de dépasser deux positions radicalement opposées qui correspondaient au

- subjectivisme : le monde n'existe qu'à travers notre esprit, le monde est une projection de notre conscience.
- objectivisme : le monde existe indépendamment de notre être, notre esprit forme une représentation de ce monde . Ici représentation réfère à une *reconstruction* symbolique du monde.

Ainsi, dans son ouvrage *La phénoménolgoie de la perception* (Merleau-Ponty, 1945), Merleau-Ponty ecrit

> Maintenant que j'ai dans la perception la chose même et non pas une représentation, j'ajouterai seulement que la chose est au bout de mon regard et, en général, de mon exploration. (Merleau-Ponty (Merleau-Ponty, 1968))

Peu à peu, les sciences cognitives, comme les phénoménologues, ont voulu dépasser ce dualisme afin de palier aux limitations des méthodes classiques dans l'étude de l'esprit et notamment un certain échec de l'objectivisme en intelligence artificielle (IA). En IA, les méthodes cognitivistes prônaient une cognition *computationnelle*, c'est à dire effectuant des calculs sur des symboles (i.e. les représentations), et ne pouvaient pas expliquer un grand nombre de phénomènes perceptifs comme certaines illusions perceptives visuelles (Noë, 2005). C'est dans les années 80 et 90 que les chercheurs en sciences cognitives et en particulier les travaux de Varela et de ses collaborateurs ont montré que l'étude de l'expérience avait une place dans la science définissant ainsi un cadre à la fois épistémologique, physiologique, neuroscientifique pour l'étude du lien entre le corps et l'esprit dans la perception et en particulier par les travaux de Varela et ses collaborateurs.

Ces derniers (Varela et al., 1991) se sont basés sur les travaux des phénoménologues afin de montrer que la *cognition* n'est ni reconstruction du monde environnant (objectivisme ou réalisme) ni projection de notre connaissance du monde (subjectivisme ou idéalisme) mais *action incarnée*. Dans l'ouvrage *L'inscription corporelle de l'esprit* (Varela et al., 1991), les auteurs expliquent l'utilisation du terme *incarnée* de la manière suivante :

> Par le mot *incarné*, nous voulons souligner deux points : tout d'abord, la cognition dépend des types d'expériences qui découlent du fait d'avoir un corps doté de diverses capacités sensori-motrices ; en second lieu, ces capacités individuelles sensori-motrices s'inscrivent elles-mêmes dans un contexte biologique, psychologique et culturel plus large. ((Varela et al., 1991), Chapitre 8, p.234)

Les propriétés sensori-motrices de l'homme, notre perception et l'action ne peuvent pas être dissociés et même évoluent ensemble. Les auteurs nomment cette approche de la cognition *enaction*, c'est à dire que notre perception consiste en une action guidée par la perception ; et les structures cognitives émergent d'un ensemble organisé d'actions (ou *schèmes*) sensori-motrices récurrentes qui permettent à l'action d'être guidée par la perception (Varela et al., 1991). En d'autres termes, l'organisme et l'esprit s'organisent mutuellement pour interagir avec l'environnement, c'est un monde et un sujet percevant qui se déterminent l'un l'autre. O'Regan (O'Regan, 1992) propose ainsi une interprétation *enactive* de la perception visuelle : le monde est comme une mémoire extérieure (aussi postulée par Turvey (Turvey, 1977)). Il n'est pas utile d'avoir une représentation (du monde extérieur) avec sa géométrie et sa métrique car il nous suffit d'explorer par des mouvements (des yeux, des mains, etc..) afin de percevoir. La théorie de l'action–perception a été abondamment étudiée dans la perception visuelle (on réfère le lecteur à l'ouvrage de Noë qui ne compte pas les avancées récentes mais renvoie à beaucoup de travaux séminaux). Dans le domaine auditif, Liberman et al. (Liberman and Mattingly, 1985; Liberman and Mattingly, 1989) proposent une théorie motrice de la perception du langage en montrant les limites des approches "classiques" basées essentiellement sur l'audition.

2.1.2 Aspects en neurosciences

Parallèlement, certains auteurs comme Berthoz (Berthoz, Alain, 1997) ou Jeannerod (Jeannerod, 1997) ont développé une théorie motrice de la perception sous un angle physiologique, c'est à dire d'un point de vue physique, mécanique et biochimique. Ainsi dans *Le Sens du Mouvement* (Berthoz, Alain, 1997) l'auteur précise l'idée selon laquelle la perception est une action **simulée** (introduit aussi par Viviani (Viviani, 1990) et supportée par Jeannerod dans (Jeannerod, 1997)). L'auteur écrit :

> Pour moi, le cerveau est un simulateur au sens du « simulateur de vol » et non celui de la « simulation sur ordinateur ». Il signifie que c'est l'ensemble de l'action qui est joué dans le cerveau par des modèles internes de la réalité physique qui ne sont pas des opérateurs mathématiques mais de vrais neurones dont les propriétés de forme, de résistance, d'oscillation, d'amplification, font partie du monde physique, sont accordées au monde extérieur. ((Berthoz, Alain, 1997), chapitre 1, p28)

Ainsi l'auteur se détache des cognitivistes et introduit la physiologie de la perception qui se détermine par le monde extérieur et qui le détermine, rejoignant l'approche conceptuelle de Varela et ses collègues. Tout au long de l'ouvrage, l'auteur reprend les avancées en neurosciences, biologie, et sciences cognitives. De cette manière il propose de redéfinir les cinq sens communs en se basant sur la perception. Il soulève le problème de la cohérence des sens, de la mémoire et du lien entre perception et émotion (qui reste encore largement non expliqué).

Ces recherches trouvent un écho avec les travaux parallèles sur le développement de modèles *computationnels* pour le contrôle moteur, et la perception. Nous reviendrons sur ces modèles dans la section 2.1.3.

La théorie de la perception comme action simulée a été supportée par des travaux en neuroimagerie, notamment la découverte des neurones miroirs par Gallese et ses collaborateurs (Gallese et al., 1996). Dans cette expérience, les auteurs ont trouvé un ensemble de neurones dans le cerveau du chimpanzé qui s'activent lorsque l'animal effectue une action avec un but (e.g casser une cacahuète) et également lorsqu'il observe cette même action effectuée par l'expérimentateur. Dans (Kohler et al., 2002), les auteurs vont plus loin et montrent que les neurones miroirs s'activent lorsqu'on observe ou fait l'action, mais aussi lorsqu'on écoute le son associé à l'action (e.g. la cacahuète cassée) sans le stimulus visuel.

2.1.3 Modèles computationnels

Parallèlement aux avancées conceptuelles et expérimentales de l'étude du comportement et sa relation avec la perception, certains auteurs se sont intéressés à des aspects de modélisation.

Un premier ensemble de modèles sont ceux basés sur le contrôle *optimal*. Dans ce cas, la question est de trouver les signaux de contrôle qui, suivant certaines dynamiques, minimiseront une fonction de coût relative à certaines contraintes (Kirk, 2004; Todorov, 2009). Ces méthodes sont bien adaptées pour la planification motrice (Bryson and Ho, 1979). Dans ce cas, les trajectoires ne sont pas explicites mais proviennent d'une optimisation. Par exemple, le minimum de secousse (ang. *minimum jerk*) (Flash and Hogan, 1985; Viviani and Flash, 1995) ou minimum de variance (Harris and Wolpert, 1998). Différentes fonctions de coût ont été étudiées (Nelson, 1983) conduisant à différentes trajectoires des paramètres cinématiques. Ces différentes fonctions de coût ont d'ailleurs étaient appliquées au geste instrumental du violoniste (Rasamimanana and Bevilacqua, 2008).

Un deuxième ensemble de modèles fait intervenir une représentation interne. Il y a deux transformations basiques dans un système de contrôle (Jordan and Wolpert, 1999) : la transformation des variables motrices en variables sensorielles et la transformation inverse des variables sensorielles en variables motrices. Dans le premier cas, une série de méthodes existe, appelée modèles *forward* pour la prédiction motrice (Wolpert and Ghahramani, 2000). Les modèles dynamiques *forward* prédisent le prochain état (dans une représentation interne) qui peut être la vitesse ou la position sachant l'état courant et la commande motrice. Ces modèles rejoignent les études par Jeannerod pour le contrôle du mouvement de l'oeil (Jeannerod, 1997). Ces méthodes peuvent aussi être utilisées pour l'apprentissage moteur (Miall and Wolpert, 1996). Dans le deuxième cas, les modèles inverses servent de contrôleur. Ces modèles transforment les variables sensorielles en contrôle moteur (Kawato et al., 1987).

Warren (Warren, 2006) note que l'approche basée sur des modèles internes, introduits précédemment, ne met pas celui qui perçoit en contact direct avec son environnement mais avec

un modèle interne. De même celui qui agit contrôle un modèle interne et non ses mouvements. L'auteur propose une approche basée sur des systèmes dynamiques dans lesquels interviennent à la fois l'environnement et le contrôle moteur. Cette approche *dynamique* a surtout eu une répercussion dans le domaine de la robotique (Schoner et al., 1995; Schaal et al., 2003). Seulement, il est souvent infaisable d'énoncer un ensemble d'équations illustrant la dynamique complète d'un système tel que le comportement humain. C'est pourquoi les méthodes basées sur des modèles à états internes ont eu plus de succès. En effet dans ce cas, des méthodes d'apprentissage efficaces existent et sont bien maîtrisées. Ceci sera discuté de manière plus détaillée dans le cas de la modélisation du geste musical 7.

2.2 Le geste musical

Dans cette section, nous revenons à un contexte musical. Nous ne parlons plus d'action mais de geste induisant le fait de véhiculer un sens, une expression. Nous présentons différents aspects du geste lié au son ou à la musique dans la littérature.

2.2.1 Préambule : détour par la parole

Le geste est souvent entendu comme lié à la parole, constituant un élément important de la communication non-verbale. Partant de la définition de l'action donnée ci-dessus, Kendon (Kendon, 2004) définit le geste par les mots suivants :

> If an action is an excursion, if it has well defined boundaries of onset and offset, and if it has features which show that the movement is not solely under the influence of gravity, then it is likely to be perceived as gestural. ((Kendon, 2004), chapitre 2, p.14)

Ainsi, Kendon précise que le geste doit être une action qui a une sémantique spatiale, et qui a un début et une fin. Le geste est un moyen que l'être humain possède pour communiquer des émotions, des pensées, des énoncés, etc. Il s'agit ici de comprendre et de classer les différents gestes que nous pouvons effectuer dans sa relation avec le phénomène acoustique qu'est la parole, et ceci à différents niveaux sémantiques. Quel que soit le domaine d'étude du geste et de sa relation avec le son, l'analyse méthodologique induit le besoin de classifier afin de définir et cibler plus clairement le sujet de l'étude. Le but de cette section est donc aussi d'inspecter les différentes taxonomies du geste et des terminologies utilisées afin de faire apparaître plusieurs aspects du geste lié à la parole.

Classification du geste lié à la parole

Nous proposons une revue des classifications du geste lorsque celui-ci est associé au langage. Tout en s'éloignant des considérations musicales, le geste associé à la parole induit deux intérêts majeurs pour notre propos. Premièrement, le processus liant les articulations motrices et la génération du langage est régie de manière cohérente au niveau cognitif. Deuxièmenent, la terminologie utilisée est pertinente pour caractériser différents types de gestes associés au son dans le cadre de nos études.

La classification reportée ici est inspirée de l'ouvrage de Kendon (Kendon, 2004). L'auteur souligne que le geste n'est devenu un sujet d'étude scientifique que tard dans le courant du XXème siècle. Nous reportons chronologiquement trois contributions majeures correspondant respectivement aux travaux de Wundt, McNeill et Kendon.

Une première classification est donnée par Wundt (Wundt, 1973). L'auteur propose la classification suivante :

- Gestes *démonstratifs* qui sont indicatifs d'objets présents, des relations spatiales ou des parties du corps.

– Gestes *descriptifs* qui peuvent se diviser en deux sous-catégories :
  • Gestes d'*imitations* : utiliser pour imiter l'action ou l'objet.
  • Gestes *connotatifs* : utilisés lorsque les caractéristiques de quelque chose sont prises pour décrire le tout. Par exemple, un homme sera décrit par un geste imitant le fait d'ôter un chapeau (exemple reporté de (Kendon, 2004)).
  • Gestes *symboliques* : les gestes se réfèrent de manière plus complexe au sujet de la description.

À partir des précédents travaux d'Ekman et al. (Ekman and Friesen, 1969), McNeill (McNeill, 2000; McNeill, 1996) base son étude sur le lien du geste avec un référent imagé.

– Gestes *imagés* : se réfèrent à une *image* de toute nature, allant de la forme d'un objet à une action ou activité.
  • Gestes *iconiques* : utilisés pour représenter un objet ou une action concrète (ces gestes sont à mettre en lien avec les gestes d'imitations de Wundt).
  • Gestes *métaphoriques* : utilisés comme moyen pour représenter une image qui se réfère à une idée abstraite (ces gestes sont à mettre en lien avec les gestes symboliques de Wundt).
– Gestes *non-imagés*
  • Gestes *déictiques* : désignation d'un objet ou quelqu'un du doigt
  • Gestes de *pulsations* : mouvements qui ne portent pas de sens mais qui marquent les segments de la paroles, son rythme (ang. *beats*).
– *Emblèmes* : signes comme bouger la main pour dire au revoir.

Dans l'étude du lien entre geste et langage, Kendon dans (Kendon, 1988), et repris dans (Kendon, 2004), porte l'attention sur le fait que le geste puisse être considéré comme couvrant plusieurs niveaux à la fois sémantiques et temporels. Il propose une étendue allant du *mot* aux représentations *pantonymiques*. Il suggère par là que le geste, dans sa relation avec la parole, peut lui être complémentaire voire s'y substituer (par exemple dans le langage des signes). En particulier, Kendon conjecture que les caractéristiques de la parole peuvent être retrouvées dans le geste.

Parallèlement, McNeill soutenait l'inséparabilité du geste et de la parole dans le processus du discours. Ceci l'a amené à définir ce qu'il appelle le *continuum de Kendon*. Ce continuum peut être considéré comme une « classification » continue du geste pris dans sa relation avec le langage parlé. Dans le domaine musical, Jensenius reprend cette observation dans son travail de thèse (Jensenius, A. R., 2007) et propose une vision schématique du continuum de Kendon que nous reprenons ici (voir figure 2.1).

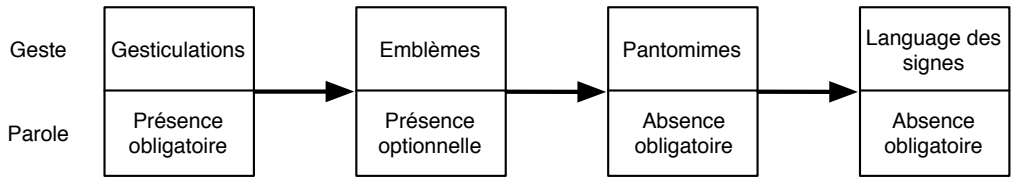

FIGURE 2.1 – Continuum de Kendon

Ainsi le geste acquiert à la fois un autre sens et prend d'autres formes parallèlement au niveau de présence de la parole. La classification prend en compte le lien fondamental entre processus moteur et processus sonore. La classification des gestes dans un contexte musical suivra la même méthodologie. Dès lors, nous pointons une différence certaine avec la musique : la parole n'est pas causée par le geste. Cette causalité geste–son dans le contexte des instruments musicaux influera sur la classification. C'est le sujet de la prochaine section.

2.2.2 Définition et taxonomie

La définition que nous avons énoncée au début de la section précédente peignait le geste comme un événement ayant un début et une fin et appartenant au domaine de la parole. Dans le cas où le geste devient signe, il transmet une sémantique claire, symbolique. Dans la pratique des instruments de musique, le geste acquiert un sens plus vaste.

Comme nous l'avons introduit précédemment, l'utilisation des technologies numériques pour l'expression musicale sépare de manière indépendante la partie *contrôle gestuel* et le *rendu sonore* portant au premier plan des questionnements sur le geste dans la musique : caractéristiques expressives, intention, etc. Cet intérêt croissant a induit une production scientifique accrue, en témoignent les travaux de thèse de doctorat liés au geste dans la musique : (Wandereley, 2001), (Volpe, 2003), (Dahl, 2006), (Jensenius, A. R., 2007), (Demoucron, 2008), (Rasamimanana, 2008), (Maestre, 2009), (Bouenard, 2009), (Schoonderwaldt, 2009), (Naveda, 2010), (Fiebrink, 2011), (Merer, 2011). À ces travaux de thèse s'ajoutent des articles ou ouvrages de référence tels que (Delalande, 1988; Cadoz, 1988; Cadoz and Wandereley, 2000; Wanderley and Depalle, 2004; Leman, 2007; Godøy, 2009). De manière générale, ces gestes sont nommés *musicaux* et se définissent, au plus large, comme des mouvements, en musique, pouvant être *physiques* ou *sonores* (ces derniers sont aussi appelés *mentaux*). Nous complétons la définition en précisant qu'à l'idée d'action s'ajoute la dimension expressive.

Dans un geste, il y a l'*expression* (aussi dans le sens d'une communication) d'une intention. Et l'expression de cette intention va susciter des émotions chez la personne qui regarde (ou qui écoute si on étend à l'expression musicale). L'expression peut se diviser entre **"ce qui est exprimé"** et **"comment cela est exprimé"** (aussi appelées expression sémantique et rhétorique (Mazzola, 2011)). Cette distinction est très importante pour la suite de nos travaux et sera reprise à la fois dans un cadre expérimental et dans la modélisation. Aussi, considérer le geste comme ayant un début et une fin est trop restrictif pour notre propos. Nous parlerons aisément du geste du musicien pour caractériser son *mouvement* tout au long de la performance musicale.

Dans le contexte des instruments de musique, différentes taxonomies du geste musical existent dans la littérature. On peut citer (dans un ordre chronologique) : (Delalande, 1988; Cadoz, 1988; Wanderley and Depalle, 2004; Jensenius et al., 2009). Nous suivrons ci-après la taxonomie la plus récente en faisant des liens avec les précédentes. Une modification sera faite par rapport à (Jensenius et al., 2009) : nous continuerons de les nommer *gestes* et non *actions* car nous pensons que ce terme est plus approprié afin de rendre compte du caractère *expressif* et non *utilitaire* (i.e. dirigé par un objectif à atteindre) du mouvement (ce qui est en accord avec les travaux précédents (Cadoz, 1988; Cadoz and Wandereley, 2000; Wanderley and Depalle, 2004)). Cette taxonomie propose de diviser les gestes musicaux en quatre sous-ensembles.

- **Gestes de production**. Ce sont les gestes qui produisent le son lors d'une performance musicale. Ils sont appelés *effecteurs* par Delalande (Delalande, 1988) ou *instrumentaux* par Cadoz (Cadoz, 1988). En outre, Cadoz propose de subdiviser cette catégorie en gestes d'*excitation* (e.g. les mouvements d'archets) ou de *modification* (e.g. l'appui des doigts sur le manche du violon).
- **Gestes ancillaires**. Ce sont les gestes qui sont auxiliaires aux gestes de production et qui ne causent pas le son. Un exemple est le mouvement du pavillon de la clarinette en situation de jeu. Delalande les nomme *gestes d'accompagnement* (Delalande, 1988) alors que Wanderley préfère *ancillaires* (Wanderley and Depalle, 2004) ou, parfois, *gestes non-évidents* (Wanderley, 2002) montrant par là qu'ils sont moins bien définis que les gestes de production mais porteurs de l'expressivité de l'interprète (cette interprétation est supportée par des travaux connexes, e.g. (Dahl and Friberg, 2003; Vines et al., 2004)). Il sont aussi parfois appelés *gestes facilitant* (Dahl et al., 2009).
- **Gestes d'accompagnement**. Ce sont les gestes qui accompagnent le son dans le sens où ce sont des gestes *en réponse* au son. Jensenius (Jensenius et al., 2009) précise qu'ils ne sont

pas liés à la pratique d'un instrument de musique mais sont les gestes qui sont associés (de manière libre) au son écouté.
– **Gestes de communication**. Ce sont les gestes effectués pour la communication notamment entre musiciens au sein d'un ensemble (par exemple lors d'une improvisation avec plusieurs instrumentistes). Ce type de geste ne sera pas détaillé dans le cadre de cette thèse.

Cette taxonomie regroupe les divers cas d'apparition du geste musical physique, mais aussi sonore (implicitement dans le cas du geste d'accompagnement, comme nous le verrons). Cette taxonomie peut alors être décrite en termes de relation entre le geste et la musique. En effet, la relation entre gestes de production et musique est causale : ces gestes produisent le rayonnement acoustique. La relation entre gestes ancillaires et la musique est complémentaire, exactement à l'image de *la gesticulation* introduite par Kendon (Kendon, 2004) (cf. figure 2.1) dans le cas de la parole. Enfin les gestes d'accompagnements sont en réponse au son : la relation de causalité est de la musique vers le geste.

2.2.3 Geste → Musique : geste de production

Ici nous considérons le geste qui produit la musique (ou plus généralement un phénomène sonore), la relation est causale : geste → musique.

Les gestes de production ont été étudiés dans le cadre du piano (Delalande, 1988; Jensenius, A. R., 2007; Loehr and Palmer, 2007), du violon (Maestre et al., 2010; Rasamimanana et al., 2006; Schoonderwaldt and Demoucron, 2009), des percussions (Bouenard et al., 2010) avec des approches différentes. Dans (Delalande, 1988; Jensenius, A. R., 2007), les auteurs adoptent un point de vue musicologique, amenant à l'édification de nouvelles taxonomies pour l'analyse. Dans le cadre du violon, (Maestre et al., 2010) proposent une modélisation par analyse / synthèse de contours des données de contrôle du violon afin de piloter la synthèse sonore. (Rasamimanana et al., 2006) se sont focalisés sur les modes de jeu du violon et les méthodes pour la reconnaissance et la modélisation (notamment en faisant le lien avec les lois motrices du mouvement (Rasamimanana and Bevilacqua, 2008)). (Schoonderwaldt and Demoucron, 2009) posent le problème de la captation du mouvement et du calcul des descripteurs. Pour les percussions, dans (Bouenard et al., 2010), les auteurs étudient la problématique de synthèse graphique du geste musical. Enfin, Loehr et al. (Loehr and Palmer, 2007) adoptent une démarche expérimentale afin de caractériser les contraintes biomécaniques et cognitives dans le jeu du piano.

2.2.4 Geste ↔ Musique : geste ancillaire

Ici nous considérons le geste qui est produit en parallèle à la musique (ou plus généralement un phénomène sonore) rendant le lien de causalité plus ambigu : geste ↔ musique.

Les gestes ancillaires ont été étudiés dans le cadre du piano (Delalande, 1988; Thompson and Luck, 2008), de la clarinette (Wanderley, 2002; Vines et al., 2004; Wanderley et al., 2005; Palmer et al., 2009), des percussions (Dahl and Friberg, 2003). Davidson et al. (Davidson, 1993) montrent que les intentions expressives des interprètes sont le plus fidèlement transmises par leurs mouvements. Cette observation est supportée par Dahl (Dahl and Friberg, 2003) dans le cas de la transmission des émotions chez les percussionnistes, de même que Vines et al. (Vines et al., 2004) dans le cas particulier des gestes ancillaires du clarinettiste. Dans cette étude, Vines et al. explorent comment les gestes expressifs d'un clarinettiste professionnel contribuent à la perception de l'information structurelle et émotionnelle dans la performance musicale. Deux études importantes sur les gestes ancillaires des clarinettistes ont été menées par Wanderley (Wanderley, 2002; Wanderley et al., 2005). L'auteur a d'abord montré que ces gestes étaient consistants pour un même joueur puis que certaines caractéristiques peuvent être partagées entre les joueurs, notamment celles liées à la partition (rythme, mesure, respirations). Un fait

marquant est le lien entre gestes ancillaires et métrique. En effet, Wanderley et al. montrent dans (Wanderley et al., 2005) que les performances des musiciens devant jouer de manière inexpressive sont plus courtes et les mouvements du pavillon quasi inexistants. D'autre part, Palmer et al. (Palmer et al., 2009) examinent le lien entre mouvements de la clarinette et expression. Les auteurs montrent notamment que les gestes ancillaires sont liés à la durée du son produit, qui s'en trouve plus grande pour les cas expressifs. Ils ajoutent que ni l'intensité ni la hauteur ne montrent une corrélation évidente avec les mouvements du pavillon.

Contrairement au cas des gestes de production, parmi les travaux sur les gestes ancillaires, aucuns (à notre connaissance) n'ont proposé des modèles pour leur étude. La recherche dans ce domaine est, à ce jour, en exploration et met en évidence plusieurs caractéristiques des gestes ancillaires que nous résumons ainsi :

- ils sont liés à l'expression d'une intention et de certaines émotions ;
- ils ne sont pas nécessaires pour le jeu mais semblent ne pas pouvoir être occultés ;
- ils sont consistants chez un même joueur ;
- ils sont partagés par plusieurs interprètes lorsqu'ils sont liés à la partition ;
- ils sont structurels et liés à la métrique plus qu'à d'autres paramètres musicaux.

La cohérence (intra-participant) des gestes ancillaires suscitent la possibilité de définir un modèle extrayant cette information. De même que la caractéristique plus générique d'être liés au temps et à la métrique, contraint le modèle à devoir prendre en compte la structure temporelle.

2.2.5 Geste ← Musique : geste d'accompagnement

Ici nous considérons le geste en réponse à la musique (ou plus généralement un phénomène sonore). Le lien de la musique vers le geste est causal : geste ← musique.

Les gestes d'accompagnement sont les gestes effectués en réaction à un *stimulus* sonore et sans instrument, ce qui nous place dans un cadre beaucoup plus large que précédemment. Dans ce contexte, des analyses qualitatives ont été menées autour du pouvoir d'expression d'une performance gestuelle telle que l'imitation d'un instrumentiste (Godøy et al., 2005) ou le pouvoir de description d'un son par un geste (Godøy et al., 2006a). Dans (Leman et al., 2009), les auteurs montrent que les mouvements effectués en réponse à une pièce musicale (traditionnelle chinoise, hauteur prédominante) sont corrélés aux mouvements du musicien. Ceci met en avant que le musicien encode ses mouvements dans le son, et ceux-ci peuvent ensuite être en partie décodés par le geste de l'auditeur. En outre, Leman et al. montrent l'importance de la vitesse dans cette relation. Les auteurs écrivent : « This suggests that the velocity of the movement is an invariant feature of embodied perception » ((Leman, 2007), chapitre 6, p.157).

En ce qui concerne les méthodes computationnelles, peu de travaux existent dans l'analyse de la relation geste d'accompagnement–son (accompagné) dans un contexte musical. Certains auteurs ont mené des études autour de la synchronisation du geste avec la musique (Large, 2000; Repp, Bruno Hermann, 2006; Styns, Frederik et al., 2007) proposant une modélisation par systèmes dynamiques de la réponse gestuelle (Large, 2000) ou encore l'analyse du geste tapant le tempo (Repp, Bruno Hermann, 2006) mettant en avant des caractéristiques importantes telles que l'asynchronie, la variabilité et la vitesse limite (repris dans (Styns, Frederik et al., 2007)). Dans (Luck and Toiviainen, 2006), les auteurs ont utilisé une mesure de corrélation pour étudier le lien entre les mouvements d'un chef d'orchestre et les caractéristiques rythmiques de la musique d'un ensemble conduit par celui-ci. Ces travaux mettent en avant l'importance des caractéristiques du mouvement liées à la synchronisation (périodes de l'accélération négative, (Repp, Bruno Hermann, 2006)) dans un contexte de direction. Dernièrement, Nymoen et al. (Nymoen et al., 2010) ont étudié le lien entre gestes d'accompagnement et sons abstraits par le biais des Machines à Support de Vecteurs. Nous reviendrons sur ces travaux publiés durant notre thèse et leur suite au travers d'une collaboration avec les auteurs. Finale-

ment, Merer (Merer, 2011) présente une étude expérimentale où les participants dessinent les mouvements inhérents au son, rejoignant l'idée d'un geste dans le son qu'on peut extraire par le mouvement.

Les gestes d'accompagnement sont ainsi liés au geste musical sonore qui ne prend forme que dans notre esprit : un mouvement mental. Ce mouvement relève de l'écoute. Métois (Metois, 1996) suggère que les gestes physiques doivent être considérés au même niveau (dans un cadre d'analyse et de modélisation) que l'information perçue par l'auditeur. Il nomme ainsi ces deux types de gestes (physiques et auditifs) : gestes musicaux. Cette approche est schématisée par la figure 2.2.

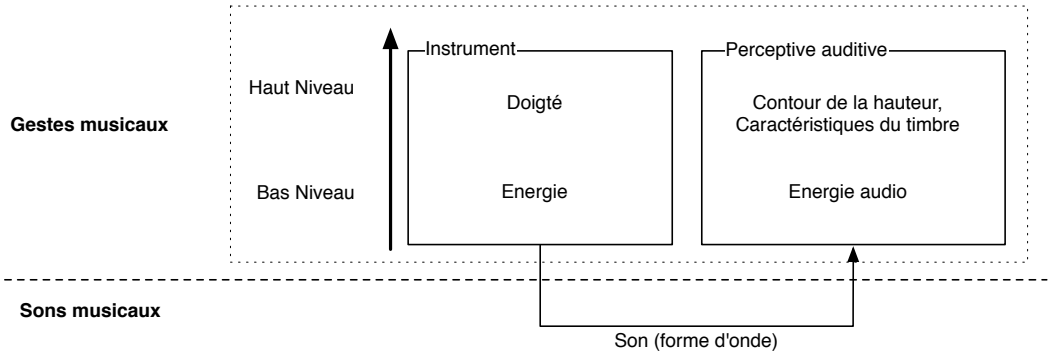

FIGURE 2.2 – Description des gestes musicaux par Métois : l'approche lie les gestes musicaux sonores et physiques en les mettant au même niveau.

Le geste musical sonore a été étudié en musicologie. Middleton (Middleton, 1993) suppose que les sons musicaux ont des formes similaires aux gestes physiques, idée partagée par Todd (Todd, 1993) qui voit dans les mouvements musicaux des mouvements dans un espace physique. Ainsi la vision de Middleton se rapproche de ce que Dobrian (Dobrian, 2004) appelle geste musical : les trajectoires des descripteurs sonores, comme la hauteur, le centroïde spectral, l'énergie audio, etc. Ceci fait écho au diagramme 2.2 proposé par Métois où les gestes musicaux auditifs se muent en contours de descripteurs sonores caractérisant le timbre. D'autre part, la vision de Todd trouve écho dans le travail de Dubnov et al. (Dubnov et al., 2006) où les mouvements musicaux (notamment les *tensions*) sont évalués par des gestes très simples des auditeurs à l'aide d'un curseur. Les auteurs ont montré la corrélation entre une mesure d'*information* [1] liée à la tension dans le musique et l'évaluation perceptive des auditeurs. Finalement, de récents travaux ont montré la pertinence de considérer des gestes mentaux liés à la perception du mouvement dans le son (i.e. le mouvement évoqué par les sons, par exemple "qui tourne") pour la l'analyse musicale (Fremiot et al., 1996).

2.3 Instruments de musique numériques

Dans le domaine de la conception d'instruments de musique numériques, il est intéressant de constater que les contributions proposées n'ont pas été dûes exclusivement aux scientifiques, mais aussi aux artistes, compositeurs et instrumentistes. Nous développerons dans cette section les contrôles gestuels mis en œuvre pour l'expression musicale. Classiquement, nous parlerons de déclenchement et de contrôle continu. Enfin nous envisagerons de nouvelles approches, dans lesquelles nous nous placerons pour la suite.

1. À entendre dans le sens de la théorie de l'information

2.3.1 Au delà d'« un geste pour un événement acoustique »

Dans l'article "*Problems and Prospects for Intimate Musical Control of Computer*" (Wessel and Wright, 2002), Wessel et al. rappellent que la plupart des instruments acoustiques traditionnels respectent le paradigme d'« un geste pour un événement acoustique » comme par exemple, le mouvement d'archet pour le violon ou l'appui d'une touche de piano. Par extension, ils mettent en avant une caractéristique fondamentale de la musique par ordinateur contrastant avec le paradigme « un geste pour un événement acoustique » : la partie *contrôle* (le geste, et les moyens de captation) est indépendante du *rendu sonore* (la synthèse). Ceci ouvre un champ infini de stratégies pour lier le geste du musicien au rendu sonore, stratégies motivées par la composition ou l'architecture de l'instrument de musique (voir aussi (Iazzetta, Fernando, 2000)).

En référence aux instruments musicaux numériques, Chadabe (Chadabe, 2002) donne un schéma conceptuel reporté par la figure 2.3, où un continuum est créé à partir du niveau d'incertitude du lien entre geste et son. À l'extrême gauche sont les systèmes totalement déterministes : ce sont les instruments traditionnels, respectant le paradigme d'« un geste pour un événement acoustique ». À droite, sont les systèmes totalement non-déterministes : l'action du musicien influera sur le son, qui à son tour, en tant que phénomène non prédictible, influera sur le musicien. Chadabe nomme ces systèmes : *interactifs*. Une inter-action est créée entre le phénomène musical et le geste du musicien. La relation entre le geste musical et le son devient complexe et rend le système expressif.

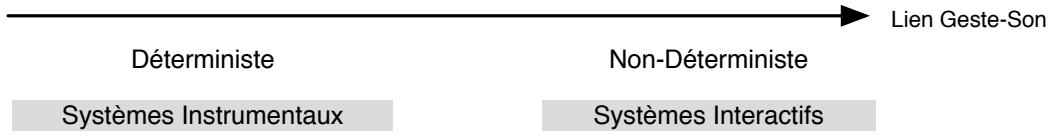

FIGURE 2.3 – Continuum pour le lien entre geste et son : de la relation déterministe à non-déterministe. Schéma conceptuel suivant l'article de Chadabe (Chadabe, 2002).

La proposition de Chadabe n'est cependant pas satisfaisante pour notre étude. Nous ne pensons pas que des systèmes interactifs non-déterministes soient gages d'une plus grande expressivité musicale. Nous pensons que si le contrôle se dissocie du rendu, il serait intéressant de considérer des systèmes **adaptatifs** c'est à dire où le rendu s'adapte au contrôle au même titre que le contrôle s'adapte au rendu.

Jordà (Jordà, 2008; Jordà, 2005) ajoute une caractéristique dans ce lien geste–musique en observant que :

> The musician performs control strategies instead of performing data, and the instrument leans towards more intricate responses to the performer stimuli, tending to surpass the note-to-note and the 'one gesture–one acoustic even' playing paradigms present in all traditional instrument, thus allowing musicians to work at different levels and forcing them to take higher level and more compositional decisions on the fly. ((Jordà, 2008), p.274)

L'auteur met en évidence deux niveaux de contrôle (appelés *micro*, *macro*). Ces niveaux sont à la fois sémantiques et à la fois temporels. En effet, le musicien devra contrôler le son à une échelle fine, ainsi que prendre en compte des mouvements plus globaux, des stratégies compositionnelles à plus long terme. Ces niveaux de contrôle devront être pris en compte dans la conception.

Ainsi, le paradigme « un geste pour un événement acoustique » est celui des instruments de musique traditionnels et peut être dépassé à bien des égards lorsqu'on considère les instruments de musique numériques : la liaison contrôle–son peut être adaptative ; le contrôle

peut porter sur plusieurs niveaux temporels à la fois. Toutefois, ce paradigme n'est pas à supprimer. Il fait partie d'une interaction possible qui peut être expressive et permettre, comme le rappellent Wessel et al. (Wessel and Wright, 2002) ainsi que Chadabe (Chadabe, 2002), une virtuosité.

2.3.2 Les paradigmes de contrôle

Le déclenchement

Une longue tradition dans le contrôle des instruments virtuels met en avant l'utilisation de la norme MIDI (*Musical Instrument Digital Interface*). Beaucoup de travaux sur les avantages et inconvénients de la norme MIDI existent dans la littérature et il ne s'agit pas ici d'en faire une revue exhaustive. Le lecteur intéressé peut se référer aux articles de McMillen (McMillen, 1994), Wessel (Wessel and Wright, 2002) ou encore Rasamimanana (Rasamimanana, 2012). Il y a cependant une caractéristique de la norme MIDI que nous aimerions discuter, à savoir qu'elle a amené à un contrôle du son, majoritairement, discret. La norme MIDI prévoit, pour autant, un grand nombre de contrôles continus, mais son utilisation a plutôt servi à des interactions basées sur le déclenchement. Un musicien peut ainsi, à partir d'un contrôleur MIDI (l'exemple classique est le clavier) contrôler le déclenchement d'un son enregistré où synthétisé par l'appui d'une des touches du clavier. Chaque touche peut être assignée à un son différent ou encore sur le même son mais où un effet serait modulé par le code MIDI de la touche (par exemple la transposition de hauteur). Ce type de contrôle s'est décliné au fil des années à travers différents types d'interface (pads, surfaces, boutons, etc), et dans les jeux vidéos grand public. Son avantage fondamental est de pouvoir être très naturellement pris en main. En d'autres termes, la relation entre le geste effectué et le son produit est très vite apprise par le musicien (ou utilisateur). Il est le contrôle utilisé à la fois dans une partie de la musique électronique populaire moderne et dans des technologies grand public.

Le contrôle continu

Sur de nombreux instruments de musique, un geste continu est exercé afin de créer le son : c'est le cas des instruments à son entretenu comme le violon. De la même manière, le contrôle de son de synthèse par un geste continu offre une expressivité accrue. La différence d'information transmise entre une commande de déclenchement et un contrôle continu est grande. Wessel et al. (Wessel and Wright, 2002) et Jordà (Jordà, 2008) parlent de bande passante pour le contrôle humain des instruments virtuels.

Un exemple célèbre de contrôle continu où la relation entre geste et son est directe est le Theremin inventé dans les années 1920. Cet instrument permet le contrôle continu de deux paramètres musicaux : la hauteur et l'intensité du son. Dans le cas du Theremin, les paramètres de position du musicien sont associés aux paramètres de la synthèse donnés par l'intensité et la hauteur. On parle d'association directe ou *mapping* direct. Dans le contexte des instruments de musique numériques, le *mapping* est la fonction liant les paramètres de contrôle aux paramètres de la synthèse sonore. Le *mapping* direct a très vite montré ses limitations pour l'expression musicale.

La littérature sur le *mapping* dans les nouvelles interfaces pour l'expression musicale (ou NIME) se concentre principalement autour des années 2000. Dans cette littérature, une classification habituelle pour l'analyse du *mapping* se divise en deux catégories de stratégies (pour plus de détails, nous référons le lecteur aux articles de Rovan et al. (Rovan et al., 1997) et Hunt et al. (Hunt and Wanderley, 2002)).

- **Explicite** (Bowler et al., 1990; Wanderley et al., 1998; Goudeseune, 2002; Van Nort et al., 2004). Le *mapping* est donné par une fonction explicite entre les paramètres entrant et les paramètres sortant.

– **Implicite** (Lee and Wessel, 1992; Modler, 2000). Le *mapping* est défini par des algorithmes génératifs.

Le *mapping* a été étudié de manière théorique (en tant que fonction mathématique entre deux espaces) ou encore musicologique (en tant qu'outil pour l'expression musicale), toujours dans l'optique du contrôle continu de la synthèse sonore. Ainsi a-t-il été lié à plusieurs types de synthèse. Seulement, très vite plusieurs limitations sont apparues. Par exemple, Chadabe (Chadabe, 2002) estime que le *mapping* perd de son sens dans le cadre des instruments interactifs tels qu'il les entend. Un autre point fondamental pour notre discours est le caractère subjectif du *mapping* qui relève bien plus de la composition que d'une stratégie méthodologique et scientifique. Enfin, le *mapping* ne prend pas en compte la structure temporelle des signaux mis en correspondance.

2.3.3 Prendre en compte la structure temporelle

Dans une motivation de contrôle continu de la synthèse sonore, la limitation majeure du *mapping*, à notre sens, est de ne pas prendre en compte la structure temporelle des modalités mises en jeu, à savoir le geste et le son. Cette idée sera omniprésente dans les choix de modélisation présentés dans la partie III.

Parmi les travaux de la littérature, nous lions ceux de Godøy (Godøy, 2006) au point que nous avons énoncé ci-dessus. En effet, l'auteur met en avant que le son et le geste (dans le cas particulier de sons abstraits) sont liés par une morphologie commune, c'est à dire une évolution temporelle commune de certains de leurs paramètres. Une morphologie induit une forme ou géométrie. Cette idée a été ensuite reprise par Van Nort (Van Nort, 2009) dans le cas qui nous intéresse de la conception des instruments de musique numériques. Son idée est de penser la sortie de l'instrument comme ce qu'idéalement il peut produire puis que cet idéal sonore contraigne le contrôle : le geste. Notons que ceci recoupe un aspect de l'*instrument* introduit dans le premier chapitre du manuscrit. L'auteur spécifie sa vision par le lien entre dynamiques du son et du geste. Par là il émet une critique similaire sur l'utilisation du *mapping* de paramètres.

2.4 Synthèse

On reprend ici l'*instrument* présenté dans l'introduction de notre thèse. Globalement l'instrument est une illustration d'une application concrète de l'*enaction* à la musique. En effet, une exploration par le mouvement permet la "prise en main" de la musique enregistrée qui va, par conséquent, influer en retour le mouvement. Il n'y a donc ni position subjectiviste de la musique, car le matériau sonore appartient à l'environnement, ni position objectiviste car il n'y a pas de représentation formelle et symbolique de la musique.

Le geste musical engagé dans l'*instrument* est un geste d'accompagnement, même si certaines des caractéristiques de la relation entre geste et musique peuvent être partagées par les gestes ancillaires. Dans la littérature, les gestes d'accompagnement et ancillaires n'ont pas reçu la même attention que les gestes de production. Les gestes d'accompagnement (et ancillaires) peuvent décoder des informations essentielles sur la perception de la musique et ces informations peuvent être ensuite utilisées pour la conception d'instruments expressifs (Leman, 2007). Par là, nous suivons la voie indiquée par les travaux précurseurs de Godøy (Godøy, 2006), Leman (Leman, 2007; Leman et al., 2009), et Van Nort (Van Nort, 2009), ces approches étant appuyées par des résultats fondamentaux en sciences cognitives et neurosciences (Zatorre et al., 2007).

Afin de concevoir un instrument de musique numérique, il faut de plus proposer une méthode computationnelle (ou modèle) permettant d'analyser le geste et contrôler le rendu sonore. Comme nous l'avons vu, il s'agira de dépasser le *mapping* classique en prenant en compte

la structure temporelle des signaux gestuels à différentes échelles (Godøy et al., 2010; Jordà, 2008). Notre approche du mapping sera de proposer des modèles pour l'interaction geste–son. Ces modèles induisent certaines possibilités de contrôle. Nous pensons que ces possibilités de contrôle doivent être inspirées par les études sur les gestes d'accompagnement et ancillaires, intrinsèquement liées à la perception du son.

La volonté d'étudier le lien entre les gestes d'accompagnement et le son, par des méthodes quantitatives, nous a amené à imaginer une expérience exploratoire engageant des gestes de participants lors de l'écoute d'un ensemble de sons. C'est le sujet du prochain chapitre.

Chapitre 3

Expériences préliminaires et questions de recherche

« If we knew what it was we were doing, it would not be called research, would it ? »
– Albert Einstein

3.1 Introduction

Une collecte de données a été menée en 2008 lors d'une mission scientifique à l'Université de Graz effectuée par Norbert Schnell (nous incitons le lecteur à lire la compilation de ces missions effecutées dans le cadre du projet COST-SID (Rocchesso, 2011)). Un ensemble de participants était invité à participer à l'expérience. Ils devaient effectuer des gestes en réponse à des sons ou des extraits musicaux. Les données collectées correspondaient aux mouvements de la main des participants, aux vidéos des performances et aux fichiers audio. Dans la lignée des travaux de Leman (Leman et al., 2009) et Godøy (Godøy, 2006) et de l'enaction (Varela et al., 1991), notre but était d'étudier les liens entre le geste musical physique et des gestes perçus dans le son à l'aide d'une méthode computationnelle et de manière non-supervisée. Ces gestes sonores pouvaient être l'*empreinte* du geste de production du musicien lors d'une performance musicale (Leman et al., 2009) ou une morphologie définie par les contours acoustiques du son (Godøy, 2006).

Les problématiques posées sont les suivantes : quels gestes les participants associent-ils aux sons écoutés dans un but de contrôle ? Sont-ils consistants dans leurs performances et par rapport à celles des autres participants ? Comment représenter à la fois le geste et le son et quelle méthode computationnelle utiliser pour retrouver automatiquement l'information sur cette association ?

Notre démarche a été de proposer une analyse quantitative non-supervisée basée sur une méthode de fouille de données (de l'anglais *data mining*). À partir des ensembles de descripteurs à la fois gestuels et sonores, la méthode permet l'extraction de nouveaux descripteurs décrivant les deux modalités dans un espace de dimension réduite et leur relation : les descripteurs extraits sont associés pour créer *un mapping*. L'analyse des descripteurs extraits et leur mise en jeu dans le *mapping* permettent une interprétation sur les stratégies de contrôle adoptées par les participants. Il est alors envisageable d'analyser leurs cohérences au travers des performances d'un ou plusieurs participants. La méthode a en revanche beaucoup de limitations, et ne peut être utilisée que dans des cas adaptés que nous spécifierons dans la suite. De ces limitations est née une série de questions de recherche qui a dessiné les différentes directions d'investigation.

Nous résumons dans ce chapitre cette première étude exploratoire et fondatrice dans notre travail de recherche. Un article a été publié (Caramiaux et al., 2010c) et est reporté en annexe C.

Nous présentons ensuite les questions issues de cette exploration auxquelles nous apporterons des éléments de réponse dans les parties suivantes du manuscrit.

3.2 Exploration par l'expérience

A partir des données collectées, deux articles ont été écrits et publiés en 2010 ((Caramiaux et al., 2010c) et (Caramiaux et al., 2010a) reportés respectivement dans les annexes C et D). Le premier article est le fruit des premiers mois de notre travail de thèse. Le second article est une application à la resynthèse et sera plus largement discuté dans le chapitre 10 dédié.

Ainsi, du travail reporté dans (Caramiaux et al., 2010c), il est intéressant de remarquer qu'au delà de la contribution scientifique, celui-ci a impliqué une série de questions de recherche qui a conduit aux contributions reportées dans les parties II et III de ce manuscrit. Dans cette section, nous allons d'abord rappeler certains éléments de l'article (Caramiaux et al., 2010c), notamment la problématique initiale envisagée, l'expérience et les résultats. Ensuite, nous énoncerons les différentes problématiques qui en sont issues et qui seront abordées dans la suite du manuscrit.

3.2.1 Problématique et état de l'art

À partir des données collectées, le geste est décrit par des descripteurs cinématiques tels que la vitesse, l'accélération (tangentielle, normale), et géométriques tels que la courbure et la torsion. Les morceaux sonores sont analysés et un ensemble de descripteurs est calculé, notamment l'intensité sonore perçue, le centroïde spectral (approximation de la brillance), la hauteur, etc. (cf. (Peeters, 2004) pour une liste abondante de descripteurs audio, leur méthode de calcul, et leur interprétation). La problématique est l'analyse de *mapping* : quelles associations de descripteurs les participants effectuent implicitement et quelle méthode utiliser pour extraire l'association ?

Les gestes étudiés sont des gestes d'accompagnement. Dans la revue des travaux portant autour du geste d'accompagnement, nous avons noté que peu de ces études proposaient l'utilisation d'une méthode computationnelle pour l'analyse du lien entre geste d'accompagnement et son accompagné. L'investigation des méthodes d'analyse de données du geste musical nous a permis de constater que, parmi les méthodes proposées dans la littérature, l'analyse en composantes principales (ou *Principal Component Analysis* (PCA)) se révèle être la plus usitée (cf. la section A.1 pour la présentation de PCA et l'état de l'art). La PCA a plusieurs avantages : c'est une méthode simple, linéaire et facilement implémentable ; les composantes principales ainsi que les poids sur les variables initiales sont généralement interprétables ; la méthode est directement utilisable sans ajustement de paramètres ; elle est non-supervisée. Notre première idée a donc été l'utilisation de la PCA sur les données gestuelles d'accompagnement. Ceci aurait permis d'extraire le mouvement principal ainsi que les paramètres gestuels pertinents. Seulement, le geste effectué porte une information supplémentaire et fondamental que nous ne pouvions pas négliger : il est associé à un son. Nous nous sommes alors orientés vers l'utilisation de l'analyse multimodale. La figure 3.1 illustre la démarche analytique adoptée pour cette première étude. Une revue de la littérature a montré que ce type d'analyse s'est fortement développé dans le domaine de la reconnaissance de la parole basée sur des données audiovisuelles. Nous nous sommes dès lors inspirés de ces travaux.

En reconnaissance de la parole basée sur des données mulitmodales audiovisuelles, un problème important est l'intégration (ou fusion) des modalités (visuelles et auditives) pour décider de la classification. C'est une instance du problème plus général de la combinaison de classifieurs statistiques (Jain et al., 2000; Potamianos et al., 2004). Dans le domaine de la reconnaissance de la parole basée sur des données audio et visuelles (habituellement le mouvement des lèvres), plusieurs stratégies sont fournies et peuvent se catégoriser en trois types

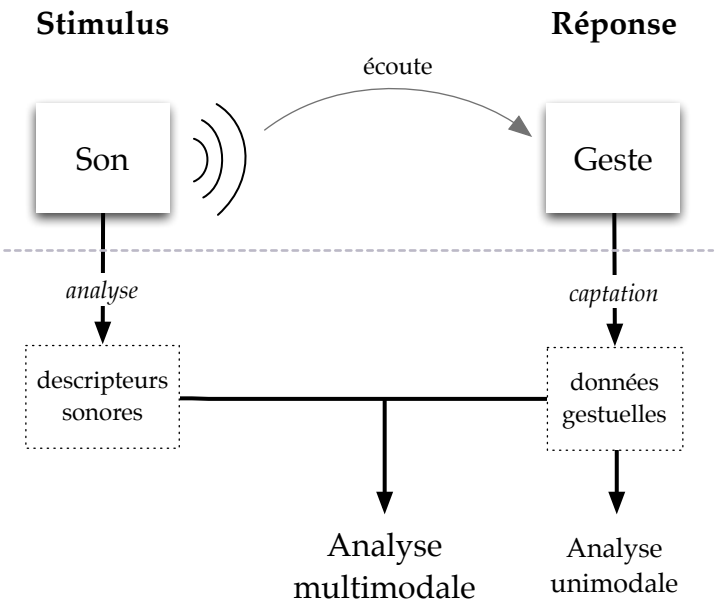

FIGURE 3.1 – Analyse multimodale entre descripteurs gestuels et sonores, adoptée dans l'étude

(Potamianos et al., 2004; Gurban, 2009) :

- *intégration anticipée* : intégration des modalités au niveau de l'échantillon ;
- *intégration intermédiaire* : intégration au niveau du phonème ou du mot ;
- *intégration tardive* : combinaison discriminante des modèles qui usuellement se basent sur des probabilités de combinaisons de différents mots / phrases.

L'analyse multimodale permettant l'extraction du *mapping* implicite entre les données de descriptions gestuelles et sonores est une intégration anticipée. Dans ce domaine, des auteurs (Sargin et al., 2007; Kidron et al., 2007) ont proposé l'utilisation de l'extension multimodale de la PCA, appelée Analyse Canonique (*Canonical Correlation Analysis* (CCA)). C'est cette méthode qui sera retenue et appliquée tout au long de cette première étude.

3.2.2 Analyse canonique

Nous proposons d'utiliser une méthode de fouille de données multimodales, appelée Analyse des Corrélations Canoniques (ang. *Canonical Correlation Analysis* (CCA)) (Hotelling, 1936). Cette méthode est une généralisation de la PCA (cf. Annexe A.1) à deux ensembles de données distincts et potentiellement de natures différentes. Dans notre cas d'étude, la CCA s'applique aux deux ensembles de données correspondant aux descripteurs sonores d'une part et aux paramètres gestuels d'autre part, pris comme des séries temporelles multidimensionnelles. Notons ces deux ensembles $\mathbf{Y}^1$, $\mathbf{Y}^2$. Les lignes de $\mathbf{Y}^1, \mathbf{Y}^2$ sont les échantillons des descripteurs gestuels et sonores considérés. Les colonnes sont les descripteurs. $\mathbf{Y}^1$ et $\mathbf{Y}^2$ doivent avoir le même nombre de lignes (même nombre d'échantillons) mais pas nécessairement le même nombre de colonnes (même nombre de descripteurs). La CCA retourne deux matrices de projection $\mathbf{A}, \mathbf{B}$ générant deux nouveaux ensembles de données $\mathbf{X}^1 = \mathbf{Y}^1\mathbf{A}$ et $\mathbf{X}^2 = \mathbf{Y}^2\mathbf{B}$ de même dimension et tel que les colonnes $\mathbf{x}_i^1, \mathbf{x}_i^2$ sont corrélées, les coefficients de corrélation r_i sont maximales et respectent $r_1 > r_2 > \ldots$. De même que PCA, les colonnes $\mathbf{x}_i^1, \mathbf{x}_j^2$ (avec $i \neq j$) sont non-corrélées. Les colonnes des matrices projetées sont appelées composantes canoniques.

Ainsi, les composantes canoniques sont des séries temporelles qui constituent le *mapping* : la i-ème composante canonique gestuelle est liée à la i-ème composante canonique sonore

associée, et la force de leur liaison est donnée par le coefficient de corrélation associé r_i. De plus, une composante canonique est une combinaison linéaire des variables originales. Ainsi, il est possible de savoir quelles sont les contributions de chacune des variables originales dans la composante et, a fortiori, dans le *mapping*.

3.2.3 Stimuli et procédure

Stimulus sonore

Les stimuli sonores utilisés étaient de types très différents. En effet, le corpus compte des extraits musicaux (de différents genres), des sons environnementaux (animaux, objets), etc. La liste complète des sons du corpus est reportée dans le tableau 3.1 ainsi que leurs durées. Tous les sons sont monophoniques, ayant une fréquence d'échantillonnage de 44.1kHz et une résolution de 16 bits.

Id.	Description	Durée (s)
1	Accordéon (mélodie polyphonique)	2.1
2	Contrebasse (mélodie monophonique)	7.8
3	Chant d'oiseau (sifflement, monophonique)	4.0
4	Croassement de corbeau	5.5
5	Valse, extrait de *Donauwalzer*	37.5
6	Pièce contemporaine pour *Flûte* (extrait 1, sequenza Berio)	6.4
7	Pièce contemporaine pour *Flûte* (extrait 1, sequenza Berio)	7.8
8	Guitare flamenco	5.7
9	Pièce rock, guitare électrique jouée par Hendrix	20.6
10	Voix (extrait d'une lecture de Kafka)	5.9
11	Chant diphonique	3.7
12	Pulsations	6.1
13	Bruits vocaux	10.0
14	Pièce contemporaine pour *Trombone* (extrait 1, sequenza Berio)	10.1
14	Pièce contemporaine pour *Trombone* (extrait 2, sequenza Berio)	4.9
16	Gouttes d'eau	2.5
17	Une vague déferlante	6.9
18	Plusieurs vagues déferlantes	22.5

TABLE 3.1 – Liste de sons utilisés durant l'expérience exploratoire réalisée en mars 2008. Pour chaque son, les participants ont dû effectuer un geste en réponse.

La raison pour laquelle nous choisissons un corpus aussi éclectique est la possibilité d'explorer un grand nombre de stratégies potentiellement différentes. En effet, la notion d'incarnation gestuelle dans le son a différents aspects suivant si le son est créé par un musicien (Leman et al., 2009) ou s'il est abstrait (Godøy, 2006).

Procédure

Dans l'expérience présentée ici, nous nous focalisons sur le mouvement de la main. Les participants tiennent un module (*solid body*) muni de marqueurs réfléchissants captés par des caméras infra-rouges (voir figure 3.2). Ainsi, les données gestuelles brutes sont : la position dans le repère cartésien 3D et les angles de rotation. La tâche donnée aux participants est d' « effectuer des mouvements de la main en s'imaginant qu'ils sont en train de produire le son ».

Deux pédales MIDI étaient placées en face d'eux (comme illustrées sur la figure 3.2). Ces pédales servaient au déclenchement des sons. A l'appui de la pédale de gauche, les sons sont joués et les participants pouvaient s'entraîner à effectuer un geste en association avec le son écouté. Durant la phase d'entraînement, les participants étaient libres d'écouter les sons un nombre quelconque de fois. Lorsqu'ils étaient satisfaits de leur geste, ils l'exécutaient à l'identique trois fois. Pour ce faire ils utilisaient la pédale de droite qui servait à déclencher le son tout

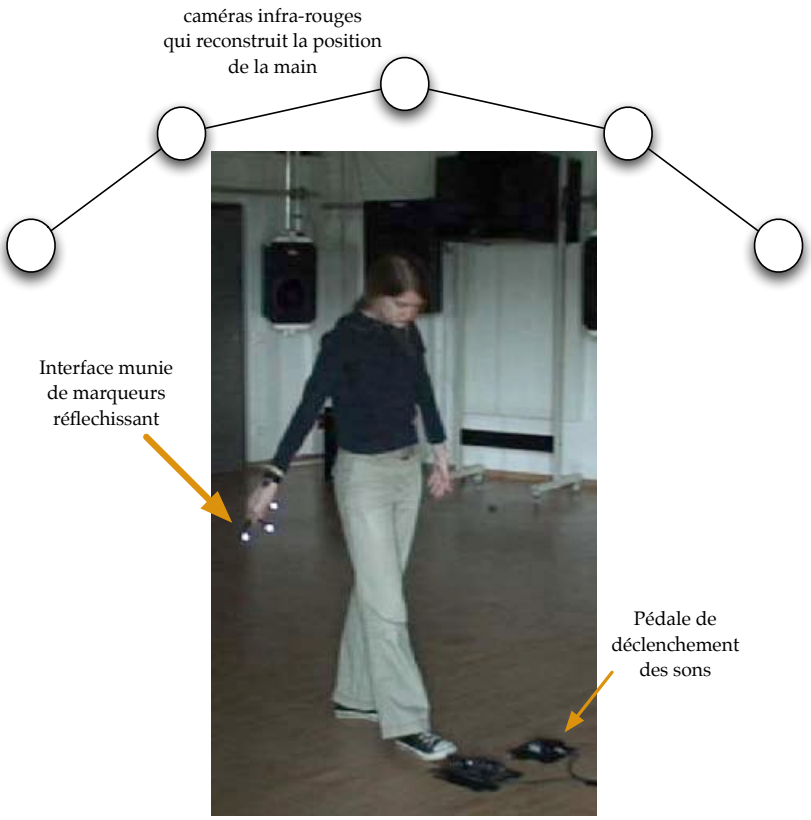

FIGURE 3.2 – Illustration de l'installation pendant l'expérience. Le participant est devant deux pédales MIDI utilisées pour le déclenchement des sons. Il effectue un mouvement avec dans la main l'interface munie de réflecteurs captés par un ensemble de caméras infra-rouges utilisées pour reconstruire le mouvement dans l'espace 3D.

en démarrant l'enregistrement. Une fois les trois gestes enregistrés, le passage au son suivant était automatique et la phase d'entraînement reprenait avec la pédale de gauche.

3.2.4 Analyse des données

Chaque son est analysé afin d'extraire un ensemble réduit de deux descripteurs sonores que sont l'énergie audio perçue et la brillance (corrélé au centroïde spectral). A chaque son est associé un geste de même durée que le son (car synchronisé à son début et à sa fin avec l'enregistrement sonore). Des paramètres gestuels capturés, nous gardons la position et nous dérivons des descripteurs cinématiques tels que la vitesse, l'accélération (tangentielle et normale), la dérivée de l'accélération et des descripteurs géométriques tels que la courbure et la torsion. Ainsi pour chaque son, nous avons un ensemble de 7 séries temporelles des descripteurs gestuels synchronisé au son, et ceci pour chaque participant.

3.2.5 Principaux résultats

Nous rappelons que les résultats reportés dans cette section sont publiés dans l'article (Caramiaux et al., 2010c). Pour chaque son, nous appliquons la CCA entre ses descripteurs sonores et les descripteurs gestuels de chaque participant. Comme énoncé précédemment, la CCA permet deux types d'analyse, complémentaires. Premièrement, les matrices de projection donnent la contribution des variables initiales dans les composantes canoniques. Deuxièmement, l'ordre des composantes canoniques et leurs coefficients de corrélation associés donnent le *mapping*. La base de données considérée comporte des sons très différents, donc on peut s'attendre à des interactions entre le geste effectué et le son très différentes.

La méthode s'est avérée très pertinente sur un sous-ensemble des sons initiaux. Nous en donnons l'exemple avec deux sons en particuliers : le son de *vague* et le son de *flûte* (extrait 2). La première analyse est la sélection des descripteurs gestuels prédominants. De manière globale pour les deux sons considérés, ces descripteurs gestuels prédominants dans le *mapping* sont : la norme de la position, la norme de la vitesse et l'accélération normale. Nous montrons que c'est cohérent avec certaines lois du mouvement qui mettent en lien des paramètres cinématiques et géométriques (loi de puissance 2/3 (Viviani and Flash, 1995)). La seconde analyse est le *mapping*, c'est à dire comment les descripteurs extraits (combinaisons linéaires des descripteurs prédominants) sont associés à l'intensité et la brillance. Pour ces deux sons et tous les participants, nous obtenons des résultats consistants pour les *mappings*. Ces résultats sont résumés de la manière suivante :

1. **La vague**. La norme de la vitesse et l'accélération normale sont corrélées avec l'intensité sonore perçue [1]. Ceci peut s'interpréter par un lien cognitif entre l'intensité sonore (qui est un descripteur sonore dominant dans l'audition (Susini et al., 2004)) et l'énergie cinétique (la vitesse étant elle-même un descripteur important pour une perception incarnée (Leman, 2007)).

2. **La flûte**. La norme de la position est corrélée avec l'énergie audio. Ce *mapping* montre que des changements fins de l'intensité sont incarnés par les participants par des changements fins de la position (à l'image d'un tracé (Godøy et al., 2006a)). Par analogie à l'énergie mécanique du mouvement, dans le cas de la flûte l'énergie potentielle du mouvement est substituée à l'énergie cinétique dans le *mapping*.

Dans les deux cas, la deuxième composante rend compte du lien entre la brillance (qui est approximée par le centroïde spectral) et la position (dans le cas de la vague) ou la vitesse (dans le cas de la flûte). Les stratégies sont inversées. La figure 3.3 illustre les corrélations, pour un participant, entre la vitesse et l'intensité sonore de la vague d'une part (graphique de gauche) et entre la position et l'intensité sonore de la flûte d'autre part (graphique de droite) [2].

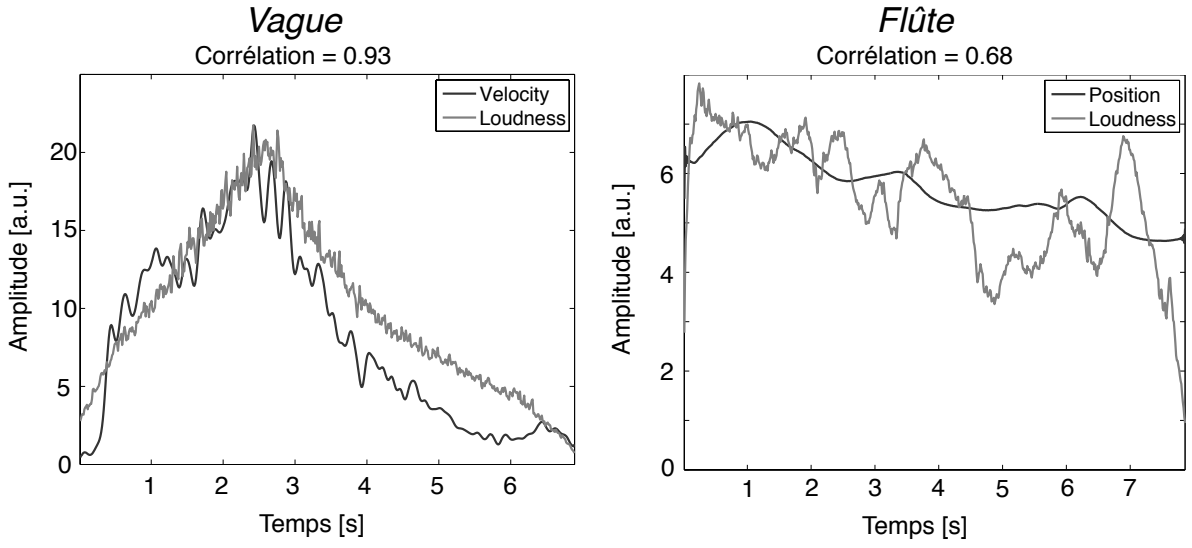

FIGURE 3.3 – Exemples de corrélations entre les descripteurs gestuels et sonores. À gauche sont reportées la vitesse du geste d'un participant et l'intensité sonore de la vague. À droite sont reportées la position du geste d'un participant et l'intensité sonore de la flûte.

1. Si deux paramètres contribuent au *mapping*, c'est qu'ils sont initialement corrélés.
2. Les exemples sont diponibles en ligne www.baptistecaramiaux.com/blog/?page_id=14

3.3 Questions de recherche

De l'expérience menée en 2008, plusieurs remarques ont pu être faites. Ces remarques ont ensuite été reformulées en questions de recherche auxquelles nous essaierons de répondre dans les parties suivantes.

3.3.1 Remarques liées à l'utilisation de CCA

Sur le type de sons

Les mouvements effectués en imaginant être en train de produire le son écouté peuvent grandement varier entre un son environnemental et un morceau de musique populaire. Au vu des vidéos, il apparaît des stratégies diverses suivant si le son comporte une hauteur prédominante (e.g. flûte, extrait 1), si le son est abstrait (e.g. pulsations), concret (e.g. vague), animé (e.g. le corbeau), non-animé (e.g. goutte d'eau), ou si le son a un rythme bien défini (e.g. Donauwalzer). Nous avons montré la cohérence des performances gestuelles pour les sons *vague* et *flûte*. Dans ce cas, les sons n'étaient pas concrètement liés à des gestes humains, soit en tant que source du son, soit dans un acte de composition. La CCA renvoie des résultats pertinents pour des sons ayant une morphologie acoustique qui peut être liée à celle du geste. De plus, ces morphologies gestuelles et sonores doivent être synchrones.

Sur la durée des sons

L'hypothèse initiale de l'utilisation d'une méthode comme CCA (ou PCA) est qu'une variable, donnée par ses réalisations, est une variable aléatoire et non un processus temporel. Ainsi, chaque variable (gaussienne) possède une moyenne et une variance constantes au cours du temps. Cependant, un geste effectué sur un son long ne peut être considéré comme stationnaire et donc sa structure temporelle doit être modélisée. Ceci a d'autant plus de sens lorsque le stimulus musical a une structure temporelle, telle que la valse *Donauwalzer*. Afin d'illustrer ce point, la figure 3.4 reporte les données de positions du geste effectué sur la valse *Donauwalzer* entre 7sec et 23sec. On voit clairement apparaître des récurrences dans le déroulement temporel. De plus cette structure est liée à la métrique du stimulus musical.

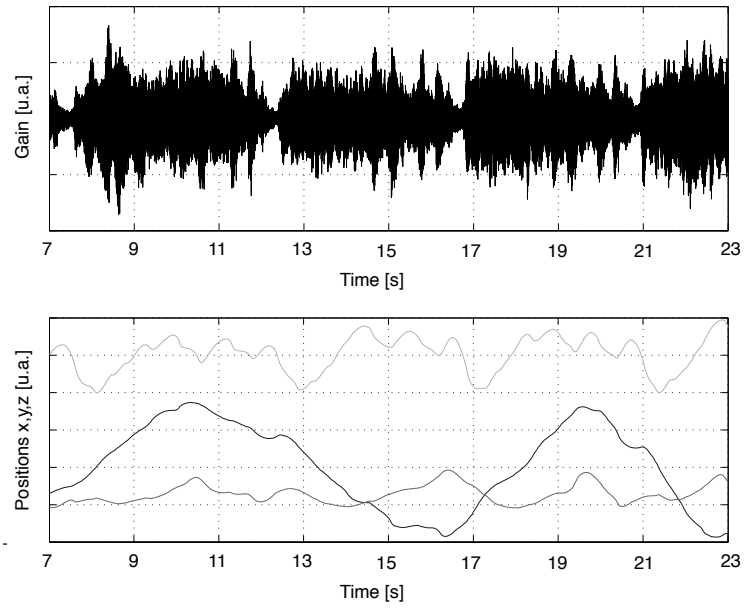

FIGURE 3.4 – Exemple de geste (en bas) effectué sur la valse *Donauwalzer* (en haut)

Ensuite, CCA fait l'hypothèse d'une relation linéaire. De la même manière, cette hypothèse ne peut tenir pour des sons longs ou la relation "globale" est grandement plus complexe. Ainsi la méthode s'applique à des sons courts, typiquement d'une durée allant de 1 à 5 secondes et ne comportant pas de structure avec des éléments récurrents. Les sons envisagés sont typiquement les *objets sonores* de Schaeffer (Schaeffer, 1966), c'est à dire ayant une morphologie temporelle bien définie (ceci sera plus amplement développé dans le chapitre 6).

3.3.2 Problématiques

De ces remarques, nous avons extrait les questions de recherche présentées ci-après. Nous apporterons des éléments de réponses au fil des chapitres du manuscrit. Les chapitres liés aux questions de recherche sont spécifiés.

Questions portant sur les réponses gestuelles à des stimuli sonores

Au vu des remarques précédentes, nous nous focaliserons sur des sons en tant que stimuli et non sur des extraits musicaux afin de mettre l'accent sur le lien geste–propriétés acoustiques plutôt que d'interroger le rythme et les structures musicales. Ici nous interrogeons les différentes stratégies cognitives de contrôle dans un contexte d'écoute. Les sons présentés dans l'étude exploratoire ne sont pas explicitement liés à une action les produisant :

- *Quel est l'impact de l'identification de la causalité sur les réponses gestuelles ?*
  [chapitre 5]

Si une action produit le son écouté :

- *Quel est le lien entre cette action et la réponse gestuelle ?*
  [chapitre 5]

S'il n'y a pas d'action perceptible.

- *Quel est le lien entre le son et la réponse gestuelle ?*
  [chapitres 5, 6 et 10]
- *Peut-on extraire un geste sonore à partir de la réponse gestuelle ?*
  [chapitres 5, 6 et 10]
- *Le lien entre geste sonore et réponse gestuelle est-il une correspondance des trajectoires des paramètres (ou morphologies) comme mis à jour par la CCA ?*
  [chapitres 5, 6 et 10]

Enfin, notre domaine de recherche étant l'informatique et non la musicologie systématique.

- *Comment peut-on pratiquement utiliser les gestes d'accompagnement et l'analyse de leur relation avec le son pour des applications musicales ou de design sonore ?*
  [chapitres 10, 11]

Questions portant sur la modélisation des structures temporelles du geste

La CCA a permis de mettre à jour les corrélations entre combinaisons linéaires des descripteurs gestuels et sonores et le besoin de modéliser la structure temporelle des descripteurs gestuels, notamment dans un but de contrôle afin de dépasser les limitations du simple *mapping*.

- *Quel modèle de structure temporelle pour la trajectoire des descripteurs ?*
  [chapitre 8]

- *Comment prendre en compte les variations de certains paramètres du geste ?*
  [chapitre 8]
- *Comment contraindre le modèle afin d'être pris en main par l'utilisateur ?*
  [chapitres 8, 9, 10]

Nous nous sommes aperçus dans le cas de la valse *Donauwalzer* (figure 3.4) que les structures temporelles gestuelles peuvent être plus complexes.

- *Comment prendre en compte des considérations structurelles à une échelle plus élevée dans la modélisation du geste ?*
  [chapitres 9]
- *Quelles différences y a t-il dans l'analyse entre une modélisation à une échelle plus élevée et des approches basées sur le signal ?*
  [chapitres 9]

L'extraction d'une description plus haut niveau peut amener à vouloir caractériser différentes échelles sémantiques.

- *Comment extraire des informations de plus haut niveau sémantique ?*
  [chapitres 10]

Enfin, les modèles ont vocation à être utilisés pour des applications concrètes.

- *Quelles sont les contraintes amenées par le temps réel ?*
  [chapitres 8, 10]

Deuxième partie

Réponses Gestuelles à des Stimuli Sonores

Chapitre 4

Perception du son et de sa source

Dans ce chapitre nous faisons l'état de l'art des travaux portant sur la perception du son et de sa source. Motivé par les théories cognitives présentées dans la section 2.1, nous pensons que le geste effectué à l'écoute d'un son est intrinsèquement lié à la cause du son et à son identification.

4.1 Introduction

Parmi les gestes d'accompagnement, plusieurs stratégies gestuelles peuvent être analysées, celles-ci dépendent du stimulus sonore envisagé. Leman (Leman, 2007; Leman et al., 2009) suggère qu'on perçoit dans la musique la trace du mouvement du musicien, comme une empreinte corporelle encodée dans le flux acoustique. Notre perception (celle de l'auditeur) est alors capable de décoder à partir de ce flux acoustique la trace du geste du musicien et d'en extraire une partie de l'information par l'exécution d'un geste en réponse au flux acoustique. D'autre part Godøy (Godøy, 2006) suggère qu'un geste est aussi perceptible dans des sons plus abstraits tels que ceux dans la musique *acousmatique* [1]. Un objet sonore se caractérise par un phénomène sonore perçu comme un tout, une unité sonore perçue dans sa matière et donc ayant une forme (comme un objet). Cette forme est donnée par l'évolution de descripteurs acoustiques tels que l'intensité sonore. En cela, la perception de l'objet sonore repose sur les formes qu'il peut prendre, ceci pouvant être mis en lien avec la notion de *trajectoire* et a fortiori de *geste*. L'idée globale de ces études est de comprendre, d'un point de vue comportemental, comment la perception du son peut être liée au geste.

Dans cette partie, nous proposerons d'analyser deux aspects liés aux stratégies de réponse gestuelle selon la perception d'un stimulus sonore. Premièrement, nous inspecterons les stratégies gestuelles liées au degré d'identification d'un son, c'est à dire de l'identification de la cause de ce son. Ensuite, nous transposerons le discours d'un point de vue plus musical en rejoignant l'approche de Godøy et ses collègues (Godøy, 2006), à savoir la spécification d'un corpus d'objets sonores abstraits dont l'action causale n'est pas identifiable. Les sons considérés sont synthétisés de manière à contrôler les profils temporels de leurs descripteurs. Le but est l'analyse du lien entre forme, ou *morphologie*, et trajectoire de la réponse gestuelle.

Ce chapitre présente un état de l'art sur la perception du son et de la musique dans son lien avec l'action et le geste. Notre volonté est de réunir les travaux relatifs à cette thématique provenant de domaines de recherche différents : étude de la parole, psychoacoustique, neurosciences, musicologie. Premièrement, la section 4.2 reporte des études en psychoacoustique et neurosciences sur l'identification des sons (identification de leur source causale). Comme l'identification est liée à l'interprétation sémantique du son, nous commençons quelques rappels pour le cas de la parole. Ensuite, la section 4.4 se concentre sur la perception musicale et

1. *"adjectif, se dit d'un bruit que l'on entend sans voir les causes dont il provient"* (Chion, 1983), p.18

plus précisément l'écoute musicale.

4.2 Perception de la parole et contrôleur moteur

La théorie sur le geste lié au langage formulée par Kendon (Kendon, 2004) montre la complémentarité du geste et de la parole pour créer une unité sémantique cohérente. L'auteur écrit :

> We shall see that speakers create *ensembles* of gesture and speech, by means of which a semantic coherence between the two modalities is attained. This is not to say that speech and gesture express the same meanings. They are often different. Nevertheless, the meanings expressed by these two components *interact* in the utterance and, through a reciprocal process, a more complex unit of meaning is the result. ((Kendon, 2004), chapitre 7, p.108–109)

Kendon met en évidence une complémentarité indissociable entre le langage parlé et le geste dans le sens de l'énoncé. Une unité sémantique complexe est créée par les deux modalités pouvant venir d'un seul processus cognitif. Cette vision est partagée par plusieurs auteurs. Une revue complète des travaux portant sur l'étude du rôle du geste et du langage dans les unités sémantiques va au delà du cadre de nos recherches. Nous citerons pour exemple les travaux de Özyürek (Özyürek, 2010) qui propose une série d'études basées sur des analyses au niveau cérébral et comportemental, mettant en avant l'interaction entre geste et parole à la fois dans la compréhension et dans la production du langage.

Dans le domaine des neurosciences, Rizzolatti et al. (Rizzolatti and Craighero, 2004) proposent une revue des études portant sur les neurones miroirs pour l'audition. Ils montrent l'existence de ce qu'ils appellent les neurones échos qui "résonnent" à l'écoute de sons verbaux. En d'autres termes, à l'écoute de stimuli verbaux, il y a une activation des centres moteurs liés à la parole. Fadiga et al. (Fadiga et al., 2002) ainsi que Watkins et al. (Watkins et al., 2003) ont proposé deux expériences qui confirment cette conclusion. Liberman et al. (Liberman and Mattingly, 1989; Liberman and Mattingly, 1985) avaient proposé qu'un tel processus neuronal serve d'intermédiaire pour la perception du langage. Xu et al. (Xu et al., 2009) précisent cette idée en montrant que le geste (pantomime ou langage des signes) est traité par les mêmes zones du cerveau que la parole, notamment l'aire de Broca.

Le lien entre le geste et le son dans le domaine du langage semble être dépendant de la *sémantique* transmise et reçue. En effet, les gestes pantomimiques et le langage semblent être traités par la même zone du cerveau, tandis que les gesticulations (pour reprendre le terme de Kendon) interagissent avec le langage pour faire émerger une sémantique complexe. Il paraît légitime de questionner la généralisation à des sons non liés à la parole. La sémantique de ces sons peut se comprendre comme l'identification de leur source.

4.3 Perception des sons environnementaux

VanDerveer (VanDerveer, 1980) définit un son environnemental comme "... any possible audible acoustic event which is caused by motions in the ordinary human environment [...]." Par cette définition, un son environnemental se définit par l'événement qui est la cause du son. VanDerveer va plus loin en précisant que cet événement est communément décrit en termes : d'actions (Warren and Verbrugge, 1984; Cabe and Pittenger, 2000), d'objets (Carello et al., 1998; Giordano and Mcadams, 2006), où du lieu où se passait l'action. La perception d'un son environnemental passe par l'identification de cette source, et l'étude psychoacoustique de la perception des sons environnementaux s'attache à caractériser l'information acoustique sur laquelle des auditeurs basent leur jugement pour identifier et catégoriser la source du son environnemental, c'est à dire des traces acoustiques de l'action, ou l'objet. Pour ce faire,

il s'agit d'analyser quelles sont les similarités perçues entre des sons par des auditeurs. Une première possibilité est une tâche de tri, permettant de spécifier certaines catégories par leurs propriétés acoustiques. Une deuxième possibilité est d'y adjoindre la verbalisation des classes constituées, permettant d'élucider les sources identifiées.

Dans une tâche de tri, une méthode standard est de représenter les sons d'une même catégorie par un ensemble de descripteurs et de réduire leur dimension par *Multidimensional Scaling* (MDS). Cette méthode permet la projection des données dans un sous-espace tout en conservant les distances respectives entre les éléments de l'espace. Cette méthode de réduction de dimension a été utilisée pour représenter les sons environnementaux. Un premier exemple sont les études pour des cas spécifiques de sons environnementaux : (Cermak and Cornillon, 1976) pour les sons de trafic et (Howard Jr, 1977) pour des sons sous-marins. Susini et al. (Susini et al., 2004) montre que les descripteurs prédominants dans la perception des sons issus des systèmes de climatisation sont le centroïde spectral et l'intensité sonore. Cette méthode a ensuite été étendue à des corpus de sons plus hétérogènes. Dans une première étude VanDerveer (VanDerveer, 1980) utilise une tâche de tri libre, sans consigne sur la similarité utilisée. Il conclut que la proximité des sons relève plus d'une proximité temporelle que spectrale. Bonebright (Bonebright, 2001) propose une représentation à 3 dimensions par l'analyse MDS. La première dimension correspond au plus haut niveau d'amplitude dans le son, l'intensité moyenne et le changement en fréquence (qui peut être corrélé au centroïde spectral). La deuxième correspond à la durée et la troisième par des pics de fréquences. Toutefois, la MDS utilisée pour le timbre a montré que la troisième dimension est moins consistante pour l'interprétation (Grey and Moorer, 1977; Krumhansl, 1989). Ces études permettent d'identifier des descripteurs sonores pertinents (comme l'intensité sonore et le centroïde du spectre) suivant le type de son environnemental.

L'identification d'un son environnemental est parfois vue comme l'illustration d'une classification cognitive : du système auditif au concept et du concept au langage (Goldstone and Kersten, 2003; McAdams, 1993; Smith, 1995). Peu de travaux existent sur la classification des sons environnementaux. VanDerveer (VanDerveer, 1980) montre que la classification obtenue semble provenir d'une similarité faite entre causes qui auraient produit le son. Guyot (Guyot, 1996) montre que la classification de bruits domestiques conduit à deux types de similarités : les sons sont groupés ensemble car ils sont produits par la même source ; ou ils sont groupés car ils sont faits par la même interaction. Marcell et al. (Marcell et al., 2000) utilise un large corpus sonore et une tâche de tri suivie d'une tâche de classification sur le résultat du tri (en suivant la méthode donnée dans (McAdams, 1993)) : les catégories obtenues sont classées dans 27 catégories plus générales avec des labels. Les catégories plus générales correspondent aux objets mis en jeux (e.g. *instrument de musique*), où l'événement se passe (e.g. *cuisine*). Mais les auteurs ont trouvé que les catégories liées aux qualités acoustiques du son ne sont que très rarement utilisées. Gérard (Gérard, 2004) propose une tâche de classification où la consigne est explicitement de classer les sons ensembles "s'ils peuvent être entendus ensemble dans la nature" ou classer les sons "sur la base de leurs propriétés acoustiques indépendamment de leur sens". Il montre que lors du premier test, les participants classent suivant des sons d'objets animés ou non-animés et lors du second test les participants classent en fonction des caractéristiques : rythmiques, de hauteurs, d'amplitudes et de motifs (cf. aussi (Guastavino, 2007)).

Ceci illustre que l'identification de sons environnementaux ne peut pas se faire exclusivement sur des propriétés acoustiques (cf. aussi (Lutfi et al., 2005)). Gaver (Gaver, 1993b; Gaver, 1993a) parle d'écoutes différentes : l'écoute du quotidien et l'écoute musicale. La première se focalise sur l'identification de la source alors que la deuxième s'attache à qualifier le son sur la base de ses propriétés acoustiques. Dans (Lemaitre et al., 2010), Lemaitre et al. proposent d'analyser ces écoutes ainsi définies et leurs répercussions dans une tâche de classification. Ils montrent que le type de similarité utilisée (acoustique ou causale) dépend : du niveau d'iden-

tification de la cause du son environnemental et de l'expertise des auditeurs.

La perception des sons environnementaux a aussi gagné de l'intérêt dans le domaine des neurosciences. Aziz-Zadeh et al. (Aziz-Zadeh et al., 2004) ont montré que les sons liés à des actions bi-manuelles montraient une plus grande activation de la structure neuronale motrice que des sons associés aux mouvements des jambes. En outre, dans cette étude, l'auteur examine la sémantique du son et son lien avec les zones cérébrales liées à la parole (comme l'aire de Broca) montrant par là que le codage de l'action peut être un précurseur au langage. Citons dans l'étude neuroscientifique de l'identification sonore, les travaux de Thierry et al. (Thierry et al., 2003) et Pizzamiglio et al. (Pizzamiglio et al., 2005). Dans cette dernière, les auteurs ont observé l'activité neuronale lors de deux tests d'écoute. Dans le premier test, les sujets écoutaient des sons causaux (i.e. dont la source du son est identifiable). Dans le deuxième test, les sujets écoutaient des sons non-causaux. Les auteurs ont pu observer deux zones neuronales pour l'identification du son. L'un traite le *sens* du son quand le système miroir audio-visuel est activé. En reprenant les termes de Berthoz (Berthoz, Alain, 1997), l'action est simulée. Le second est distinct du premier et est sollicité lorsque l'identification du son doit se baser seulement sur des propriétés acoustiques et perceptuelles du son lui-même sans la possibilité d'activer le système miroir. L'action n'est pas simulée. Ainsi, les écoutes *musicale* et *du quotidien* sont ici supportées par des travaux en neurosciences.

4.4 Perception musicale

Ici nous focalisons le discours sur la perception de la musique. En premier lieu, nous reportons de manière concise certains travaux sur la perception du timbre et le lien avec les descripteurs acoustiques. Ensuite, en cohérence avec la démarche de ce chapitre, nous inspecterons le lien entre perception musicale et source. Enfin, nous occulterons la source pour présenter l'*écoute réduite* de Schaeffer et son implication pour la relation entre geste et son.

4.4.1 Préambule

En psychoacoustique, des études se sont intéressées à l'analyse des catégories créées lors d'une tâche de tri et notamment comment décrire ces catégories à l'aide d'un nombre réduit de descripteurs acoustiques (Caclin et al., 2005; Krumhansl, 1989; McAdams et al., 1995). Ces études utilisent aussi la MDS. La problématique est que, mises à part l'intensité et la hauteur (qui sont des caractéristiques unidimensionnelles), le timbre est fondamentalement multidimensionel. Caclin et al. (Caclin et al., 2005) montrent que le temps d'attaque, le centroïde spectral, et un résiduel rendant compte de la structure fine du spectre sont des descripteurs prédominants pour le timbre. Ceci est cohérent avec les études précédentes, par exemple (McAdams et al., 1995), qui ont montré la prédominance du centre de gravité spectral, et aussi de la durée d'attaque (qui permet d'isoler les impacts des sons entretenus).

La perception musicale est particulièrement sensible à la hauteur (à la fois le contour global, des tailles d'intervalles spécifiques entre hauteurs ou localement des rapports de durées de notes) (Patterson et al., 2002) et au rythme (Zatorre et al., 2007). Sur ce dernier point, Zatorre et al. précisent que le concept émergeant semble être que l'analyse du rythme dépend de l'interaction entre les systèmes auditifs et moteurs. Ces recherches montrent l'importance dans la musique des structures hiérarchiques de rythmes et de hauteurs.

4.4.2 Lien avec la source

Dans *Embodied Music Cognition and Mediation Technology* (Leman, 2007) Leman écrit

> However, it is clear that a straightforward mapping between perceived stimuli features and sounding objects, for example, is not evident. Nor can all aspects of music perception

> be attributed to the recognition of the source of the stimulus. Music is often made of abstract sounds whose source may be difficult to define. Instead of the sources themselves, music often seems to focus attention on structures and relationships between sounds. It is possible therefore, that the emphasis on source should be considered from the point of view of human corporeality and corporeal synthesis, rather than from the viewpoint of external sources (Godøy, 2001). ((Leman, 2007), chapitre 2, p.48)

Ainsi la perception de la musique ne peut-elle se borner à l'identification de la source ou à un objet sonore abstrait. L'auteur propose de repenser la source du son (dans le contexte musical) comme incarnée plutôt que comme une source extérieure. En cela, la source est liée aux actions corporelles effectuées pour produire le son ainsi que sa représentation liée à la construction de notre propre perception. L'auteur va plus loin que les précédentes énonciations de *l'écoute musicale* et pose clairement l'approche d'une cognition *incarnée* de la musique.

La perception musicale liée à la production a eu un écho en neurosciences. Lahav et al. (Lahav et al., 2007) ont fait l'expérience suivante. Des sujets non-musiciens ont dû apprendre une mélodie simple en quelques jours. Ensuite, cette même mélodie leur était diffusée (sans qu'eux-mêmes ne jouent) et un système d'imagerie cérébrale observait l'activité neuronale. Les auteurs ont alors pu observer que le système de neurones miroirs s'active au moment de l'écoute. Ensuite les sujets ont écouté les mêmes notes de la mélodie mais dans un ordre différent. Pour ce second stimulus, l'activation était moindre. Enfin, le dernier test était l'écoute d'une mélodie construite sur la même base logique que les premières mais dont les notes constitutives de la mélodie n'appartiennent pas à la mélodie initiale. Dans ce cas, le réseau de neurones miroirs ne s'active pas. Cette expérience montre ainsi que le lien cognitif entre action et perception est lié à l'apprentissage sensori-moteur.

4.4.3 Écoute réduite

Ceci nous amène à étudier le cas où l'auditeur n'identifie pas la source du stimulus sonore musical perçu. Dans ce contexte nous nous focalisons sur une approche musicologique appelée *écoute réduite*.

Les modes d'écoute de Gaver (Gaver, 1993b; Gaver, 1993a) font écho aux écoutes définies par Schaeffer (Schaeffer, 1966). Schaeffer théorise les écoutes possibles dans son "Traité des Objets Musicaux" (Schaeffer, 1966), qui a été repris par Chion (Chion, 1983) dans un guide didactique. Schaeffer définit quatre types d'écoute qui sont résumés dans la figure 4.1 repris de l'ouvrage de Chion. Comme le rappelle très clairement l'auteur ((Chion, 1983), p.25), ils reprennent deux dualismes connus : abstrait/concret et subjectif/objectif. Le premier dualisme a été présenté dans la section précédente sur les sons environnementaux, mais aussi avec (Pizzamiglio et al., 2005). Le second dualisme concorde avec les premiers écrits sur l'enaction dans le livre de Varela et al. (Varela et al., 1991) : la définition d'une voie moyenne entre subjectivisme et objectivisme, l'interaction action–perception.

Les quatre modes d'écoute définis par Schaeffer se définissent par :

1. **Écouter** : c'est prêter l'oreille afin d'identifier la source, ou trouver des indices qui nous amèneraient à cette source
2. **Ouïr** : c'est la perception brute du rayonnement acoustique dans le système auditif. C'est une attitude en soi passive.
3. **Entendre** : c'est qualifier, c'est à dire manifester une intention d'écoute.
4. **Comprendre** : c'est saisir un sens, traiter comme un signe qui renvoie à une entité sémantique.

De cette catégorisation, Schaeffer définit l'écoute réduite comme étant l'écoute du son pour lui-même, comme ce qu'il appelle l'*objet sonore*, sans chercher à identifier sa source, sa provenance ou son sens. Comprendre l'écoute réduite c'est donner accès à l'objet sonore. Schaeffer propose

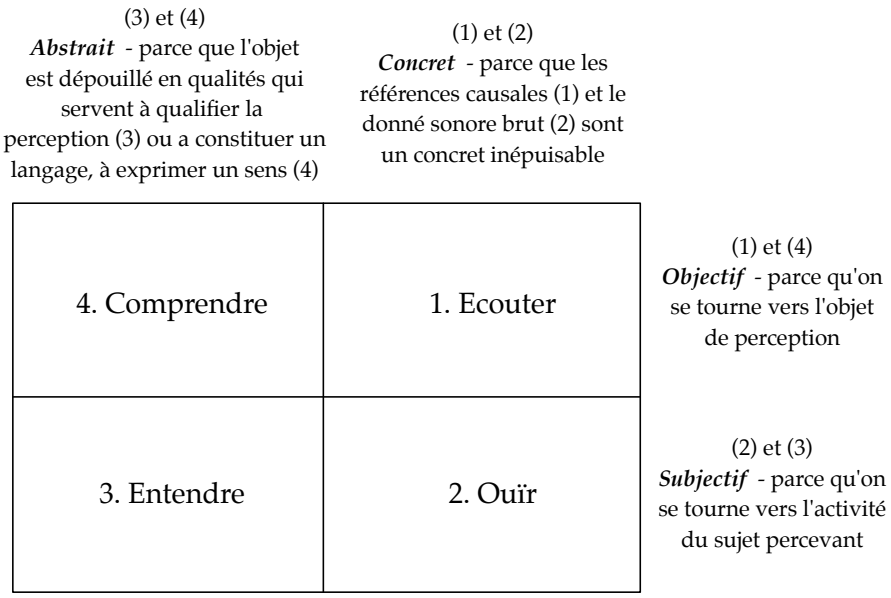

FIGURE 4.1 – Tableaux des écoutes de Schaeffer, repris du guide de Chion (Chion, 1983)

ensuite une typomorphologie, c'est à dire une taxonomie des objets sonores basée sur leur forme temporelle. Le lecteur intéressé peut se référer aux travaux complémentaires suivants : (Smalley, 1997; Thoresen and Hedman, 2007; Peeters and Deruty, 2009). La typomorphologie des objets sonores peut se résumer globalement par :

1. **Impulsif** : son bref sans entretien
2. **Itératif** : son dont l'entretien se prolonge par itérations c'est à dire une suite d'impulsions
3. **Entretenu** : son dont l'entretien se prolonge de manière continue dans le temps

Cette typologie de son est étroitement liée avec le geste. Schaeffer parle de facture gestuelle et lie cette typologie avec des gestes ponctuels, itératifs ou continus. Godøy (Godøy, 2006) étend le concept d'objet sonore à l'*objet gestuel–sonore* (ou *gestural–sonorous objects*) c'est à dire la prise en compte de la perception de l'objet sonore en tant que trajectoire gestuelle dans le son, traçant par ses contours acoustiques, la morphologie de l'objet.

4.5 Synthèse

Dans ce chapitre nous avons étudié le lien entre perception du son et perception de la source, notamment la cause du son en termes d'action ou d'objet produisant le son. Les travaux précédents sur la perception des sons environnementaux ont mis en exergue deux types d'écoute à savoir l'écoute du quotidien et l'écoute musicale. La première se focalise sur l'identification de la source alors que la deuxième s'attache à qualifier le son sur la base de ses propriétés acoustiques. En musique, nous avons vu que la perception musicale ne peut pas être que liée à la source mais plutôt incarnée. L'extrême est alors l'écoute réduite qui occulte toute source et se focalise sur les morphologies des qualités acoustiques des sons.

Dans la suite de cette partie nous présentons deux contributions. La première se focalise sur cette distinction entre écoute du quotidien et écoute musicale pour analyser dans quelle mesure les gestes en réponse à des stimuli sonores environnementaux sont affectés par ces modes d'écoute. Cette première étude expérimentale correspond au chapitre 5 reprenant l'article (Caramiaux et al., 2012b).

La seconde étude sera liée à la perception musicale et notamment à l'écoute réduite. Elle analysera les gestes associés à des stimuli sonores abstraits (sans source identifiable) mais dont

les morphologies sont explicites et contrôlées. Cette deuxième étude expérimentale correspond au chapitre 6 reprenant l'article (Nymoen et al., 2011).

Chapitre 5

Study of the impact of sound causality on gesture responses

B. Caramiaux[1], **P. Susini**[1], **O. Houix**[1], **F. Bevilacqua**[1]

[1] UMR IRCAM-CNRS, Paris, France

Abstract : In this paper, we investigate the gesture responses of participants listening to sound stimuli whose causal action can be identified or not. We aim at providing behavioral issues on the gestural control of sound according to the sound causal uncertainty permitting the design of new sonic interaction to be enhanced. We present a first experiment conducted to build two corpora of sounds. The first corpus contains sounds related to identifiable causal actions. The second contains sounds for which causal actions cannot be identified. Then, we present a second experiment where participants had to associate a gesture to sounds from both corpora. Also they had to verbalize their perception of the listened sounds as well as their gestural strategies by watching the performance videos during an interview. We show that when the sound causal action can be identified, participants mainly mimic the action that has produced the sound. On the other hand, participants trace the acoustic contours of timbral features when no action can be associated as the sound cause. In addition, we show from gesture data that the inter-participants gesture variability is higher if considering causal sounds rather than non-causal sounds. Variability demonstrates that in the first case, people have several ways of producing the same action whereas in the non-causal case, the common reference is the sound stabilizing the gesture responses.

Keywords : Environmental Sound, Sound Perception, Gesture Analysis

5.1 Introduction

In recent years, there has been a great deal of interest in understanding the effect of music on human motion. Different approaches have been followed involving : behavior (Leman et al., 2009; Godøy et al., 2005) ; psychology (Eitan and Granot, 2006) ; and neuroscience (Zatorre et al., 2007). The motivations are various including purely theoretical aspects and strategies for sonic interaction design. Our goal is to understand what in the sound people aim at revealing when performing gestures in response to sound stimuli. We are inspired by previous works driven by embodied music cognition notions (Leman, 2007) that make theoretical links between the perceived music and body motion. For instance, music embodiment can refer either to instrumentalists' gesture in pitch melody (Leman et al., 2009) or to abstract gestures from acousmatic sounds (Godøy et al., 2005). In both examples, gestures occur intrinsically in the acoustic flux. Both stimuli can be differentiated according to their causality, i.e. the identification of the sound cause. In the instrumental case, the causality is clear while in abstract sound, no specific cause can be associated to the sound. Our focus in this paper is to link the sound's causality to gesture responses.

Sound causality has been particularly studied for environmental sound perception. VanDerveer pointed out that environmental sounds are defined and perceived by the identification of their sources, the events having caused the sound (VanDerveer, 1980). The definition induces that the source is identifiable during the listening process. Gaver proposed to differentiate two types of listening (Gaver, 1993a; Gaver, 1993b) defined as : *musical listening* that focuses on the sound qualities of the acoustic signal ; and *everyday listening* that focuses on the event causing the sound. Gaver added "the distinction between everyday listening and musical listening is between experiences, not sounds" ((Gaver, 1993b), p. 1).

A challenging task is to study how the listener is able to identify the sound sources (either environmental and caused by an action or musical) based on acoustic properties. The common methodology to respond to this problem is to ask listeners to sort sounds between categories that are further labeled (classification) or not (McAdams, 1993). Early works have shown that such categories reflect mostly the action or the object having caused the sound (VanDerveer, 1980). This has been later specified by showing that during a free sorting task, the event having caused the sound (kind of action/interaction, type of excitation, source) is a criterion to categorize sounds but less often their acoustic qualities (Marcell et al., 2000).

The distinction in environmental sound perception between perceiving based on events causing the sound or acoustic qualities has received little attention. Gerard et al. (Gérard, 2004) explicitly did the distinction in the task by asking one group of listeners to sort together sounds "which they may hear together in the environment" while another group had to sort together sounds "on the basis of their acoustical characteristics, independently of their meaning". Interestingly, Lemaitre et al. (Lemaitre et al., 2010) showed recently that the categorization based on either action/object or acoustic qualities depends on the sound identification and listener's expertise. In other words, the more we can identify a sound source, the more the sounds will be categorized based on the action/object that caused the sound, while no identifiable source leads to categorization based on acoustic properties. Also, experts from sound related fields will more easily use acoustic features for categorization rather than non-experts. Finally, certain authors pointed out the importance of the context (Li et al., 1991; Repp, 1987).

Based on the previous mentioned studies involving two main types of sound identification, we are interested in inspecting how gestures performed in response to environmental sound stimuli are affected by the type of identification used when perceiving a sound.

Studies of body responses to sound stimuli mainly belong to musically motivated approaches. Music highly influences timing in motor behaviors (see Zatorre et al. for a review on the link between auditory and motor systems in music performance and perception (Zatorre et al., 2007)). The first experiments investigated tapping task on beats (Large, 2000; Large and Palmer, 2002) highlighting anticipations and asynchronies. More explorative studies tried to inspect free body gestures related to sound stimuli (Godøy et al., 2006c; Godøy et al., 2006b; Nymoen et al., 2010; Caramiaux et al., 2010c; Nymoen et al., 2011). Godøy et al. (Godøy et al., 2006c) studied how piano performance mimicry can give insight on pianists' musical expersion. In a second study, Godøy et al. (Godøy et al., 2006b) asked participants to *trace* sound stimuli on a 2-dimensional surface leading to highly varying strategies and the need to specify the corpus. Nymoen et al. (Nymoen et al., 2010) focused on abstract synthesized sounds (controlled by evolution of the acoustic features : loudness, brightness and pitch) and investigated the link between the sounds and gestures performed with a rode. Support Vector Machines were used for the analysis but did not discriminate between gesture–sound relationships. In parallel, Caramiaux et al. (Caramiaux et al., 2010c) analyzed multi-modal relationships between gesture kinematic features and sound features by using Canonical Correlation Analysis (CCA). A wide sound corpus was used leading to consistent control strategies for short and non action-related sounds. Finally, CCA has been recently used between gesture responses and controlled abstract synthetic sounds confirming the intrinsic mapping between 1/ spectral features or energy with the movement's velocity and 2/ the pitch and the vertical position (Nymoen et al., 2011). Also

position and velocity were never used simultaneously for the control of sound features.

From reviewed works, it remains unclear how the perception of action-related and non action-related sounds, either linked to *everyday* and *musical* listening, are linked with gestural responses. Our hypothesis, based on the recent work done by (Lemaitre et al., 2010), is that people mimic the action that has produced the sound when the cause, in terms of action, can be identified (*action-related sounds*). On the other hand, people draw the sound contours (or follow acoustic features' evolutions) if responding to sounds whose causes in terms of action cannot be identified (*non action-related sounds*). To validate this hypothesis, two experiments are presented in this paper. In experiment 1 (section 5.2), we report how the two sound corpora are built (i.e. action-related vs. non action-related) based on a listening task during which participants had to rate their confidence in identifying the action causing the sounds. The main idea here was to provide two sets of sounds for the second experiment that tests our hypothesis. Experiment 2 (section 5.3) was conducted with the corpora created from experiment 1. Participants were asked to associate a gesture synchronously to the sound (i.e. while they are listening to it) and were then asked some questions on their performance. Two groups have been made : one assigned to each corpus. We investigated from interviews either if the participants consciously performed gestures that either mimic the action causing the sound or draw acoustical contours. Also, we investigated if the gestural responses allowed for discriminating action and non-action related sounds. Finally, we discuss the results in section 5.4 and conclude in section 5.5.

5.2 Experiment 1 : Building a non-causal sound corpus

Both action related and non-action related sound corpora have been created based on a listening test inspired by the one used in (Lemaitre et al., 2010). Lemaitre et al. conducted three studies inspecting the classification of sounds according to their level of identification and listener's expertise. The authors proposed an ordering of sounds based on their *causal uncertainty* (Ballas, 1993).

5.2.1 Participants

Twenty one non-musician participants (ten males and ten females) were recruited outside of the institute for this study which took place at Ircam – Centre Pompidou in Paris. None of them reported having hearing problems. All participants gave informed consent to participate in the study. The experiment took approximately thirty minutes, and participants were given a nominal fee.

5.2.2 Stimuli

An initial sound corpus was taken from the study by Lemaitre et al. (Lemaitre et al., 2010). The sounds used belonged to a domestic context (usual objects found in a kitchen), to ensure that the sources of the sounds were likely to be known to all listeners. The sounds were classified according to the *causal uncertainty* (Ballas, 1993) which measures how consistent listeners were at describing the sound cause in terms of action or object. A low value indicates that they are consistent, and a high value indicated the opposite. Each sound identification is calculated through the causal uncertainty index (noted H_{cu}) (Ballas, 1993), (Lemaitre et al., 2010) which measures the identification of the cause in terms of the action's and/or object's verbalized description. Each sound has a H_{cu} index scaled between 0 (i.e. all the participants provided the same description of the sound in terms of action or object) and 4.75 (all the participants provided a different description in terms of action or object). From the whole set of sounds we retained the 10 best identified sounds (mean1.92s, std 1.51s). These sounds form the action

related corpus a with low causal uncertainty measured in the study by (Lemaitre et al., 2010), meaning the ten selected sounds were supposed to be easily identified in terms of action or object causing them.

From this corpus, we created ten transformed versions of the previous sounds as follows. The original sound is analyzed in Mel bands and white noise is convoluted by the analyzed bands. The resulting sound had a similar spectral evolution than the original (for instance preserving the energy, the first Mel coefficient) while avoiding high frequencies. Inspired by previous works on sound identification based on spectral resolution (Gygi et al., 2004; Shafiro, 2008), this process aims to affect recognition of the causality of the original sounds. In the following, the corpora will be called non-transformed and transformed corpus.

We used the same ecological adjustment of sound levels given in (Lemaitre et al., 2010). As pointed out by the authors, the sounds were recorded with different techniques, including near field and far field recordings. The relative levels of the sounds were not coherent. In (Lemaitre et al., 2010), the authors asked for participants to evaluate the level of the sounds according to a reference. Sound volumes are set to respect the ecological sound level.

All the sounds from both corpora were monophonic and had 16-bit resolution and a sampling rate of 44.1kHz.

Id.	Description (Lemaitre et al., 2010)	Max. level(dB)	Duration(s)
1	Cutting bread	61	1.2
2	Closing a cupboard door	72	1.4
3	Opening a drawer with castors	71	1.7
4	Champagne cup shocked	59	1.0
5	Knife removed from his case	66	1
6	Pouring cereals into a bowl	68	5.1
7	Bottle top	58	0.72
8	Removing a cork stopper	63	1.5
9	Crushing a metallic can	65	1.2
10	Crumpling a plastic bag	67	4.4

TABLE 5.1 – Corpus of the ten most identified causal sounds in terms of object or action, from (Lemaitre et al., 2010).

5.2.3 Material

To be consistent with study (Lemaitre et al., 2010), we used a very similar apparatus. The sounds were played diotically by a Macintosh Mac Book Pro workstation with a MOTU firewire 828 sound card. A pair of YAMAHA MSP5 loudspeakers were used to amplify the sounds. Participants were seated in a double-walled IAC sound-isolation booth. The sounds were played and the study was run using Cycling'74's Max/MSP. The ecological adjustment of sound levels with a *sound level meter* Cyrrus.

5.2.4 Procedure

The goal of the experiment is to analyze if listeners can identify the cause of a sound after the transformation described above. To investigate the uncertainty in the cause of identification, a common tool is to measure the index of causal uncertainty H_{cu}. The procedure of measuring H_{cu} is very time-consuming and needs for a precise semantic analysis of verbalizations. Instead in (Lemaitre et al., 2010), the authors proposed to measure the confidence in the identification using a scale between 1 and 5, reported in table 5.2.

Statement	Scale
"I don't know at all"	**1**
"I am really not sure"	**2**
"I hesitate between several causes"	**3**
"I am almost sure"	**4**
"I perfectly identify the cause of the sound"	**5**

TABLE 5.2 – Scale for rating the participants' confidence in identifying the sound cause. The scale is reported both in english and french

They show that the resulting measure is correlated to H_{cu} even if both measures do not provide exactly the same information. While the H_{cu} evaluates the identification of causality, the rating measures the confidence in identification of the sound causal action.

The twenty one non-musician candidates were split into two groups : one of ten and another of eleven. Each group was randomly assigned to one of the corpora. At the beginning, the participants were told that they would have to listen to sounds corresponding to several actions performed in the context of a kitchen. The participants sat in the sound-isolation booth and read the instructions alone. The participants were asked to rate their confidence in identifying the sound cause in terms of its action for each sound. Each sound was judged twice (test and retest) and the participants were told that they may listen to the sounds several times.

For each participant, the order of sounds in the sequence was randomized. The experiment was divided into three steps. First, the participants were provided with two examples of sounds to get used to the interface and the nature of sounds. Second, they heard sequentially the ten remaining sounds and had to rate on a scale from 1 to 5 their confidence in identifying the cause.

Only the assertions were displayed on the screen together with black buttons. When they rated one sound, they could validate by pushing a button and switch to the next sound. Third, the participants had to answer a short interview on the experiment. The scores were stored as a column of integers between 1 and 5. A second file associated to the scores stored the corresponding sound order.

5.2.5 Results

The consistency of the participants' answers are tested by the test/retest scores. Regarding the resulting scores for the sounds from the non-transformed corpus. The confidence scores remained the same for 64.6% of the sounds (STD$= 14.3\%$) and had less than 1 category of difference for 94.6% (STD$= 6.9\%$). Regarding the resulting scores for the sounds from the transformed corpus. The score between the test and the retest remained the same for 56.0% (STD$= 17.1\%$) and was different with less than 1 category for 89.0% (STD$= 9.9\%$). Even if the scores were not the same between the test and the retest for half of the sounds, they remained close (less than 1 category). The mean values (between test and retest) could be computed resulting to one score per participant per sound. The following analysis was conducted using them.

The confidence values were submitted to a repeated-measure analysis of variance (ANOVA) with one within-subject factor (the ten sounds) and one between-subject factor (the two corpora). The analysis revealed that the participants are globally significantly less confident in identifying the sound cause in terms of action while doing the listening test with the transformed corpus rather than the non-transformed one ($F(1,19) = 14.64, p < 0.01$). The analysis also revealed that the confidence in the cause identification depends on the sound ($F(9,171) = 8.22, p < 0.01$). Finally, an interaction existed between the conditions non-transformed and transform and the sound considered ($F(9,171) = 4.34, p < 0.01$) meaning

that the confidence in identification depends on both the sound and if it is transformed or not. Figure 5.1 reports the mean values for each sound and both transformations. It illustrates that globally, the confidence scores are lower for the transformed sounds (except from sounds 1 and 2).

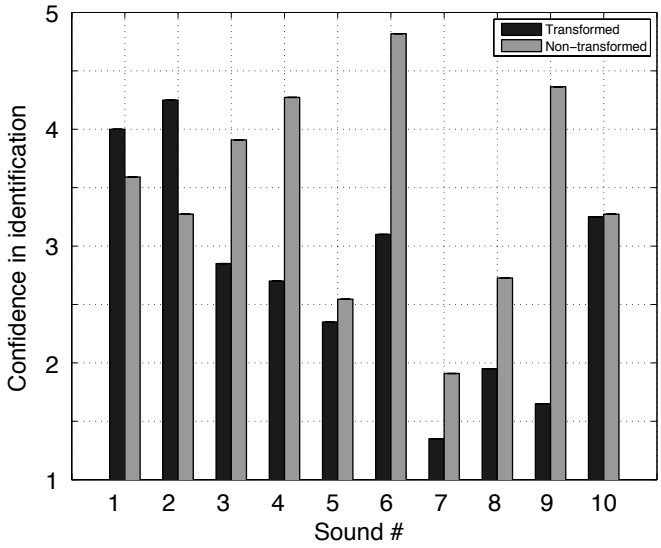

FIGURE 5.1 – Experiment 1 scores. Confidence in sound identification in terms of action was rated by participants. Means are depicted on the figure for both non-transformed and non-transformed sounds.

For each sound, we further tested if the difference in participants' confidence in the sound cause identification is significant between the transformed and non-transformed versions of this sound. To do so, we apply a t-test with a Boneferroni correction leading to a global significance level of $\alpha = 0.005$. The analysis led to the discrimination of three sounds from each corpus among the initial set of ten. Theses sounds are 4, 6 and 9, whose descriptions are reported in table 5.3.

Id.	Sound description	t value ($df = 19$)
4	Champagne cup shocked	2.88
6	Pouring cereals into a bowl	3.49
9	Crushing a metallic can	4.89

TABLE 5.3 – Sounds selected after analysis. Three sounds are significantly discriminated ($\alpha = 0.005$) after applying the audio transformation.

5.2.6 Discussion

The first experiments allowed us to build one subset from each corpus considered. Each subset is made of three sounds indexed as 4, 6 and 9. The first subset contains the original three sounds, the second subset the transformed versions of these three sounds. In terms of temporal profile, the sound 4 is an impact, the sound 6 has a continuous profile without impact and the last is an irregular sound with several articulated impacts. Profiles of the original sounds can be seen on figure 5.2.

The contrasts in score between sounds from both corpora are low : only three sounds have been discriminated from the listening test. This can be explained by the fact that both corpora have been evaluated independently by their respective group of participants. Hence, each set of sounds has been rated according to a confidence scale related to intra-group differences instead of inter-group differences. Merging together both corpora into a single corpus and

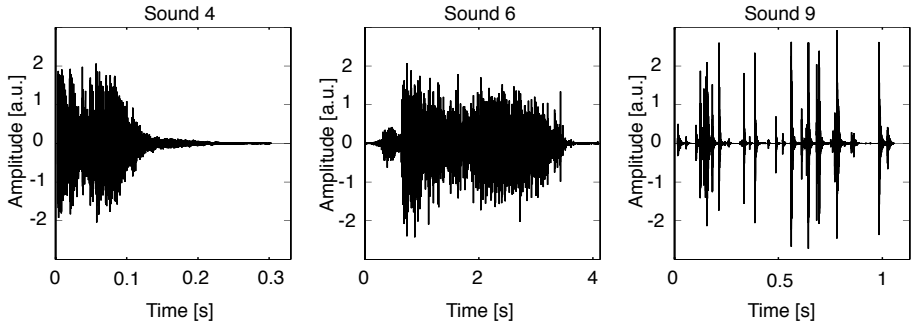

FIGURE 5.2 – Waveforms of sounds 4, 6 and 9

evaluating the confidence in identification with a similar listening test applied to the resulting corpus, would lead to a better contrast between transformed and non-transformed sounds. In this case, asking for participants to rate their confidence in identifying the sound cause will be based on their relative causality. Hence, it would not be possible to interpret which non-transformed sounds have an identifiable cause and which transformed sounds have not an identifiable cause. In the following, sounds from the subset of non-transformed sounds will be called *action related* sounds and sounds from the subset of transformed sounds will be called *non-action related* sounds.

5.3 Experiment 2 : Gesture responses

In the previous experiment, sound corpora have been validated by participants in a listening test as action and non-action related sound corpora. In this section we present an experiment that aims at analyzing how gesture response to a sound stimulus is affected according to the level of identification of the sound stimuli causes in terms of action.

5.3.1 Participants

Twenty-two non-musician participants (11 women and 11 men) were recruited outside of the institute for this study which took place at Ircam – Centre Pompidou in Paris. The participants were not those that have passed the first study presented before (section 5.2). All but three participants were right-handed. None of them reported having hearing problems. The study was carried out in accordance with the Declaration of Helsinki. All participants gave informed consent to participate in the study. The experiment took approximately one hour, and the participation was retributed with a nominal fee.

5.3.2 Stimuli

The stimuli are two corpora corresponding to the sound subsets created in the first experiment. A first corpus contained the original action related sounds selected (sounds 4, 6 and 9). A second corpus contained the transformed versions of the sounds 4, 6 and 9. As before, the first corpus will be called *action related corpus* while the second one *non-action related corpus*.

To compare the implication of causality on gestural responses of sound stimuli, we chose to add an additional sound that was not discriminated by the audio transformation. Since no formal criterion existed for the selection process, we chose a sound that had a distinct acoustic profile. Among the non-discriminated sounds in experiment 1, we chose to add sound 8 "*Removing a cork stopper*". Overall, the four original sounds 4, 6, 9 and 8 belonged to the first corpus and the transformed versions of them to the second. In the following they are re-indexed by

1, 2, 3 and 4. Their mean duration is 2.2s (STD=1.9s). All the sounds from both corpora were monophonic and had 16-bit resolution and a sampling rate of 44.1kHz.

5.3.3 Material

The hand's position was captured by tracking on-hand placed markers with an ARTtrack 3D video infra-red motion capture system at 100Hz sample rate. The same loudspeakers than in experiment 1 were used : a pair of YAMAHA MSP5. A video camera recorded each performance. Two MIDI pedals as well as the MIDI controller Berhinger BCF2000 were used for triggering the sounds. Motion, audio and video were recorded synchronously at each trial using the real time programming environment Max/MSP. Participants were standing in a studio at Ircam in front of a screen displaying a visual countdown that marked the sound beginning (detailed in the next section). The figure 5.3 summarizes the whole experiment set up.

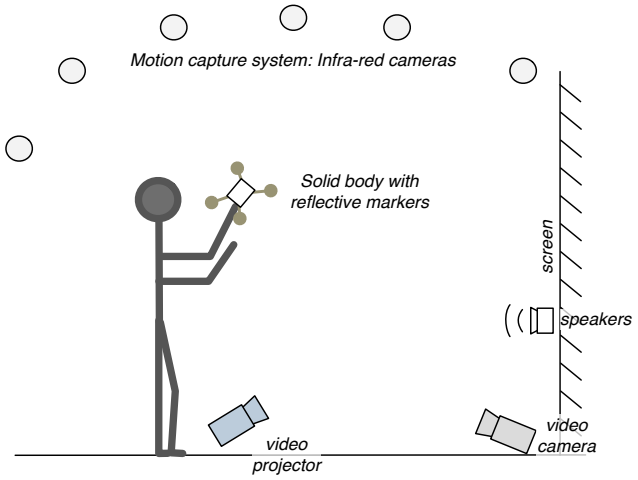

FIGURE 5.3 – Simplified set up used in the experiment.

Finally, we employed the same procedure for ecological adjustment of sound levels.

5.3.4 Design and procedure

The experiment followed a between-participants design. The participants were splitted into two groups : action-related sounds were presented to 11 participants and non-action related sounds to the remaining 11 participants. Participants were assigned to one corpus randomly. For each participant from both corpora, the sequence of sounds were randomized.

The participant stood up in front of the screen. At the beginning, the participants were told that the experiment contained two phases. During the first phase, participants were asked to perform a gesture associated and synchronous to the sound they will hear. Each performance was video recorded to be used in the second phase. Here "associated" means performing a gesture that mimics the action causing the sound or that traces the temporal acoustic evolution of the sound. Two examples for the two distinct strategies that can be adopted in the performance were given by the experimenter. The strategies are explicitly told to the participants to avoid participants to be lost when being faced to such a non-usual experience. Then, the global setting was explained. The participant was placed in front of a large-scale screen and two MIDI pedals. Each MIDI pedal is used to trigger the current sound. Pushing down a MIDI pedal started a 2-seconds countdown displayed on the screen. The participants were told that the sound is played when the countdown reached 0. Pushing down a MIDI pedal also created a new track used to stored the video (of the participant's gesture performance),

the audio played, and the gesture. Hence the recording started when the pedal is pushed (2 seconds before the sound onset) and continued until 0.7 second after the end of the sound. For each sound of the corpus, there were three sequential steps : *training*, *selecting*, *validating*. The left MIDI pedal was used for the steps *training*, *validating*, the right MIDI pedal for *selecting*. In the first step (*training*), the participant could listen to the sound any number of times by pushing the left pedal. Synchronously, any number of rehearsals could be performed in order to find the gesture that was, for the participant, well associated to the sound. All the trials were recorded. When the candidates felt confident, they had to *select* the gesture by doing the same pushing the right pedal. The selected gesture is called *candidate gesture*. The final step was the validation of the candidate gesture. The participant must perform three times exactly the same gesture. This step validated that the candidate gesture is stabilized.

During the second phase, participants sat on a chair in front of the screen. They were told that they will watch the video of their *candidate gesture*, sound after sound, and for each sound the experimenter will ask some questions helping the participant to comment their actions. The interview was conducted following the interviewing techniques defined in (Vermersch, 1990) and practically used in (Tardieu et al., 2009). The technique aimed at helping the interviewee to describe the sounds and their actions engaged by an open dialogue preparing restarts when the participant encountered difficulties in the verbalized description. Together with the participants, the videos were sequentially visualized for each sound. Only the candidate gestures were watched. Overall, a total of four videos were visualized. For each video a dialogue was opened according three pre-defined topics. First, we discussed what came spontaneously to their mind when they first listened to the sound. Second, we discussed the gesture they performed, for example *Was it difficult to find the gesture ? What are the different steps in your gesture ?* etc.. Finally, we discussed the relationships between the performed gesture and the corresponding sound, for example *Did you try to be synchronous ?* etc. The interviews were recorded externally to be further analyzed as explained in the next section 5.3.5.

5.3.5 Data Analysis

The experiment dealt with two types of data involving two distinct types of analysis. On the one hand, we collected interviews. On the other hand, we recorded 3-dimensional gesture data from the motion capture system.

Regarding to the interviews, a general grid for the analysis was designed to be filled by the verbal description (verbs, nouns, adjectives) given by the participants. It consisted in two main categories : *causal level*, *acoustic level*. The causal level referred to the identified sound cause while the acoustic level referred to the sound quality. Both levels were divided into subcategories as follows :

- causal level : *action*, *object* (that produced the sound)
- acoustic level : *temporal description*, *profile*, *timbre*

The decomposition was inspired by (Lemaitre et al., 2010). The interviews was analyzed by three independent experts that filled the grid of analysis while listening to the interviews. Two of them are co-author of this paper and we asked on purpose an other expert who was not aware of the goal of the study. The three independent experts have listened independently all the recorded interviews. The resulting analysis were reported in a general table for both corpora leading to a *portrait* of each sound. The other data captured were each performance's gesture data. This defined the first set of analysis. The second was focused on gesture data analysis captured by the motion capture infra-red system.

Regarding to the gesture data, each gesture trial is a 3-dimensional time series corresponding to the positions along the three axis (x, y, z). From the whole set of gesture data collected, we kept each participant's *candidate* gesture. We computed the norm of the first order derivatives using functional data analysis techniques (Ramsay and Silverman, 1997). Velocity profiles

interpret the kinetic energy that participants employed during their performance. As shown in previous studied, this dynamical feature hold relevant information on gesture (Leman, 2007; Caramiaux et al., 2010c). We used a B-spline basis of order 6 to fit the discrete gesture data. One spline is placed at each discrete sample. A roughness penalty is applied on the fourth derivative with the penalty coefficient $\lambda = 1e2$. This allowed for the accurate computation of smooth first and second order derivatives (Ramsay and Silverman, 1997). Note that Loehr et al. (Loehr and Palmer, 2007) used a similar analysis of the discrete data of pianists' movements. Finally, each participant's *candidate* gesture is segmented according to three gesture parts corresponding to the *preparation*, the *stroke* and the *release* (Kendon, 2004). These parts are defined as follows :

- preparation : gesture occurring before the stroke, used for preparing the movement as bringing the arms in a specific position.
- stroke : gesture synchronous to the sound (e.g. producing the sound).
- release : gesture release after the stroke, occurring after the sound's end, e.g. relaxing the arms on both sides of the body.

Note that some gestures could not have a *release* part. For this purpose, in the analysis we will only use both preparation and stroke gestures. Note also that, in this paper, stroke does not solely refer to impact but more generally to gesture part intentionally synchronous to the sound.

5.3.6 Results

We remind the reader that in this section, sounds 1, 2, 3 and 4 refer to sounds previously indexed by 4 (champagne cup shocked), 6 (pouring cereals into a bowl), 9 (crushing a metallic can) and 8 (removing a cork stopper).

Analyzing the Interviews

Action related sounds. The first step is the analysis of the sounds at a causal level. The participants used verbs to characterize the action causing the sound and they mostly used the same verb for sounds $1, 2$ and 3 : *to hit* (sound 1) ; *to pour* (sound 2) ; *to crush* (sound 3). These verbs were the ones used in (Lemaitre et al., 2010) to describe the original sounds from the kitchen (see table 5.1). Sound 4 did not show a consensus in the description of the action. The verbs used by the participants were *to force*, *to rub*, *to pull*, or *to scrape*, showing a higher variability in the description. Regarding to the object description, no consensus could be found. Sound 1 referred to either one object *glass, bell, jar* or two objects hitting *glass–glass*, *knife–glass*. Sound 2 referred either to several small objects by *grains, marbles, peas, coins* or an interaction between small objects and a static one *stones–floor*, *rice–bowl*. Sound 3 was described in terms of a single object : a consensus was found for a *metal can*. Finally, sound 4 referred to objects from distinct lexical fields. Isolated objects are described by a *cork stopper*, a *zip*, *rack-and-pinion*. Interaction between objects are *stick–box with sprockets*, *ruler–table*.

The second step is the analysis of the sounds at an acoustic level. Temporal description of sounds are minimal : *short* (sound 1) ; *long* (sound 2) ; *discontinuous* (sound 3). Sound 4 had no temporal description. The profiles were described as : *changes in loudness* (sound 2) ; *irregular* (sound 3) ; *continuous* (sound 4). The timbre was not described by the participants, only a rough description of sound 1's timbre as *high–pitched* was asserted.

Finally, the gestures were described by the action : "I was cheering with a glass in my hand" (sound 1) ; "I am pouring grains of something" (sound 2) ; "I am crushing a can" (sound 3). Gesture related to sound 4 is described as "pulling", "rubbing".

Non-action related sounds. First, we analyzed the description at a causal level. Semantic description of transformed sounds was metaphorical in the sense that no precise actions or objects

were emphasized but rather qualitative description of conscious representation of the sound evocation. Sound 1 was characterized by an object that *is falling*, *is hitting*. This object was described differently according to the participants but referred to the metaphor, for example *a box*, *a javelin*, *a knife* or *a ball* (petanque) and could be in interaction : *box–tiled floor*, *hammer–sheet metal*. Sound 2 was described as *waves* or more globally the *sea*, *water* that was going *back and forth*, *up and down*. The lexical field used is related to the sea. Sound 3 showed a higher variability between descriptions. Participants described it by *something* that *is walking*, *is hurtling down*. Several "objects" are reported like *a horse, book pages, rope [lasso], a whip*. Finally, sound 4 was described as either *a geyser, a wagon, a cupboard* or *a flame* that accomplished the action of *being sucked up, being teared up*.

The acoustic level description showed more detailed features compared to non-transformed sounds. Indeed temporal description was described as follows. Sound 1 was *brief*. Sound 2 was decomposed into three distinct parts : the beginning (*louder, sudden, brutal*), the middle (*jolt*), the end (*slower, descrescendo, quiet*). Sound 3 temporal evolution was qualified as *Rhythmic, jerky*. Sound 4 was *linear* and decomposed into two distinct parts : the beginning (*open, intense*), the end (*abrupt, is closing*). The second sub-category was the sound profile. Sound 1 was *brutal*. Sound 2 had a profile that *goes up and goes down*, with *peaks of intensity* and *irregular, recurrent*. Sound 3 had an *irregular* profile. Finally, sound 4 had a *crescendo/descrescendo* profile (i.e. trajectories of the loudness). The timbre was more verbalized for sounds 3 and 4 than sound 1. Sound 1 timbre was *high–pitched* (same description as in the non-transformed case). Sound 3 had a timbre that was *wide* and *diffuse* while sound 4 had a timbre that was a *switching between high-pitched and low-pitched*.

Gestures are described as follows. Gesture associated to sound 1 is brief and precise and act the impact on an object or by an object. Gesture associated to the sound 2 is waving, is *mimicking this object* (in this case the *wave*) and is describing as drawing the sound profile. Gesture associated to sound 3 is repetitive and *follows the sound*. Finally, gesture associated to sound 4 is also described as *mimicking the object* and temporally *following* the sound.

Variability of velocity trajectories

The hand gesture data were analyzed in terms of the norm of velocity from the 3-dimensional positions coordinates. Figure 5.4 illustrates all the performances for each sound from each corpus. Each plot represents from top to bottom : The waveform for the action related sound i ; The *candidate* gestures associated to the action-related sound i by all the participants : upper bound is the third quartile limit, lower bound is the first quartile limit and the curve is the median evolution ; The corresponding non-action related sound i ; The *candidate* gestures associated to the non-action related sound i by all the participants.

Each sound has 11 velocity profiles associated that are the *candidate* gestures of all participants. Normalizing by the standard deviation, all gesture signals are of variance 1. We aimed at analyzing the effect of the corpora (i.e. the audio transformation) and the sounds themselves on the gesture variability between participants. The distance measure between gestures used in this paper is the Dynamic Time Warping (DTW) (see 5.6.1 for the technical description of the method). DTW measures the cost to time stretch one given signal to another of different length. This distance allows for considering variation in time (time stretching) and in amplitude since the DTW relies on a distance measure between samples. Hence for each pair of gesture signals we have a real value between 0 and ∞ that illustrates how close these two signals are. In this paper, each gesture considered is described as the 1-dimensional velocity profile. The rationale behind this choice is based on our previous works that highlighted the importance of dynamic and kinetic energy in free hand gestures (Caramiaux et al., 2010b; Caramiaux et al., 2010c; Nymoen et al., 2011)).

The testing procedure was as follows : one velocity profile represents the participant's can-

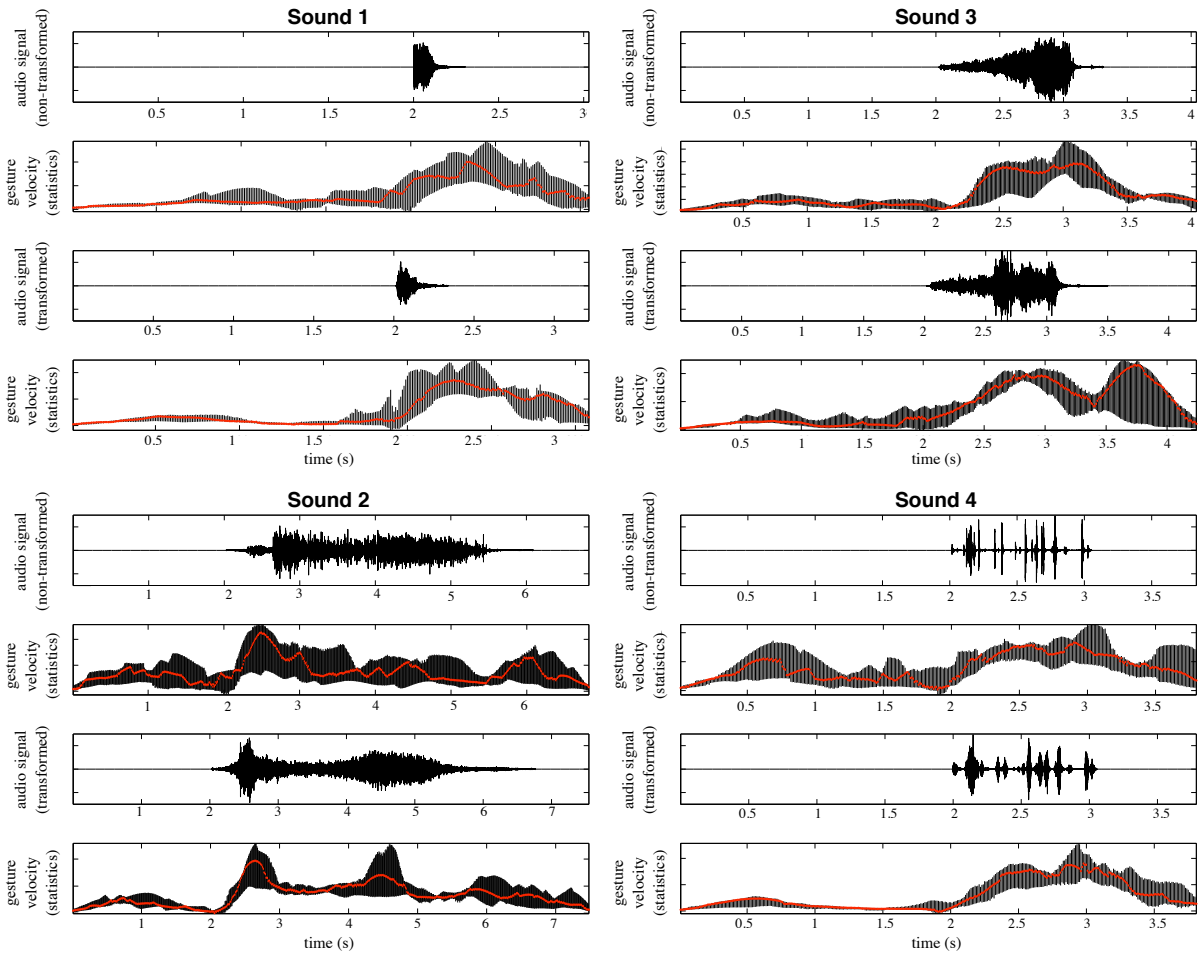

FIGURE 5.4 – Gestures' velocity associated to each sound from each corpus. Each plot represents from top to bottom : The waveform for the action related sound i ; The *candidate* gestures associated to the action related sound i by all the participants : upper bound is the third quartile limit, lower bound is the first quartile limit and the curve is the median evolution ; The corresponding non-action related sound i ; The *candidate* gestures associated to the non-action related sound i by all the participants.

didate gesture. Each measure computed the distance between two profiles leading to the need to compute the distance between all combinations of two profiles among the 11 available per corpus. However, since the distances considered were symmetric, it was not necessary to compute the 11^2 combinations. The number could be reduced to $(11 \times (11-1))/2 = 55$ measure values per sound.

We examined the preparation gestures. A repeated-measure ANOVA with one within-subject factor (the four sounds) and one between-subject factor (the two corpora) is computed on the warping distances between participants' gestures. It resulted that no significant effect of the corpus (in other words, the audio transformation performed on sounds) was found but there was a significant effect of the sound considered ($F(3, 324) = 12.88, p < 0.01$).

In the same way, we examined the strokes. The warping distance values were submitted to a repeated-measure ANOVA with one within-subject factor (the four sounds) and one between-subject factor (the two corpora). The analysis revealed that the treatment (change of corpus) affected significantly the warping distances ($F(1, 108) = 36.67, p < 0.01$) as well as the sounds ($F(3, 324) = 38.95, p < 0.01$). The distances are significatively lower for the non-action related sounds than action-related sounds, as shown with figure 5.5 (left). There is interaction between the treatment and the sounds ($F(3, 324 = 13.82, p < 0.01)$) meaning that the effect of the treatment on distance values were not equivalent over the sounds. We examined these

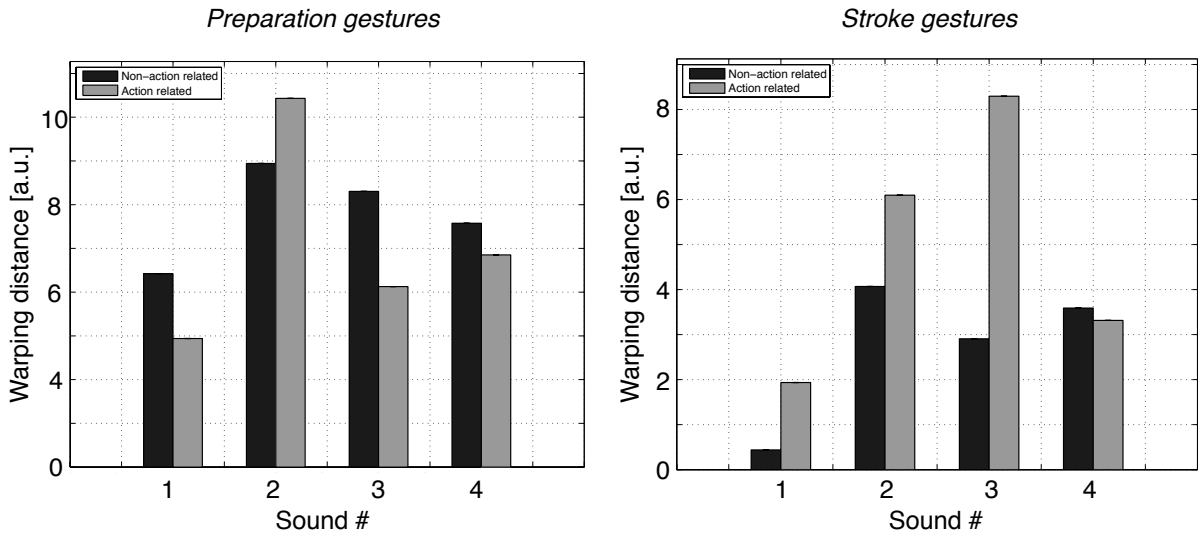

FIGURE 5.5 – Mean results of the warping distances for both the preparation gestures (left) and the stroke gestures (right). Means are reported for each sound (columns) and each corpus (black–gray).

differences in further details using using a Student's t'test with a Boneferroni correction : the significance level is set to $\alpha = 0.01$. The analysis revealed that the treatment is significant for sounds 1, 2 and 3 but not for sound 4. For the three first sounds, the variability is then lower for non-action related sounds than for action related sounds.

5.4 Discussion

Qualitative verbalization

Description at a causal level of action-related sounds selected in the first experiment (sounds 1, 2, 3) is in terms of action that has produced the sound rather than the object producing the sound : verbalization of the action was consensual and participants used *verbs* to describe the sound in terms of action ; in comparison, the description of the object related to the sound was fuzzier. The added sound 4 (corresponded to sound 8 on figure 5.1) did not fulfil this conclusion, coherently with the listening test presented in section 5.2 (and the one by Lemaitre et al. (Lemaitre et al., 2010)). Indeed, the listening test showed no significant difference between the mean confidences in identification between the original and the transformed sound. In addition, action related sounds were not described at the acoustic level. Here, we found a similar outcome than in (Lemaitre et al., 2010) where (non-musician) participants focused more on sound identification in terms of action (or object) rather than acoustic qualities (also coherent with (Marcell et al., 2000)). Participants made use of *everyday listening* strategy (Gaver, 1993b; Gaver, 1993a) because the sound were easily identified. Gestures were also described in terms of action by using the same terminology as for sounds creating a direct link between gesture and sound : the gesture is the action that produced the sound, hence action was anticipating the sound.

Non-action related sounds were described at an acoustic level by temporal description and profile characteristics. The causal level was subjective and metaphorical description of sounds in terms of action and object. Action was not described by infinitive verbs, such as **to do** something, but rather a qualitative form indicating something that **is doing** the action. The emphasis was translated from the action itself to the object or event that is doing such action. In addition, terminology used to describe the object was highly variable. Participants changed of lexical fields highlighting subjective perception : metaphor of the sound. These metaphors see-

med to repose on perception of the sound acoustic qualities. For example the metaphor of the wave for the transformed sound 2 refer to the continuous and oscillating profile of the sound's loudness. However, these acoustic qualities were more detailed than for action-related sounds. They verbalized the temporal description, highlighting different parts in sounds 2 and 4, and the profiles. We assume therefore that participants made use of another listening strategy that we refer to the *musical listening* strategy (Gaver, 1993b; Gaver, 1993a).

Gestures associated to non-action related sounds were described by object mimicking where *object* did not refer to the causal level but should be rather linked to *sound object* defined by Schaeffer (Schaeffer, 1966). Hence mimicking sound object was precised by terms like *following* or *drawing* meaning mimicking the shape of the object. This outcome supports previous works (Godøy, 2006; Godøy et al., 2006b; Nymoen et al., 2011). First, Godøy (Godøy, 2006) postulates the link between sound morphologies (temporal profiles) and a gestural equivalency. This was then experimented in (Godøy et al., 2006b) and precised by Nymoen et al. (Nymoen et al., 2011). In this latter work, synthesized sound based on audio feature evolution served as stimuli for an experiment where people were aksed to move spontaneously in response to the sound. The authors made use of a cross-modal analysis tool (firstly used in our previous work (Caramiaux et al., 2010c)) to extract the relationship between gesture kinematic features evolution and audio descriptors evolution, validating a *tracing* link between sound stimuli and gesture responses.

To conclude on interviews analysis, they validated our hypothesis : action related sounds were perceived as an action causing the sound while non-action related sounds were perceived relying solely on acoustic features. In addition, gestures were either mimicking the action causing the sound or the *sound object* (in the sense of Schaeffer (Schaeffer, 1966)) by drawing, tracing or following its acoustic profile. With the sounds considered, it seems that occulting high frequencies on audio signals and retaining the low-frequency part allows for taking off the causal action and highlighting the intrinsic *sound object*.

Quantitative gesture data analysis

The results from the analysis of inter-participants stroke variability showed that strokes were more consistent across participants when referred to non-action related sounds than action related sounds. Indeed the variability (in terms of warping distance) is lower after treatment for sounds $1, 2$ and 3. Sound 4 did not induce difference in variability for both measure. This confirmed that if the audio transformation did not discriminate between action and non-action related sounds through a listening test, the gesture responses to these sounds could not be differentiated based on a consistency measured across participants. Our interpretation of variations between gesture responses for both corpora is as follows. Participants mimic the action causing the sound when the action can be identified. This leads to variations due to subjective strategies (linked to habits or to the absence of a physical object in hand) for executing the action. Even if they identified the same causal action, the references are distinct and subjective. We can relate this type of gesture with the *iconic gestures* postulate by McNeil in (McNeill, 1996) defined as the non-verbal elements used in communication to represent an object or an concrete action.

Otherwise, participants mimic the *sound object* (terminology used in the sense of specified by the interviews) leading to a common reference that is the sound itself. The participants relied on the temporal evolution of the sound to perform the gesture. As discuss previously, they are often derived from a metaphorical image of the stimulus. In that sense, they can be related to the *metaphorical gestures* defined by McNeil in (McNeill, 1996) as the gestures linked to an abstract idea.

We then inspected if the same result is still suitable for preparation gestures. Firstly, one can observed more preparation (gesture velocity profiles varied more) when the sound cause

is identified as an action. In the results, no significant differences arised between the sounds after and before treatment. It means that participants consistently prepare the stroke gesture for mimicking action related sounds. On the other hand, participants were consistent in not preparing the stroke gesture for non-action related sounds. The variability however significantly changed between sounds, and we could observe more variability for longer sounds (even if the countdown was set at 2 seconds for each sound) : sound $2, 3$ and 4 gave rise to higher variability than for sound 1 and sound 2 induced higher variability than sounds 3 and 4.

5.5 Conclusion

In this paper, we presented an experimental study on gesture responses to sound stimuli. The sound stimuli were environmental sounds where the action that have produced the sound is clearly identified and transformed versions of the same sounds where the sound source could not be identified anymore. We aimed at validating the following hypothesis : gesture associated to action related sound mimics the action causing the sound while gesture associated to non-action related sound follows the sound acoustic contours. In the paper, we have first built two sound corpora by means of a proposed audio transformation validated by a listening test. The main audio feature retained in the transformed sounds was the loudness. Then we proposed an experiment where participants were asked to perform gestures associated to sounds from one of the two corpora and to verbalize their performance. We showed that verbalization from the interviews validated our hypothesis and introduced the concept of mimicking both the action or the *sound object*. In addition, gesture data showed lower variations between strokes (gestures linked to the sound) for non-action related sounds than action related sounds. Preparation gestures showed consistency across transformed sounds that is not retrieved for the original sounds. This is supported by a better matching between velocity profiles and loudness profiles for non-action related sounds. Our interpretation is that sound causality as action is represented by an *iconic gesture* (McNeill, 1996) that can be performed under distinct forms (depending on the participants' habits in doing the action). The participants have distinct *references*. On the other hand, when the sound cause is not (or difficult to) identify, the participants perform a *metaphoric gesture* (McNeill, 1996) that follow the acoustic energy contour of the sound and have a common reference that is the sound itself.

Future works will be dedicated to inspect the variations of the gestures performed on isolated sounds (as presented in this paper) when executed in sequence. We aim at proposing models that learn these gestures and can extracting them from sequences by modeling coarticulation aspects.

5.6 Annex

5.6.1 Dynamic Time Warping (DTW)

The DTW performed a temporal alignment between two curves (here we considered two 1-dimensional velocity curves). The DTW is based on the Euclidean distance. Consider two velocity profiles, denoted $\mathbf{v}_1$ and $\mathbf{v}_2$, or respective lengths L_1 and L_2. The method first computes a cost matrix $C(v_1, v_2)$ such as :

$$C(v_1, v_2) = (c(v_1, v_2)_{ij})_{1 \leq i \leq L_1, 1 \leq j \leq L_2}$$

Where

$$c(v_1, v_2)_{ij} = \sqrt{\sum_{i=1}^{L_1} \sum_{j=1}^{L_2} (v_{i,1} - v_{j,2})^2}$$

Where $v_{i,1}$ is the i-th element of the mulitdimensional vector $\mathbf{v}_j$. Then a dynamic programming based algorithm find the optimal path (i.e. minimizing the cumulative cost) in the matrix $C(v_1, v_2)$ leading to the temporal alignment between both velocity profiles. The cumulative sum's end point is the global cost and was used as the measure value.

Chapitre 6

Analyzing Sound Tracings – A multimodal approach to music information retrieval

K. Nymoen[1], **B. Caramiaux**[2], **M. Kozak**[3], **J. Torresen**[1]

[1] University of Oslo, Department of Informatics, 0316 Oslo, Norway
[2] UMR IRCAM-CNRS, Paris, France
[3] University of Chicago, Department of Music, Chicago, IL 60637, USA

Abstract : This paper investigates differences in the gestures people relate to *pitched* and *non-pitched* sounds respectively. An experiment has been carried out where participants were asked to move a rod in the air, pretending that moving it would create the sound they heard. By applying and interpreting the results from Canonical Correlation Analysis we are able to determine both simple and more complex correspondences between features of motion and features of sound in our data set. Particularly, the presence of a distinct pitch seems to influence how people relate gesture to sound. This identification of salient relationships between sounds and gestures contributes as a multi-modal approach to music information retrieval.

Keywords : Sound Tracing, Cross-Modal Analysis, Canonical Correlation Analysis

6.1 Introduction

In recent years, numerous studies have shown that gesture, understood here as voluntary movement of the body produced toward some kind of communicative goal, is an important element of music production and perception. In the case of the former, movement is necessary in performance on acoustic instruments, and is increasingly becoming an important component in the development of new electronic musical interfaces (Van Nort, 2009). As regards the latter, movement synchronized with sound has been found to be a universal feature of musical interactions across time and culture (Nettl, 2000). Research has shown both that the auditory and motor regions of the brain are connected at a neural level, and that listening to musical sounds spontaneously activates regions responsible for the planning and execution of movement, regardless of whether or not these movements are eventually carried out (Chen et al., 2008).

Altogether, this evidence points to an intimate link between sound and gesture in human perception, cognition, and behavior, and highlights that our musical behavior is inherently multimodal. To explain this connection, Godøy (Godøy, 2006) has hypothesized the existence of *sonic-gestural objects*, or mental constructs in which auditory and motion elements are correlated in the mind of the listener. Indeed, various experiments have shown that there are correlations between sound characteristics and corresponding motion features.

Godøy et al. (Godøy et al., 2006b) analyzed how the morphology of sonic objects was reflected in sketches people made on a digital tablet. These sketches were referred to as *sound tracings*. In the present paper, we adopt this term and expand it to mean a recording of free-air movement imitating the perceptual qualities of a sound. The data from Godøy's experiments was analyzed qualitatively, with a focus on the causality of sound as impulsive, continuous, or iterative, and showed supporting results for the hypothesis of gestural-sonic objects.

Godøy and Jensenius (Godøy and Jensenius, 2009) have suggested that body movement could serve as a link between musical score, the acoustic signal and aesthetic perspectives on music, and that body movement could be utilized in search and retrieval of music. For this to be possible, it is essential to identify pertinent motion signal descriptors and their relationship to audio signal descriptors. Several researchers have investigated motion signals in this context. Camurri et al. (Camurri et al., 2003) found strong correlations with the quantity of motion when focusing on recognizing expressivity in the movement of dancers. Furthermore, Merer et al. (Merer et al., 2008) have studied how people labeled sounds using causal descriptors like "rotate", "move up", etc., and Eitan and Granot studied how listeners' descriptions of melodic figures in terms of how an imagined animated cartoon would move to the music (Eitan and Granot, 2006). Moreover, gesture features like acceleration and velocity have been shown to play an important role in synchronizing movement with sound (Luck and Toiviainen, 2006). Dimensionality reduction methods have also been applied, such as Principal Component Analysis, which was used by MacRitchie et al. to study pianists' gestures (MacRitchie et al., 2009).

Despite ongoing efforts to explore the exact nature of the mappings between sounds and gestures, the enduring problem has been the dearth of quantitative methods for extracting relevant features from a continuous stream of audio and motion data, and correlating elements from both while avoiding *a priori* assignment of values to either one. In this paper we will expand on one such method, presented previously by the second author (Caramiaux et al., 2010c), namely the Canonical Correlation Analysis (CCA), and report on an experiment in which this method was used to find correlations between features of sound and movement. Importantly, as we will illustrate, CCA offers the possibility of a mathematical approach for selecting and analyzing perceptually salient sonic and gestural features from a continuous stream of data, and for investigating the relationship between them.

By showing the utility of this approach in an experimental setting, our long term goals are to quantitatively examine the relationship between how we listen and how we move, and to highlight the importance of this work toward a perceptually and behaviorally based multimodal approach to music information retrieval. The study presented in the present paper contributes by investigating how people move to sounds with a controlled sound corpus, with an aim to identify one or several sound-gesture mapping strategies, particularly for pitched and non-pitched sounds.

The remainder of this paper will proceed as follows. In Section 6.2 we will present our experimental design. Section 6.3 will give an overview of our analytical methods, including a more detailed description of CCA. In Sections 6.4 and 6.5 we will present the results of our analysis and a discussion of our findings, respectively. Finally, Section 6.6 will offer a brief conclusion and directions for future work.

6.2 Experiment

We have conducted a free air sound tracing experiment to observe how people relate motion to sound. 15 subjects (11 male and 14 female) participated in the experiment. They were recruited among students and staff at the university. 8 participants had undergone some level of musical training, 7 had not. The participants were presented with short sounds, and given the task of moving a rod in the air as if they were creating the sound that they heard. Subjects

first listened to each sound two times (more if requested), then three sound tracing recordings were made to each sound using a motion capture system. The recordings were made simultaneously with sound playback after a countdown, allowing synchronization of sound and motion capture data in the analysis process.

6.2.1 Sounds

For the analysis presented in this paper, we have chosen to focus on 6 sounds that had a single, non-impulsive onset. We make our analysis with respect to the sound features *pitch*, *loudness* and *brightness*. These features are not independent from each other, but were chosen because they are related to different musical domains (melody, dynamics, and timbre, respectively) ; we thus suspected that even participants without much musical experience would be able to detect changes in all three variables, even if the changes occurred simultaneously. The features have also been shown to be pertinent in sound perception (Misdariis et al., 2010; Susini et al., 2004). Three of the sounds had a distinct pitch, with continuously rising or falling envelopes. The loudness envelopes of the sounds varied between a bell-shaped curve and a curve with a faster decay, and also with and without tremolo. Brightness envelopes of the sounds were varied in a similar manner.

The sounds were synthesized in Max/MSP, using subtractive synthesis in addition to amplitude and frequency modulation. The duration of the sounds were between 2 and 4 seconds. All sounds are available at the project website [1]

6.2.2 Motion Capture

A NaturalPoint Optitrack optical marker-based motion capture system was used to measure the position of one end of the rod. The system included 8 Flex V-100 cameras, operating at a rate of 100 frames per second. The rod was approximately 120 cm long and 4 cm in diameter, and weighed roughly 400 grams. It was equipped with 4 reflective markers in one end, and participants were instructed to hold the rod with both hands at the other end. The position of interest was defined as the geometric center of the markers. This position was streamed as OSC data over a gigabit ethernet connection to another computer, which recorded the data and controlled sound playback. Max/MSP was used to record motion capture data and the trigger point of the sound file into the same text file. This allowed good synchronization between motion capture data and sound data in the analysis process.

6.3 Analysis Method

6.3.1 Data Processing

The sound files were analyzed using the MIR toolbox for Matlab by Lartillot et al. [2] We extracted feature vectors describing *loudness*, *brightness* and *pitch*. Loudness is here simplified to the RMS energy of the sound file. Brightness is calculated as the amount of spectral energy corresponding to frequencies above 1500 Hz. Pitch is calculated based on autocorrelation. As an example, sound descriptors for a pitched sound is shown in Figure 6.1.

The position data from the OptiTrack motion capture system contained some noise ; it was therefore filtered with a sliding mean filter over 10 frames. Because of the big inertia of the rod (due to its size), the subjects did not make very abrupt or jerky motion, thus the 10 frame filter should only have the effect of removing noise.

1. http ://folk.uio.no/krisny/mirum2011
2. http ://www.jyu.fi/hum/laitokset/musiikki/en/research/ coe/materials/mirtoolbox

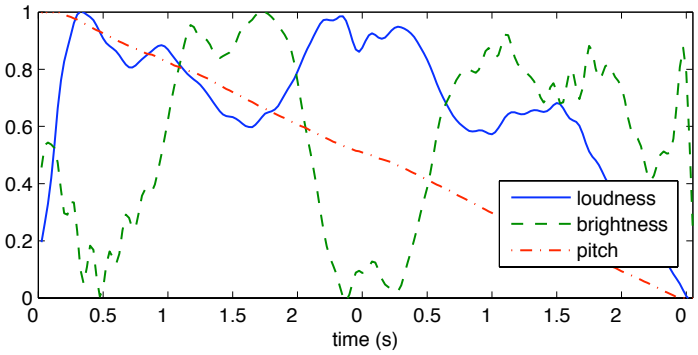

FIGURE 6.1 – Sound descriptors for a sound with falling pitch (normalized).

From the position data, we calculated the vector magnitude of the 3D velocity data, and the vector magnitude of the 3D acceleration data. These features are interpreted as the velocity independent from direction, and the acceleration independent from direction, meaning the combination of tangential and normal acceleration. Furthermore, the vertical position was used as a feature vector, since gravity and the distance to the floor act as references for axis direction and scale of this variable. The horizontal position axes, on the other hand, do not have the same type of positional reference. The subjects were not instructed in which direction to face, nor was the coordinate system of the motion capture system calibrated to have the same origin or the same direction throughout all the recording sessions, so distinguishing between the X and Y axes would be inaccurate. Hence, we calculated the mean horizontal position for each recording, and used the distance from the mean position as a one-dimensional feature describing horizontal position. All in all, this resulted in four motion features : *horizontal position*, *vertical position*, *velocity*, and *acceleration*.

6.3.2 Canonical Correlation Analysis

CCA is a common tool for investigating the linear relationships between two sets of variables in multidimensional reduction. If we let X and Y denote two datasets, CCA finds the coefficients of the linear combination of variables in X and the coefficients of the linear combination of variables from Y that are maximally correlated. The coefficients of both linear combinations are called *canonical weights* and operate as projection vectors. The projected variables are called *canonical components*. The correlation strength between canonical components is given by a correlation coefficient ρ. CCA operates similarly to Principal Component Analysis in the sense that it reduces the dimension of both datasets by returning N canonical components for both datasets where N is equal to the minimum of dimensions in X and Y. The components are usually ordered such that their respective correlation coefficient is decreasing. A more complete description of CCA can be found in (Hair et al., 1998). A preliminary study by the second author (Caramiaux et al., 2010c) has shown its pertinent use for gesture-sound cross-modal analysis.

As presented in Section 6.3.1, we describe sound by three specific audio descriptors[3] and gestures by a set of four kinematic parameters. Gesture is performed synchronously to sound playback, resulting in datasets that are inherently synchronized. The goal is to apply CCA to find the linear relationships between kinematic variables and audio descriptors. If we consider uniformly sampled datastreams, and denote $\mathbf{X}$ the set of m_1 gesture parameters ($m_1 = 4$) and $\mathbf{Y}$ the set of m_2 audio descriptors ($m_2 = 3$), CCA finds two projection matrices $\mathbf{A} = [\mathbf{a}_1 \dots \mathbf{a}_N] \in (\mathcal{R}^{m_1})^N$ and $\mathbf{B} = [\mathbf{b}_1 \dots \mathbf{b}_N] \in (\mathcal{R}^{m_2})^N$ such that $\forall h \in 1..N$, the correlation

3. As will be explained later, for non-pitched sounds we omit the *pitch* feature, leaving only two audio descriptors.

coefficients $\rho_h = correlation(\mathbf{Xa}_h, \mathbf{Yb}_h)$ are maximized and ordered such that $\rho_1 > \cdots > \rho_N$ (where $N = \min(m_1, m_2)$).

A closer look at the projection matrices allows us to interpret the mapping. The widely used interpretation methods are either by inspecting the canonical weights, or by computing the canonical loadings. In our approach, we interpret the analysis by looking at the canonical loadings. Canonical loadings measure the contribution of the original variables in the canonical components by computing the correlation between gesture parameters $\mathbf{X}$ (or audio descriptors $\mathbf{Y}$) and its corresponding canonical components $\mathbf{XA}$ (or $\mathbf{YB}$). In other words, we compute the gesture parameter loadings $\mathbf{l}^x_{i,h} = (\text{corr}(\mathbf{x}_i, \mathbf{u}_h))$ for $1 \leq i \leq m_1, 1 \leq h \leq N$ (and similarly $\mathbf{l}^y_{i,h}$ for audio descriptors). High values in $\mathbf{l}^x_{i,h}$ or $\mathbf{l}^y_{i,h}$ indicate high correlation between realizations of the i-th kinematic parameter $\mathbf{x}_i$ and the h-th canonical component $\mathbf{u}_h$. Here we mainly focused on the first loading coefficients $h = 1, 2$ that explain most of the covariance. The corresponding ρ_h is the strength of the relationship between the canonical components $\mathbf{u}_h$ and $\mathbf{v}_h$ and informs us on how relevant the interpretation of the corresponding loadings is.

The motion capture recordings in our experiment started 0.5 seconds before the sound, allowing for the capture of any preparatory motion by the subject. The CCA requires feature vectors of equal length ; accordingly, the motion features were cropped to the range between when the sound started and ended, and the sound feature vectors were upsampled to the same number of samples as the motion feature vectors.

6.4 Results

We will present the results from our analysis starting with looking at results from pitched sounds and then move on to the non-pitched sounds. The results from each sound tracing are displayed in the form of statistical analysis of all the results related to the two separate groups (pitched and non-pitched). In Figures 6.2 and 6.3, statistics are shown in box plots, displaying the median and the population between the first and third quartile. The rows in the plots show statistics for the first, second and third canonical component, respectively. The leftmost column displays the overall correlation strength for the particular canonical component (ρ_h), the middle column displays the sound feature loadings ($\mathbf{l}^y_{i,h}$), and the rightmost column displays the motion feature loadings ($\mathbf{l}^x_{i,h}$). The + marks denote examples which are considered outliers compared with the rest of the data. A high value in the leftmost column indicates that the relationship between the sound features and gesture features described by this canonical component is strong. Furthermore, high values for the sound features *loudness* (Lo), *brightness* (Br), or *pitch* (Pi), and the gesture features *horizontal position* (HP), *vertical position* (VP), *velocity* (Ve), or *acceleration* (Ac) indicates a high impact from these on the respective canonical component. This is an indication of the strength of the relationships between the sound features and motion features.

6.4.1 Pitched Sounds

The results for three sounds with distinct pitch envelopes are shown in Figure 6.2. In the top row, we see that the median overall correlation strength of the first canonical components is 0.994, the median canonical loading for *pitch* is 0.997 and for *vertical position* 0.959. This indicates a strong correlation between pitch and vertical position in almost all the sound tracings for pitched sounds. The overall correlation strength for the second canonical component (middle row) is 0.726, and this canonical function suggests a certain correlation between the sound feature *loudness* and motion features *horizontal position* and *velocity*. The high variances that exist for some of the sound and motion features may be due to two factors : If some of these are indeed strong correlations, they may be less strong than the pitch-vertical position correlation For this reason, some might be pertinent to the 2nd component while others are

pertinent to the 1st component. The second, and maybe the most plausible, reason for this is that these correlations may exist in some recordings while not in others. This is a natural consequence of the subjectivity in the experiment.

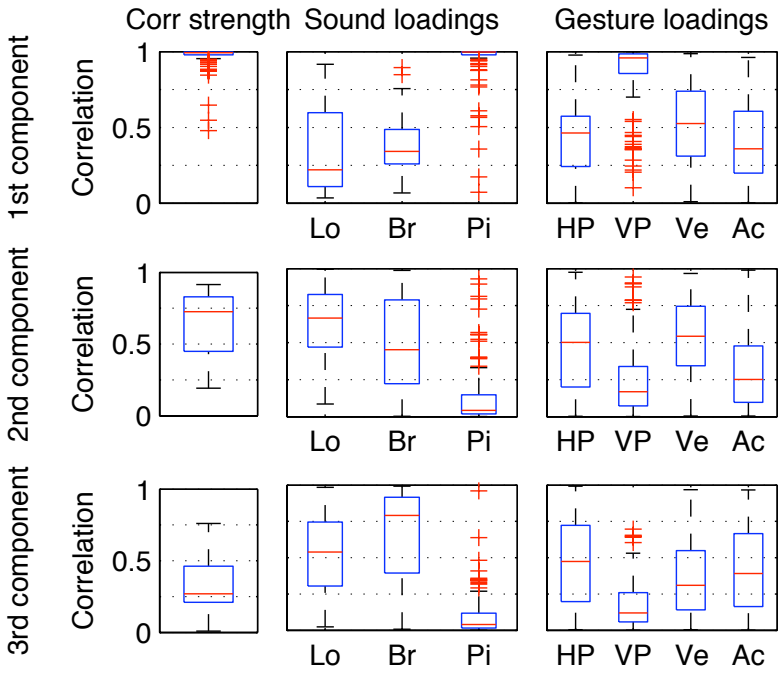

FIGURE 6.2 – Box plots of the correlation strength and canonical loadings for three pitched sounds. *Pitch* (Pi) and *vertical position* (VP) have a significantly higher impact on the first canonical component than the other parameters. This indicates a strong correlation between pitch and vertical position for pitched sounds. The remaining parameters are : *loudness* (Lo), *brightness* (Br), *horizontal position* (HP), *velocity* (Ve) and *acceleration* (Ac).

6.4.2 Non-pitched Sounds

Figure 6.3 displays the canonical loadings for three non-pitched sounds. The analysis presented in this figure was performed on the sound features loudness and brightness, disregarding pitch. With only two sound features, we are left with two canonical components. This figure shows no clear distinction between the different features, so we will need to look at this relationship in more detail to be able to find correlations between sound and motion features for these sound tracings.

For a more detailed analysis of the sounds without distinct pitch we investigated the individual sound tracings performed to non-pitched sounds. Altogether, we recorded 122 sound tracings to the non-pitched sounds ; considering the first and second canonical component of these results gives a total of 244 canonical components. We wanted to analyze only the components which show a high correlation between the sound features and motion features, and for this reason we selected the subset of the components which had an overall correlation strength (ρ) higher than the lower quartile, [4] which in this case means a value ≤ 0.927. This gave us a total of 99 components.

These 99 components all have high ρ-values, which signifies that they all describe some action-sound relationship well ; however, since the results from Figure 6.3 did not show clearly

4. The upper and lower quartiles in the figures are given by the rectangular boxes

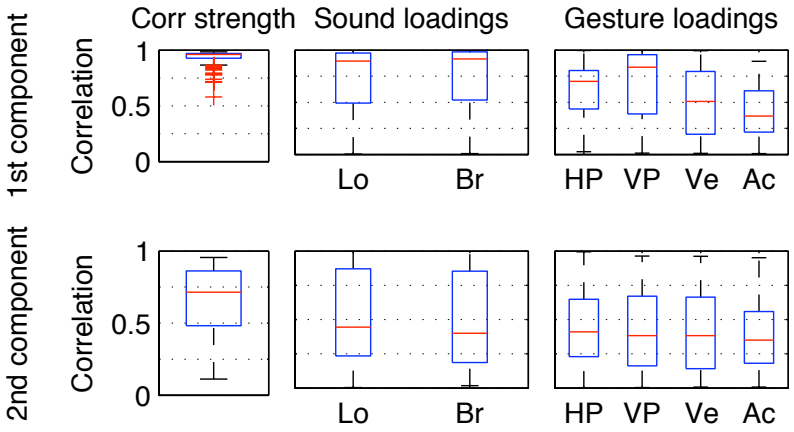

FIGURE 6.3 – Box plots of the correlation and canonical loadings for three sounds without distinct pitch.

which sound features they describe, we have analyzed the brightness and loudness loadings for all the recordings. As shown in Figure 6.4, some of these canonical components describe loudness, some describe brightness, and some describe both. We applied k-means clustering to identify the three classes which are shown by different symbols in Figure 6.4. Of the 99 canonical components, 32 describe loudness, 30 components describe brightness, and 37 components showed high loadings for both brightness and loudness.

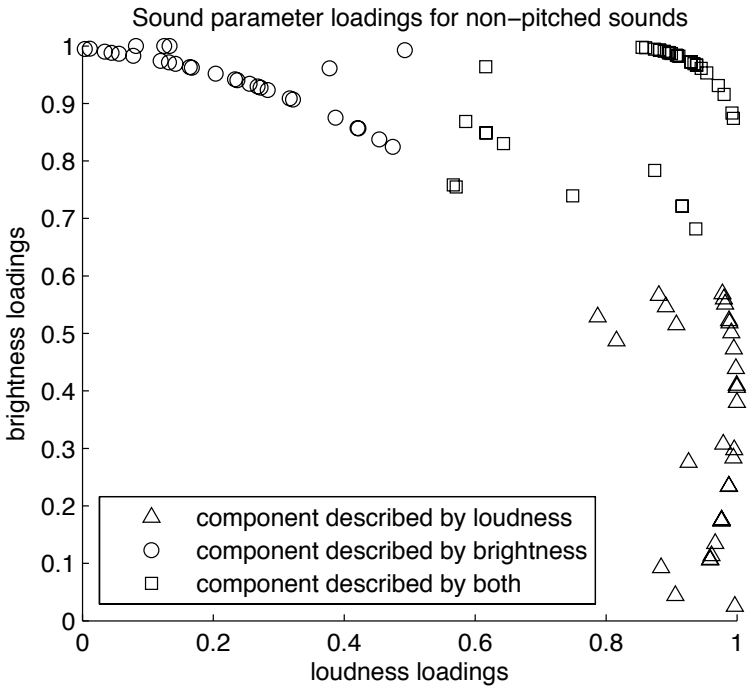

FIGURE 6.4 – Scatter plot showing the distribution of the sound feature loadings for brightness and loudness. Three distinct clusters with high coefficients for brightness, loudness, or both, are found.

Having identified the sound parameters' contribution to the canonical components, we can further inspect how the three classes of components relate to gestural features. Figure 6.5 shows the distribution of the gesture loadings for *horizontal position*, *vertical position* and *velocity* for the 99 canonical components. *Acceleration* has been left out of this plot, since, on average, the acceleration loading was lowest both in the first and second component for all sounds. In

the upper part of the plot, we find the canonical components that are described by vertical position. The right part of the plot contains the canonical components that are described by horizontal position. Finally the color of each mark denotes the correlation to velocity ranging from black (0) to white (1). The three different symbols (triangles, squares and circles) refer to the same classes as in Figure 6.4. From Figure 6.5 we can infer the following :

- For almost every component where the canonical loadings for both horizontal and vertical positions are high (cf. the upper right of the plot), the velocity loading is quite low (the marks are dark). This means that in the instances where horizontal and vertical position are correlated with a sound feature, velocity usually is not.
- The lower left part of the plot displays the components with low correlation between sound features and horizontal/vertical position. Most of these dots are bright, indicating that velocity is an important part in these components.
- Most of the circular marks (canonical components describing brightness) are located in the upper part of the plot, indicating that brightness is related to vertical position.
- The triangular marks (describing loudness) are distributed all over the plot, with a main focus on the right side. This suggests a tendency towards a correlation between horizontal position and loudness. What is even more interesting is that almost all the triangular dots are bright, indicating a relationship between loudness and velocity.
- The square marks (describing both loudness and brightness) are mostly distributed along the upper part of the plot. Vertical position seems to be the most relevant feature when the canonical component describes both of the sound features.

6.5 Discussion

As we have shown in the previous section, there is a very strong correlation between vertical position and pitch for all the participants in our data set. This relationship was also suggested when the same data set was analyzed using a Support Vector Machine classifier (Nymoen et al., 2010), and corresponds well with the results previously presented by Eitan and Granot (Eitan and Granot, 2006). In our interpretation, there exists a one-dimensional intrinsic relationship between pitch and vertical position.

For non-pitched sounds, on the other hand, we do not find such prominent one-dimensional mappings for all subjects. The poor discrimination between features for these sounds could be due to several factors, one of which is that there could exist non-linear relationships between the sound and the motion features that the CCA is not able to unveil. Non-linearity is certainly plausible, since several sound features scale logarithmically. The plot in Figure 6.6, which shows a single sound tracing, also supports this hypothesis, wherein brightness corresponds better with the squared values of the vertical position than with the actual vertical position. We would, however, need a more sophisticated analysis method to unveil non-linear relationships between the sound features for the whole data set.

Furthermore, the scatter plot in Figure 6.5 shows that there are different strategies for tracing sound. In particular, there are certain clustering tendencies that might indicate that listeners select different mapping strategies. In the majority of cases we have found that loudness is described by velocity, but also quite often by the horizontal position feature. Meanwhile, brightness is often described by vertical position. In one of the sounds used in the experiment the loudness and brightness envelopes were correlated to each other. We believe that the sound tracings performed to this sound were the main contributor to the class of canonical components in Figures 6.4 and 6.5 that describe both brightness and loudness. For this class, most components are not significantly distinguished from the components that only describe brightness. The reason for this might be that people tend to follow brightness more than loudness when the two envelopes are correlated.

In future applications for music information retrieval, we envision that sound is not only

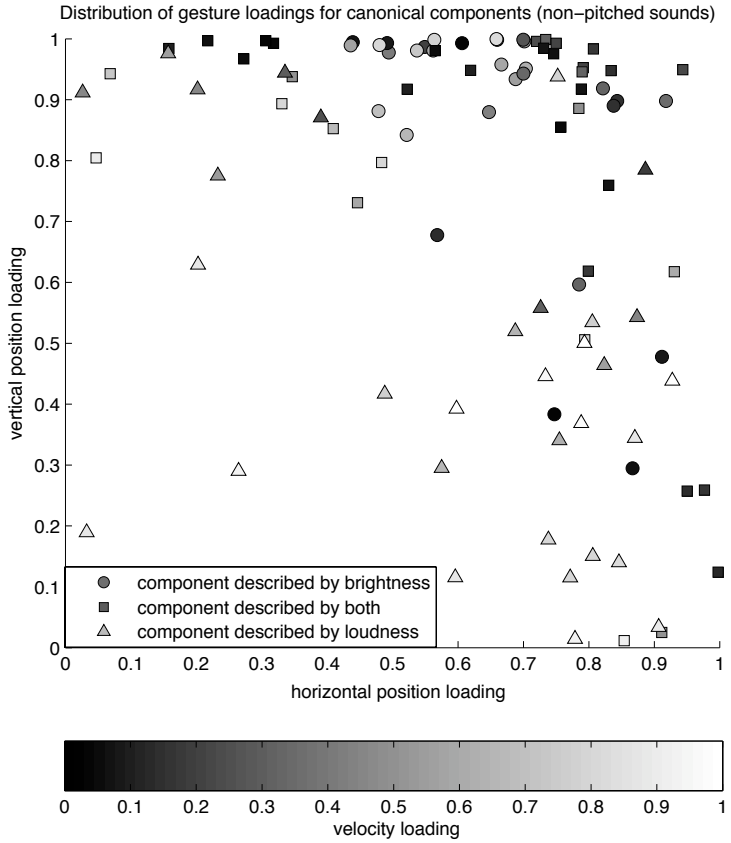

FIGURE 6.5 – Results for the 99 canonical components that had high ρ-values. X and Y axes show correlation for horizontal position and vertical position, respectively. Velocity correlation is shown as grayscale from black (0) to white (1). The square boxes denote components witch also are highly correlated to brightness.

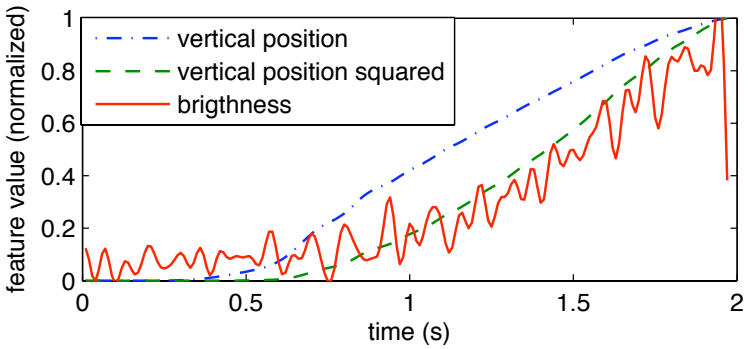

FIGURE 6.6 – Envelopes of brightness, vertical position, and vertical position squared. The squared value corresponds better with brightness than the non-squared value, suggesting a non-linear relationship.

described by audio descriptors, but also by lower-level gesture descriptors. We particularly believe that these descriptors will aid to extract higher-level musical features like affect and effort. We also believe that gestures will play an important role in search and retrieval of music. A simple prototype for this has already been prototyped by the second author (Caramiaux

et al., 2011). Before more sophisticated solutions can be implemented, there is still a need for continued research on relationships between perceptual features of motion and sound.

6.6 Conclusions and future work

The paper has verified and expanded the analysis results from previous work, showing a very strong correlation between pitch and vertical position. Furthermore, other, more complex relationships seem to exist between other sound and motion parameters. Our analysis suggests that there might be non-linear correspondences between these sound features and motion features. Although inter-subjective differences complicate the analysis process for these relationships, we believe some intrinsic action-sound relationships exist, and thus it is important to continue this research towards a cross-modal platform for music information retrieval.

For future directions of this research, we propose to perform this type of analysis on movement to longer segments of music. This implies a need for good segmentation methods, and possibly also methods like Dynamic Time Warping to compensate for any non-synchrony between the sound and people's movement. Furthermore, canonical loadings might be used as input to a classification algorithm, to search for clusters of strategies relating motion to sound.

Troisième partie

Modélisation des Structures Temporelles du Geste

Chapitre 7

Contexte théorique pour la modélisation du temps

Dans les chapitres précédents nous avons montré par des études expérimentales qu'il existe une stratégie cognitive entre geste humain, et évolution temporelle d'un son enregistré. Prendre en compte cette stratégie pour la conception d'un instrument de musique numérique requiert la modélisation des données gestuelles en entrée et de leur évolution dans le temps. Dans ce chapitre nous présentons les outils mathématiques formels permettant la modélisation des structures temporelles dans le geste à des fins d'analyse et de reconnaissance.

7.1 Introduction

L'étude exploratoire présentée dans le chapitre 3 a mis en évidence le besoin de modéliser la structure temporelle des signaux gestuels. L'analyse canonique, comme l'analyse en composantes principales, sont des méthodes statiques (cf. A.1). Elles font l'hypothèse que chaque variable gestuelle ou sonore est une variable aléatoire gaussienne de moyenne et de variance constantes, ne donnant pas accès à la structure temporelle des signaux considérés. C'est une limitation importante pour un but de contrôle, comme discuté dans la section 2.3.2. Dans une performance musicale, les stratégies de contrôle ne sont pas qu'instantanées et certaines relèvent d'une planification à plus long terme induisant des motifs récurrents à la fois dans le geste et dans le son. Ceci avait déjà été pointé dans le chapitre 3. Dans le cas de la valse (appelée *Donauwalzer* dans le corpus) présentée aux participants, les gestes avaient une structure récurrente qui était liée à la structure temporelle de la valse. Ainsi, une des ambitions serait de pouvoir détecter ces motifs, et pour ce faire nous devons modéliser la structure temporelle à une échelle plus étendue.

Dans cette partie, nous présentons deux modèles ayant pour but de détecter les structures temporelles à différentes échelles temporelles. Chaque modèle fait l'objet d'une contribution. La première contribution est la conception d'un modèle dynamique pour les trajectoires de descripteurs gestuels. Le modèle étant dynamique, il repose sur la structure continue du signal d'observations et possède une précision temporelle à une échelle fine correspondant à l'échantillon. De cette manière le modèle est capable de s'adapter localement aux variations temporelles du geste. Nous inspecterons précisément son utilisation pour la reconnaissance du geste avec invariants. Une modélisation à cette échelle peut être utilisée pour le contrôle *micro* du son. La seconde contribution est l'utilisation d'un modèle pour la représentation d'un geste en une séquence de formes géométriques prises dans un dictionnaire donné. La structure temporelle prise en compte est à l'échelle du segment (a fortiori du *symbole*, même si nous n'utiliserons que très peu ce terme). L'intérêt est de pouvoir analyser la séquence de segments inférée d'un point de vue syntaxique faisant apparaître des motifs récurrents qu'il serait diffi-

cile d'extraire des signaux bruts. Notre motivation est donc de faire émerger la structure *macro* des gestes musicaux. Celle-ci peut-être utilisée pour le contrôle du son avec une stratégie à plus long terme. La figure 7.1 schématise les deux niveaux de modélisation introduits. Sur cette figure, nous commençons avec le signal de geste échantillonné. Le niveau qu'on appelle *micro* modélisant la trajectoire du geste (et de ses caractéristiques) est décrit par la courbe intermédiaire. Enfin au niveau de la séquence, des segments de courbes sont assimilés faisant apparaître les récurrences et la structure « grammaticale » sous-jacente.

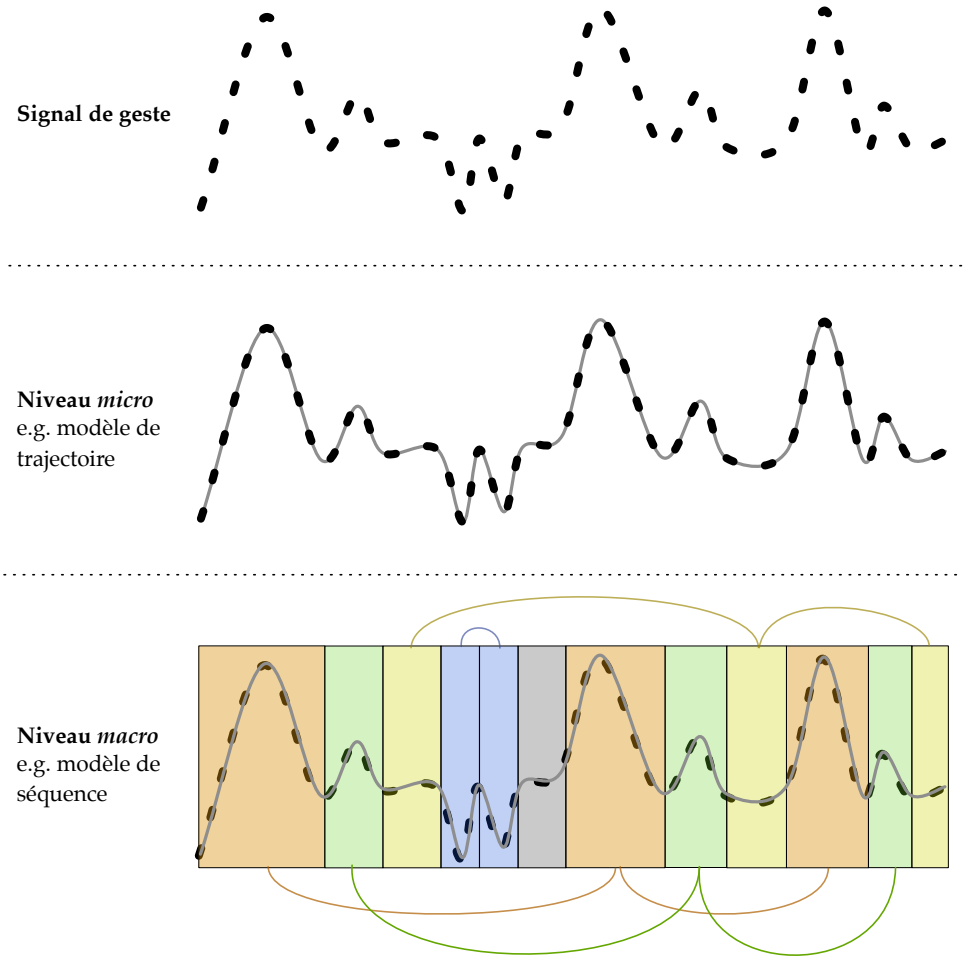

FIGURE 7.1 – Modélisation des structures temporelles dans le signal gestuel (en haut). La figure du milieu reporte un modèle de trajectoire. La figure du bas reporte un modèle de reconnaissance des motifs dans le signal impliquant une modélisation et une analyse au niveau symbolique.

Les modèles que nous présenterons dans cette partie utilisent des méthodes d'inférence Bayésienne, l'hypothèse étant que les données utilisées véhiculent une incertitude qu'il nous faut prendre en compte. Cette incertitude provient à la fois des procédés de mesure et de l'incapacité pour un être humain de refaire exactement le même geste. Nous avons voulu limiter ce chapitre à un état de l'art des modèles pour la reconnaissance ou le suivi de geste. Les justifications théoriques des modèles présentés sont détaillées dans l'annexe A. Ainsi, nous proposons tout d'abord d'introduire le contexte probabiliste pour la modélisation du temps (section 7.2). Ensuite, nous présentons un état de l'art général pour introduire les deux articles qui seront reportés dans les chapitres 8 et 9. Nous présenterons dans la section 7.3.1 les chaînes de Markov cachées puis leur extension au cas continu avec les systèmes dynamiques (section 7.3.2). De là, nous présenterons des modèles qui prennent en compte la structure temporelle à plusieurs échelles. Nous introduirons les modèles segmentaux (section 7.3.3), les modèles hiérarchiques

(section 7.3.4) et enfin les modèles de séquences de systèmes dynamiques (section 7.3.5).

7.2 Contexte probabiliste pour la modélisation du temps dans les signaux gestuels

L'application de méthodes statiques, telles que l'analyse en composantes principales (*Principal Component Analysis* (PCA), reportée dans l'annexe A.1 avec son état de l'art), pour l'analyse des données de captation néglige la structure temporelle sous-jacente du geste. De même que considérer une variable gestuelle comme une variable aléatoire (moyenne et variance constantes au cours du temps) néglige le caractère non-stationnaire du geste. Cette structure temporelle est pourtant une connaissance importante en vue d'une tâche de *modélisation*. Modéliser consiste ici à donner une description qui capture les caractéristiques du comportement du système à étudier sur le long terme (Taylor, 2009). Ainsi, dans ce manuscrit, modéliser suggère de trouver une structure capturant la dépendance temporelle entre les données observées. Comme introduit dans la section 2.1.3, il est souvent difficile de définir un ensemble d'équations dynamiques explicites pour la modélisation des observations gestuelles. Un grand nombre de méthodes utilisent plutôt le formalisme probabiliste afin d'apprendre la structure à partir des données ou, ayant une structure déjà définie, de trouver les paramètres les mieux adaptés aux données expérimentales (ang. *fitting*). Dans le cadre de cette thèse, nous nous plaçons dans un formalisme probabiliste Bayésien permettant d'inférer les dépendances temporelles à partir des données d'observations.

7.2.1 Cadre Bayésien pour la modélisation du temps

Le cadre Bayésien suggère la définition d'une distribution a priori qui encode notre croyance initiale et qui *infère* une distribution a posteriori laquelle synthétise l'apport des données dans l'étude du système. On parle ainsi d'inférence Bayésienne. L'inférence Bayésienne est très utilisée dans divers domaines de recherche qui utilisent des modèles statistiques (ou *stochastiques*, c'est à dire ayant une dépendance temporelle) pour le filtrage, la prédiction ou lissage de données expérimentales. La loi fondamentale sur laquelle s'appuie ce processus d'inférence, et qui apparaîtra sous diverses formes dans cette thèse, est le théorème de Bayes définit par l'équation suivante :

$$p(X|Y) = \frac{p(Y|X)}{p(Y)} p(X) \tag{7.1}$$

On réfère le lecteur aux divers ouvrages pour plus de détails, comme par exemple (Bishop, 2006). Dans l'étude de séries temporelles par inférence Bayésienne, une variable aléatoire Y (resp. X) est la donnée d'un processus stochastique à un certain instant t : Y_t (resp. X_t). On considère Y_t comme étant observable et X_t comme étant nos hypothèses sur les observations. L'équation de Bayes lie notre croyance a priori avec une probabilité sur notre hypothèse étant données les observations. Ceci a conduit à un grand nombre de modèles dits *modèles à états* (un exemple sera donné dans la section suivante avec la figure 7.2).

Les modèles à états modélisent la dépendance entre les variables observées et des variables latentes qui factorisent nos hypothèses. Afin d'éclaircir notre propos, prenons l'exemple de la PCA. Dans l'annexe A.1, nous évoquons l'existence d'une formulation probabiliste de la PCA que nous analysons ici. L'équation de la PCA est la suivante :

$$\mathbf{X} = \mathbf{YA}$$

Où $\mathbf{Y}$ sont les variables observées corrélées, $\mathbf{X}$ les variables extraites non corrélées et $\mathbf{A}$ la matrice de projection (orthogonale). De cette équation nous pouvons écrire :

$$\mathbf{Y} = \mathbf{X}\hat{\mathbf{A}}$$

Où $\hat{\mathbf{A}} = \mathbf{A}^{-1}$. La version probabiliste définit **X** comme la variable latente et **Y** la variable observée. La figure 7.2 permet une visualisation graphique de la relation entre variable observée et variable latente. Pour plus détails, on réfère le lecteur à la littérature sur les modèles graphiques (Bishop, 2006) (chapitre 8), les réseaux Bayésiens (Jensen, 1996) et les réseaux Bayésiens dynamiques (Murphy, 2002). De manière simplifiée, un cercle grisé indique une variable continue observable alors qu'un cercle blanc indique une variable latente. Une arête de A vers B indique une dépendance entre les distributions de A et B. Une arête orientée indique une causalité entre deux variables.

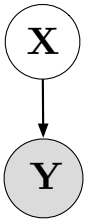

FIGURE 7.2 – Modèle graphique pour une variable observée dépendant d'une variable latente.

Dans un cadre Gaussien la densité de probabilité conditionnelle de **Y** sachant **X** s'écrit :

$$p(\mathbf{Y}|\mathbf{X}) \sim \mathcal{N}(\mathbf{X}\hat{\mathbf{A}} + \mu; \sigma)$$

Où μ est une constante permettant de décentrer **Y**. Une méthode classique pour trouver les paramètres optimaux (la matrice $\hat{\mathbf{A}}$) sachant les observations, repose sur la maximisation de la log-vraisemblance (on réfère le lecteur à l'article (Tipping and Bishop, 1999) pour plus de détails).

De cet exemple simple, plusieurs points de conception de modèles se dessinent.

1. les variables latentes synthétisent notre hypothèse, l'information qu'on cherche à extraire des observations ;
2. cette information peut être quelconque à partir du moment où on peut écrire les densités de probabilité conditionnelles (DPC) dans le cadre de la sémantique Bayésienne ;
3. si au lieu de **X** et **Y** nous posons $\mathbf{x}_t$ et $\mathbf{y}_t$ nous pouvons lier certaines variables temporellement dans la mesure où on peut écrire les DPC ;
4. le cadre Bayésien offre des outils puissants pour l'apprentissage des paramètres (fréquemment basés sur la maximisation de la vraisemblance).

7.2.2 Architecture à plusieurs niveaux temporels

Nous avons introduit dans la première partie de la thèse, la pertinence de prendre en compte différents niveaux temporels dans la structure du geste. Ceci a été justifié à la fois pour la perception du mouvement, pour les contraintes cognitives du programme moteur et pour le contrôle expressif dans le contexte musical. Dans le formalisme Bayésien des modèles à états, un niveau temporel peut-être modélisé par l'ajout d'un état supérieur générant les états inférieurs (Figure 7.3).

Ainsi les observations (Y dans la Figure 7.3) sont-elles générées par un état (X dans la Figure 7.3), par exemple représentant la dynamique sous-jacente. Nous dirons que la génération se fait au niveau *signal*. L'état représentant la dynamique dépend lui-même d'un état supérieur (Z dans la Figure 7.3) qui gouverne par exemple un certain type de dynamique ou des paramètres haut-niveau. Ces états peuvent donc encoder une structure temporelle à plus long terme incluse dans le signal. Nous dirons que la génération se fait au niveau *symbolique*.

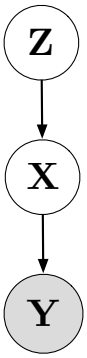

FIGURE 7.3 – Modèle graphique à plusieurs niveaux (ici deux niveaux) : une variable observée **Y** dépendant d'une variable latente **X** qui dépend elle-même d'une autre variable latente **Z**.

7.3 Modèles Bayésiens Dynamiques

Dans cette section nous présentons différents modèles pour la structure temporelle du geste. Nous inspecterons leur utilisation dans la littérature. En début de chaque partie, nous introduirons le modèle étudié sous forme de modèle graphique, comme introduit dans la section précédente. Une description formelle plus détaillée peut être trouvée dans l'annexe A.

7.3.1 Chaîne de Markov cachée

Les modèles de Markov cachés (ou chaînes de Markov cachées, *Hidden Markov Models* (HMM)) peuvent se représenter graphiquement comme indiqué par la figure 7.4 (un carré indique une variable discrète). Les variables latentes sont les variables cachées qui ne dépendent que de la variable précédente dans la chaîne (propriété de Markov).

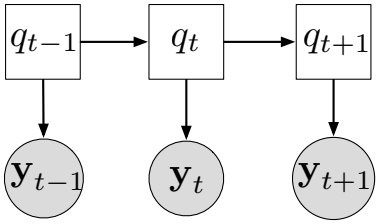

FIGURE 7.4 – Modèle graphique de HMM. Les variables q_i sont discrètes.

Les HMMs ont été largement utilisés dans des domaines liés au traitement du signal pour les raisons suivantes : ils peuvent apprendre à partir des données ; ils sont robustes aux variations temporelles des données d'entrée (dilatation temporelle dynamique) ; et la structure probabiliste basée sur une inférence Bayésienne leur permet de prendre en considération des incertitudes dans les variables observées et inférées. Ils ont été utilisés avec succès pour la reconnaissance d'écriture manuscrite : on renvoie le lecteur aux travaux pionniers de Nag et al. (Nag et al., 1986), Chen et al. (Chen et al., 1994) ou encore Hu et al. (Hu et al., 1996). À notre connaissance, la première tentative d'utilisation des HMMs pour la reconnaissance du mouvement a été faite par Yamato et al. dans (Yamato et al., 1992). Les auteurs proposent la reconnaissance du mouvement à partir d'images vidéo. Une séquence d'images est convertie en séquence de descripteurs du mouvement puis en une séquence de symboles donnée par quantification de vecteurs (ang. *vector quantization*). Cette séquence de symboles est ensuite utilisée pour apprendre les paramètres d'un HMM.

Par la suite, les HMMs ont connu plusieurs extensions qui ont permis d'outrepasser leurs limitations inhérentes :

- Les données observées sont supposées issues d'un seul processus : Brand et al. (Brand et al., 1997) ont alors proposé des HMMs couplés. Le couplage permettant de modéliser l'action de plusieurs processus sur des données observées (par exemple interaction entre personnes).
- La capacité de représentation est limitée : représenter k bits d'information sur le passé nécessite 2^k états. Ghahramani et al. proposent un modèle à états distribués et l'appellent HMM Factoriel (Ghahramani and Jordan, 1997). Brown et al. (Brown and Hinton, 2001) proposent le produit de HMMs qui permet l'inférence et l'apprentissage de manière indépendante sur chaque HMM ce qui permet une inférence exacte sur le produit des HMMs (à la différence des HMMs factoriels). Taylor et al. (Taylor and Hinton, 2009) proposent son utilisation pour la reconnaissance du mouvement dansé.
- L'apprentissage ne se fait qu'à partir des observations : l'*Input/Output* HMM proposé par Bengio et al. (Bengio and Frasconi, 1995) permet de faire un apprentissage sur les données d'observation et sur la relation entre des données d'entrée (*input*) et des données d'observation (*output*).
- Les paramètres du modèle sont statiques : le HMM Paramétrique permet de prendre en compte les variations spatio-temporelles des données d'entrée par adaptation des paramètres et a été appliqué pour la reconnaissance de geste (Wilson and Bobick, 1999), et la modélisation du style (Brand and Hertzmann, 2000).
- La modélisation de durée d'un état : la modélisation des durées dans un HMM est restrictive car elle est fixe et suit une exponentielle décroissante. Les modèles semi-Markov (Yu, 2010) outrepassent cette limitation. Nous reviendrons sur cette limitation dans la section 7.3.3 où nous proposerons l'utilisation des modèles segmentaux.

7.3.2 Système dynamique et filtrage particulaire

Les systèmes dynamiques peuvent se représenter par un modèle à états comme indiqué sur la figure 7.5. Dans le cas où les transitions (entre variables latentes) et émissions (des variables latentes vers les observations) sont régies par des matrices, l'inférence est exacte et optimale par filtrage de Kalman. Dans le cas général, le filtrage de Kalman ne peut pas être utilisé.

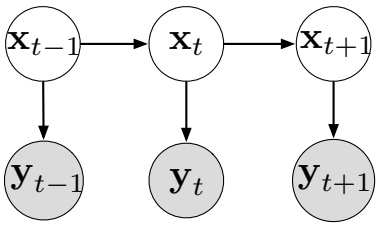

FIGURE 7.5 – Modèle graphique de système dynamique. Les variables continues x_t sont les valeurs d'un état. La transition entre états est faite par le système dynamique.

Le filtrage particulaire est une méthode incrémentale pour l'estimation de distributions de probabilité. En cela, cette méthode a été principalement utilisée dans des applications temps réel et particulièrement pour le suivi, la localisation et l'alignement (Arulampalam et al., 2002; Comaniciu et al., 2003; Dellaert et al., 1999; Doucet et al., 2000). C'est une méthode d'inférence non-exacte pour les réseaux Bayésiens dynamiques.

Même si l'apparition des filtres particulaires date des années 60-70 (Handschin and Mayne, 1969), cette méthode n'a véritablement connu un essor que dans les années 90 lorsque la puissance de calcul devint assez grande pour que la méthode puisse être, de manière pragmatique, utilisée en temps réel (on réfère le lecteur à l'article par Doucet et al. (Doucet et al., 2000) qui

fait l'examen des premiers travaux liés aux filtres particulaires). Par son inférence incrémentale, son utilisation s'est développée pour le suivi d'objets. Le suivi d'objets par filtrage bayésien est une problématique bien connue dans le domaine de la vision. Isard et al. ont proposé l'algorithme CONDENSATION (*CONDitional dENSity propagATION*) (Isard and Blake, 1998a; Isard and Blake, 1998b) où un système dynamique linéaire (basé sur un modèle auto-régressif AR) régit la transition entre les états $\mathbf{x}_{t-1}$ et $\mathbf{x}_t$. L'échantillonnage est effectué à partir de la distribution *a priori*. À la différence des filtres de Kalman, l'algorithme basé sur le réechantillonnage d'importance (ou *Sequential Importance Sampling* (SIS) en anglais) permet de suivre plusieurs cibles à la fois (grâce à l'estimation d'une distribution non gaussienne). L'algorithme est testé pour le suivi de mains dans une scène où il y a beaucoup d'objets parasites de même que pour le suivi de plusieurs personnes dans une séquence vidéo. La simplicité de la méthode et son efficacité pour des situations concrètes en vision ont fait naître une attention particulière sur des extensions possibles. Parmi celles-ci, Black et al. (Black and Jepson, 1998b) ont proposé un algorithme permettant la reconnaissance et l'alignement de trajectoires basés sur l'estimation d'un paramètre de position, d'un paramètre d'amplitude et d'un paramètre de vitesse. Les auteurs proposent une application pour la reconnaissance de gestes simples bidimensionnels pris comme commandes pour les IHMs tels que "copier", "coller", "supprimer".

Du fait de leur nature continue, les filtres particulaires peuvent s'insérer comme une entrée pour le contrôle continu de médias numériques tel que le son. Dans (Hermann et al., 2001), Hermann et al. explorent la sonification des densités de probabilités multidimensionnelles à l'aide du filtrage particulaire. L'objectif est l'écoute informative des simulations et la compréhension (par l'audition) des densités de probabilités. Dans (Williamson and Murray-Smith, 2005), les auteurs proposent un *feedback* sonore de l'incertitude dans la reconnaissance gestuelle afin d'améliorer la qualité de l'interaction. Le système de reconnaissance se base sur le filtrage particulaire et la synthèse granulaire est choisie pour le feedback sonore.

Visell et al. (Visell and Cooperstock, 2007a; Visell and Cooperstock, 2007b) proposent un système de sonification des mouvements humains basé sur la synthèse spectrale pilotée par le résultat de la reconnaissance. La reconnaissance est basée sur des systèmes dynamiques non-linéaires (issus des travaux de Schaal et al. (Schaal et al., 2003)) pris comme références (ou *templates* en anglais). Les auteurs se focalisent sur des mouvements de marche pour lesquels la synthèse sonore est une aide au mouvement. Le domaine d'application est la réhabilitation.

7.3.3 Modèle segmental

Les modèles segmentaux peuvent se représenter graphiquement comme indiqué par la figure 7.6.

Une limitation majeure des modèles basés sur HMM pour l'analyse de série temporelle réside dans la distribution implicite de la durée dans chaque état caché. En effet, dans une chaîne de Markov (d'ordre 1), la probabilité de rester pendant k observations dans un état i avant de passer à un état j ($j \neq i$) est une exponentielle décroissante. Or, cette hypothèse est très restrictive pour beaucoup de systèmes naturels notamment les mouvements humains. Ceci a amené le développement de modèles ou la distribution de probabilité sur les durées est explicite, notamment par l'utilisation d'une chaîne semi-Markovienne. Un cas particulier des modèles de semi-Markov cachés sont les modèles segmentaux. Ces modèles permettent de contrôler la durée dans les états en définissant ces états non plus comme émettant une observation mais une séquence d'observations. Ainsi ces modèles permettent de capter la structure temporelle du signal non plus au niveau de l'état mais au niveau d'un segment.

Ces modèles ont été utilisés pour la reconnaissance et tout particulièrement pour la reconnaissance de la parole (Russell, 1993; Ostendorf et al., 1996; Achan et al., 2004). L'engouement dans cette communauté pour des modèles segmentaux provient du lien acoustique / phonétique. Glass l'explique dans (Glass, 2003) et le passage mérite d'être reporté in extenso :

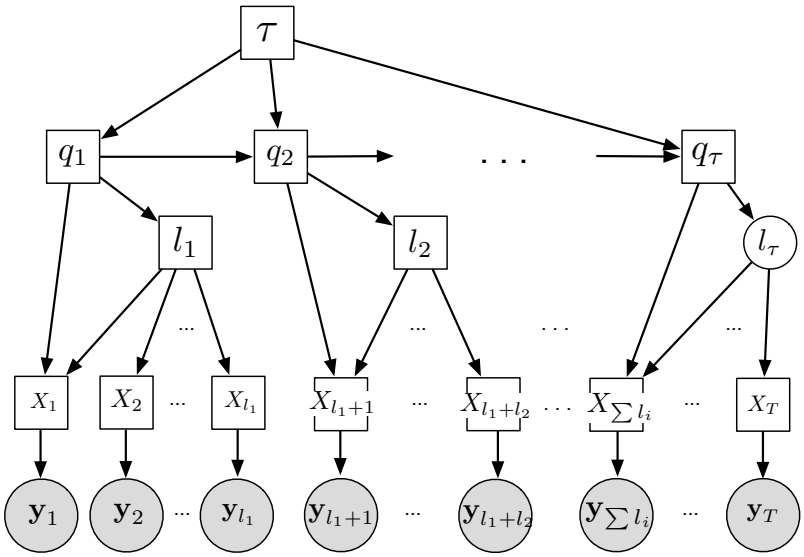

FIGURE 7.6 – Modèle graphique de modèle segmental. Les variables l_i représentent les durées du segment i. Les variable q_i sont des variables discrètes ou $q_i = k$ indique le k-ème segment. Enfin τ est le nombre de segments pour représenter le signal.

> *While HMMs have shown themselves to be highly effective, it is reasonable to question some of their basic structure. For example, almost all HMM-based ASR [Automatic Speech Recognition] formulations restrict their acoustic modelling to an observation space defined by a temporal sequence of feature vectors computed at a fixed frame rate, typically once every 10 ms. As a result, adjacent feature vectors, especially those within the same phonetic segment, often exhibit smooth dynamics and are highly correlated, violating the conditional independence assumption imposed by the HMM model (Digalakis, 1992). The relationship between features computed in different phonetic segments is weaker, however. These observations motivate a framework which makes fewer conditional independence assumptions between observation frames ; especially for those occurring within a phonetic segment.* (cf. (Glass, 2003) p.138)

Ceci est le cas des travaux de Russell et al. (Russell, 1993) dans lesquels les auteurs veulent réduire l'impact de l'hypothèse d'indépendance entre des observations consécutives. Une étude comparative entre leur modèle et un HMM classique a été menée afin de convaincre de son utilisation (Holmes and Russell, 1995). A cette période un panel d'approches segmentales pour la reconnaissance de parole apparaît (Digalakis, 1992; Gales and Young, 1993; Kannan and Ostendorf, 1993; Zue et al., 1989). Ostendorf et al. (Ostendorf et al., 1996) ont formalisé les approches segmentales précédentes sous forme d'un modèle unique, présenté en annexe A.3, mettant en jeu une forme générale de segment. En dehors du domaine de la parole, peu de travaux ont utilisé ce modèle. Ceci s'explique par le fait que les formes de segments utilisés dans le modèle ne sont pas toujours bien définies. On retrouve ce modèle pour la détection de sauts (dans le sens de singularité) (Ge and Smyth, 2000), la reconnaissance de vues de camera pour le sport (Ding and Fan, 2007), ou encore pour la reconnaissance de formes dans des mouvements fluides (Kim and Smyth, 2006).

Dans un domaine connexe au nôtre, Artières et al. (Artières et al., 2007) ont utilisé un modèle segmental pour la reconnaissance d'écriture manuscrite. Les auteurs proposent de reconnaître les caractères comme des séquences de traits pris dans un dictionnaire. Chaque caractère correspond à une séquence particulière apprise sur une base d'entraînement. Cette même architecture a été reprise par Bloit et al. dans (Bloit et al., 2010) pour la reconnaissance de profils de descripteurs audio. La motivation étant une modélisation morphologique du son et de la musique afin de sortir du carcan des éléments symboliques (comme les notes) qui forment le

système musical occidental.

Ainsi le modèle segmental a-t-il été utilisé avec des séries temporelles pour lesquelles des primitives étaient bien définies. Dans le cas du geste musical (et particulièrement le geste ancillaire), les primitives peuvent dépendre du musicien, de l'instrument, du style, etc... De fait, ce type de modèle n'a jamais été utilisé dans ce contexte. Dans la première partie de ce manuscrit, nous avons montré que le geste, en réaction à la musique, peut contenir des éléments récurrents, des motifs. L'approche segmentale semble pertinente pour la modélisation de cette structure temporelle haut-niveau.

7.3.4 Modèle hiérarchique

Les modèles hiérarchiques peuvent se représenter graphiquement comme indiqué par la figure 7.7. Les carrés *double* indique un état terminal, permettant de changer de primitives à un niveau supérieur. Le retour au niveau supérieur se fait automatiquement, sans probabilité, d'où la flèche pointillée.

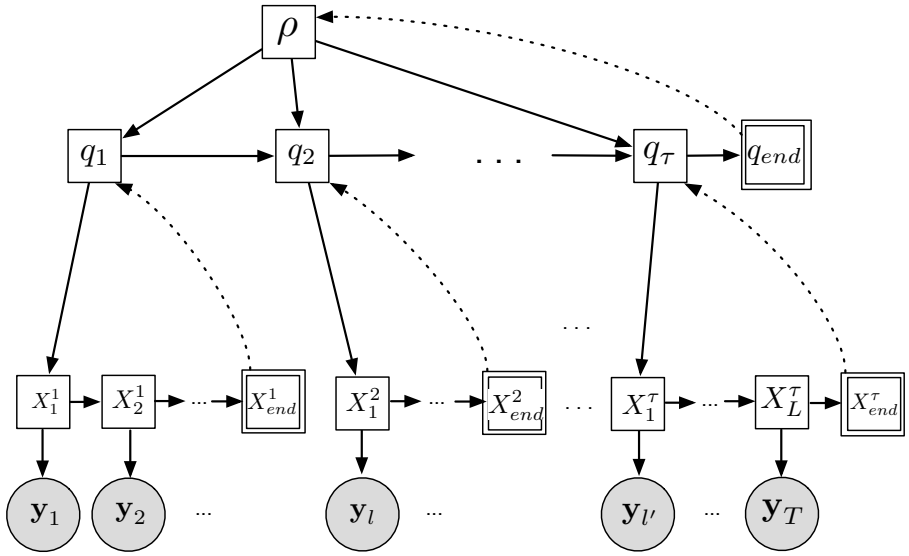

FIGURE 7.7 – Modèle graphique de modèle hiérarchique. Les variable q_i sont des variables discrètes ou $q_i = k$ indique le k-ème segment. La vriable ρ est la racine, pointant vers le prochain segment. Il n'y a pas de variables de durée, la transition à un prochain segment se fait à la fin du segment courant.

Les modèles hiérarchiques ont été peu utilisés dans la littérature. Le modèle présenté par Fine et al. (Fine et al., 1998) est appliqué à des données de la parole et à l'écriture manuscrite. Bien que l'architecture soit très intéressante, les méthodes d'inférence (exacte) proposées sont cubiques en temps et rendent le modèle peut attractif. Murphy et al. proposent dans (Murphy and Paskin, 2001) une version linéaire en temps de l'inférence basée sur la représentation du modèle hiérarchique sous forme de réseau bayésien dynamique. La perte se traduit dans la complexité en espace qui devient quadratique par rapport au nombre de niveaux et au nombre maximum d'états dans chaque niveau.

Le modèle hiérarchique a été utilisé pour la reconnaissance d'activités (Bui, 2003; Bui et al., 2004; Nguyen et al., 2005). En effet, l'activité humaine est une séquence d'actions, de gestes ou de mouvements hiérarchiquement organisés (Turaga et al., 2008) définissant un cadre d'application propice pour ce type de modèles. La même structure a ensuite été appliquée pour la reconnaissance de gestes de manipulation avec contexte (Li et al., 2005). Les auteurs définissent trois niveaux. Le niveau supérieur correspond à la tâche de manipulation, par exemple "prendre" ou "déplacer". Le niveau intermédiaire correspond à l'objet manipulé, par exemple "un stylo". Le dernier niveau est la primitive. Rajko et al. (Rajko et al., 2007) proposent une

variante des modèles hiérarchiques, qu'il nomme modèle sémantique pour la reconnaissance de gestes bidimensionnels.

7.3.5 Système dynamique linéaire par morceaux

Les systèmes dynamiques (linéaires dans notre cas) par morceaux peuvent se représenter graphiquement comme indiqué par la figure 7.8.

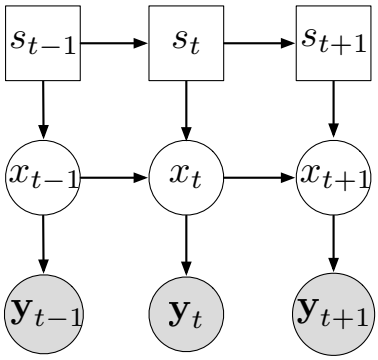

FIGURE 7.8 – Modèle graphique de systèmes dynamiques par morceaux. Les variables x_t sont des variables continues d'un système dynamique k donné par la valeur de la variable discrète s_t. Ici à chaque échantillon on a l'indication du système dynamique courant.

A notre connaissance, les premiers travaux portant sur l'utilisation des systèmes linéaires par morceaux (ou *Switching Linear Dynamic Systems* (SLDS)) à la reconnaissance de gestes sont ceux par Bregler (Bregler, 1997). Dans son article, il ne les présente pas précisément de la même manière mais plutôt comme plusieurs systèmes dynamiques définissant des classes (symboliques) et l'émission d'une classe ou transition entre les classes est effectuée par un HMM. Ensuite Murphy définit plus précisément la structure des SLDS comme un commutateur entre filtres de Kalman (Murphy, 1998). L'hypothèse est qu'un système naturel n'est pas linéaire et qu'il est pertinent de l'approcher par une séquence de systèmes linéaires. L'inférence est malheureusement non-exacte. Pavlovic et al. (Pavlovic et al., 2001) explore l'utilisation de plusieurs méthodes d'inférence pour la reconnaissance de mouvements humains qui sont des séquences de *marche* et de *course*. Ce modèle obtient de meilleurs résultats qu'un HMM ou qu'un filtre de Kalman étendu. Enfin, Oh et al. (Oh et al., 2008) proposent de nouvelles méthodes d'inférence testées sur le mouvement des abeilles qui montrent une succession de mouvements périodiques dans une direction et un retour curviligne.

7.4 Synthèse

Nous avons présenté dans ce chapitre un panel de modèles à états induisant des variables discrètes (HMM) ou continues (Système dynamique) et permettant de considérer la dépendance temporelle des variables à plusieurs échelles (Modèle segmental, Modèle hiérarchique, SLDS). L'ajout de complexité dans le réseau bayésien dynamique (continuité, hiérarchie entre états) augmente le *pouvoir* de modélisation mais comporte plusieurs inconvénients, notamment la complexité dans l'apprentissage et l'impossibilité d'utiliser des méthodes d'inférence exacte.

Dans la suite de cette partie, nous présenterons deux études portant sur l'utilisation de modèles à états pour l'analyse du geste. Dans la première nous décrirons une nouvelle méthode adaptative pour la reconnaissance. Cette méthode utilise un modèle de système dynamique non-linéaire. L'inférence n'est pas exacte et se fera par filtrage particulaire. L'étude correspond au chapitre 8 et à l'article (Caramiaux et al., 2012a).

La deuxième étude propose l'utilisation du modèle segmental pour la segmentation et l'analyse syntaxique du geste du musicien. Le modèle segmental nous permet de choisir une forme de régression pour chaque état et ainsi décomposer un geste en entrée sur ces formes. L'étude correspond au chapitre 9 et à l'article (Caramiaux et al., 2012c).

Chapitre 8

Realtime Adaptive Recognition of Continuous Gestures

B. Caramiaux[1], **N. Montecchio**[2], **F. Bevilacqua**[1]

[1] UMR IRCAM-CNRS, Paris, France

[2] University of Padova, Department of Information Engineering, Padova, Italy

Abstract : This paper presents a recognition method for gestures that are represented as multidimensional time series. The proposed method relies on defining gesture classes using single templates, to allow users to define their own gesture with a simple procedure. The method simultaneously aligns the input gesture to the templates using a Sequential Montecarlo inference technique. Contrary to standard template-based methods based on dynamic programming, such as Dynamic Time Warping, the algorithm can adapt and track in real-time gesture variations. Different evaluations were performed on synthetic data, 2D pen gestures and 3D gestures captured with accelerometer sensors. The results show that the proposed method performs equally or better compared to other standard template-based methods, due to its dynamic adaptive capability. Moreover, the method continuously updates, during the gesture, the estimated parameters and recognition results which offers key advantages for continuous man-machine interaction.

Keywords : Gesture Recognition, Particle Filtering, Continuous Gesture Modeling, Adaptive Decoding, Gesture Analysis

8.1 Introduction

The now ubiquitous use of multitouch interfaces and inertial measurement sensors has created new usages of gestures in Human-Machine Interaction (Jordà, 2008; Rasamimanana et al., 2011). Several methods for gesture recognition have been previously proposed (Mitra and Acharya, 2007), the most efficient ones being generally tied to cumbersome training procedures. This implies that users must conform to a predefined fixed gesture vocabulary, which might represent an important hurdle in comprehending and learning the system. Moreover, most systems are not designed to adaptively take into account variations that occur during the gesture performance.

The gesture recognition system we propose was designed to avoid the aforementioned shortcomings : the method operates with single templates to define gesture classes with a simple procedure and it can accommodate a large set of gesture variations within each class, for example in the speed, the amplitude or the orientation. Importantly, these variations can be estimated continuously during the performance. Therefore, the system is designed to allow users to personalize the system, by choosing themselves gesture templates while being flexible to gesture variations. It also fits in cases where the gesture expressivity is a concern, since this implies recognizing the gesture unit itself (as a symbol) and how the gesture is performed.

This system is suited for continuous gestures, such as finger trajectories on a tablet or the hand motions manipulating an interface (e.g. game interface or mobile phones). It can be seen as an extension of a previous system we developed called *Gesture Follower* (Bevilacqua et al., 2010; Bevilacqua et al., 2011b), that was found effective to recognize and synchronize continuous gestures to digital media such as sounds and visuals (Bevilacqua et al., 2012). For example, this system was found successful for music control using tangible interfaces (Rasamimanana et al., 2011), for music and dance pedagogy (Bevilacqua et al., 2011a; Bevilacqua et al., 2007) and for gesture-based gaming systems (Rasamimanana and Bevilacqua, 2012).

This body of work allowed us to establish a series of requirements that guided the method we report in this paper :

1. Training procedure must be based on single templates, to allow users to define their gesture vocabulary with simple and direct procedures.
2. The results should be updated continuously during the gesture. This allows the results to be used in continuous interaction paradigms or for anticipation (as proposed for example by Bau et al. (Bau and Mackay, 2008) and Appert et al. (Appert and Bau, 2010)). This generally requires taking into account the gesture's fine temporal structure.
3. The gesture variations that occur during the performance should be taken into account and estimated.

The *gesture follower* was designed to handle the first two points above, but fails to handle gesture variations properly. The method presented here, based on a Sequential Monte Carlo inference, tackles this issue of continuously adapting to the gesture variations.

The paper is structured as follows. First, we review the state of the art on continuous gesture recognition (Section **??**). Second, we present the model and algorithm (Section **??**). In Section 8.4, we show results on synthetic data, illustrating the basic mechanisms of the algorithm. In Section 8.5, we present evaluations performed on real data in two cases : 2D pen-gestures and 3D gestures captured with accelerometer sensors. Finally, we discuss in Section 8.6 the different features of our method and its applications.

8.2 Related Work

We review in this section the most often used methods to recognize gestures represented as multidimensional times series.

Considering 2D drawing gestures, several basic methods take advantage of simple distance functions between gestures. Rubine (Rubine, 1991) proposes a geometric distance measure based on examples of single-stroke gestures. Wobbrock et al. propose a simple template-based method that makes use of Euclidean distance (Wobbrock et al., 2007), after a pre-processing stage in order to take into account geometric variations (such as scaling and rotation) and speed variations (by uniformly resampling the data).

Several methods are based on Dynamic Programming (DP) to handle local time variations. The most widely used technique is Dynamic Time Warping (DTW), that requires the storage of the whole gesture temporal structure (Gavrila and Davis, 1995; Liu et al., 2009). A similarity matrix is computed between the test gesture and a reference template (typically using Euclidean distance) and the optimal path is computed, representing the best alignment between the two time series. Applications are various, such as gesture control (Merrill and Paradiso, 2005), communicative gesture sequences (Heloir et al., 2006), querying based on human motion (Forbes and Fiume, 2005). An extension of DTW method has been proposed by Bobick et al. (Bobick and Wilson, 1997), to take into account several examples, using principal curve in the DP computation. One of the main drawback of methods based on DP is that it does not provide an explicit noise model, and does not prevent from errors due to unexpected or lost observations in the incoming sequence.

Statistical methods prevent such shortcomings, such as the well-known Hidden Markov Models (HMM) (Rabiner, 1989). HMMs are based on a probabilistic interpretation of the observation and can model its temporal behaviour using a compact representation of gesture classes. HMMs have been successfully applied in human motion recognition from vision-based data as explained in the review (Mitra and Acharya, 2007). HMM-based methods are generally robust since they rely on learning procedures based on large databases, modeling thus the variations occurring within a gesture class (Bilmes, 2002).

Several extensions of HMM have been proposed. Wilson highlights the need to adapt gesture in realtime for interacting with machines (Wilson, 2000). In (Wilson and Bobick, 1999), Wilson and Bobick propose a model that takes into account parametric changes in execution. They describe an application where bi-handed gesture semantics are related to the global trajectories (for example actions on an object) while variations provide with additional meaning (for example the size of the object). In this case, the amplitude is defined globally on the whole gesture (see also (Brand and Hertzmann, 2000)). In (Wilson and Bobick, 2000), Wilson and Bobick describe an online learning method that can be applied to each different user. A case study is described, where simple gestures such as "rest", "down", and "up" are recognized. Nevertheless, this method does not allow for the continuous adaptation of gesture class.

A recent method proposed by Bevilacqua et al. (Bevilacqua et al., 2010; Bevilacqua et al., 2011b), called *Gesture Follower* makes use of the HMM statistical framework, but with an approach that differs from standard implementations. In this paper, this method will be denoted GF–HMM. Initially, the aim of the GF–HMM method was to estimate the time progression of a gesture in real-time, using a template reference (Bevilacqua et al., 2012). Hence, similarly to DTW, this method uses the whole time series and assigns a state to each sample. This allows for the modeling of fine-temporal gesture structure (similarly to the approach of Bobick and Wilson in (Bobick and Wilson, 1997)). The system makes use of a forward procedure simultaneously on several template gestures, which allows for the estimation, during the gesture performance, of its time progression and likelihood related to each template.

The GF–HMM method fulfills the requirements 1) and 2) we noted in the Introduction. However, GF–HMM can take into account gesture variation in a limited way : the parameters variations are considered as noise around a fixed mean value. As we will show in this paper, an adaptive approach using an extended state model and a different decoding scheme is possible. This corresponds to considering the recognition problem as a tracking problem, where Particle Filtering (PF) techniques have been widely used, and prove effective to adapt continuously the shape of the tracked objects.

An exhaustive review of particle filtering literature is beyond the scope of this paper, and we refer the reader to (Arulampalam et al., 2002) and (Doucet et al., 2001) for more specific theoretical works on PF. For example, methods based on PF for tracking were used on hand gestures and faces (Bretzner et al., 2002; Zhou et al., 2004; Mitra and Acharya, 2007). In these previous works, PF is used to estimate the position of the considered area of importance in image sequences.

The method we propose is close to the work of Black et al. (Black and Jepson, 1998a), based on the *condensation* algorithm (Isard and Blake, 1998a), for the recognition of spatio-temporal gesture templates. The model was applied to data recorded using a 2-dimensional augmented whiteboard. The implementation allowed for the tracking of speed and scaling variation. It will be denoted PF–condensation.

In this paper, we generalize the approach by (Black and Jepson, 1998a), by estimating scaling and rotation invariance in both 2D and 3D gestures, and making different choices in the PF implementation. Since this method can be seen as an extension of our *Gesture Follower* system from a functional point of view, we will refer it as GF–PF.

8.3 Gesture Model and Recognition

In this work we define gestures as temporal series of a fixed number of parameters. For a given *input* gesture, the recognition task corresponds to selecting the best match among a set of pre-recorded *template* gestures. The input gesture is denoted $\mathbf{z} = \mathbf{z}_1 \dots \mathbf{z}_N$ (or $\mathbf{z}_{1:N}$) and the template gesture is denoted $\mathbf{g} = \mathbf{g}_1 \dots \mathbf{g}_T$ (or $\mathbf{g}_{1:T}$). $\mathbf{z}$ can be of different length than $\mathbf{g}$. As described in the next section, we use a Bayesian approach with a continuous state representation.

8.3.1 Continuous state model

The model can be formulated with the following dynamical system :

$$\begin{cases} \mathbf{x}_k &= f_{\text{TR}}(\mathbf{x}_{k-1}, \mathbf{v}_{k-1}) \\ \mathbf{z}_k &= f_{\text{OB}}(\mathbf{x}_k, \mathbf{w}_k; \mathbf{g}) \end{cases} \tag{8.1}$$

where, at discrete time k,

- $\mathbf{x}_k$ is a vector representing the *system state* ;
- f_{TR} is a (possibly non linear) function that governs the evolution of the system state, depending on $\mathbf{x}_{k-1}$ and an independent and identically distributed (i.i.d.) process noise sequence $\mathbf{v}_k$;
- f_{OB} is a (possibly non-linear) function that generates the *observations* $\mathbf{z}_k$, depending on the system state $\mathbf{x}_k$, an i.i.d. measurement noise sequence $\mathbf{w}_k$ and a template gesture $\mathbf{g}$.

The form of Equation (8.1) implies that the process to be tracked is Markovian, i.e. the system state at each instant depends only on the state at the previous instant.

The problem is thus formulated as a tracking problem, i.e. tracking and adapting the values of $\mathbf{x}_k$. Precisely, state variables $\mathbf{x}_k$ are gesture features to be chosen, as detailed in Section 8.3.2. They are estimated by comparing the incoming gesture $\mathbf{z}$ with a template gesture $\mathbf{g}$.

The particular form of transition between states, f_{TR}, is described in Section 8.3.3 and the observation function, f_{OB}, is described in Section 8.3.4.

The algorithm, based on particle filtering, infers the gesture features as described in Section 8.3.5. The extension of the tracking algorithm for the recognition task is detailed in Section 8.3.6.

8.3.2 State space model

The state of the system is composed of gesture features that have to be estimated over time. The process is adaptive since the features are updated at each time step. The system state at instant k is denoted as :

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k(1) \\ \vdots \\ \mathbf{x}_k(D) \end{pmatrix} \in \mathbb{R}^D$$

where D is the dimensionality of the state space.

The first dimension $\mathbf{x}_k(1)$ is set to be the *phase* p_k at discrete time k, which represents the alignment between the template gesture and the incoming gesture at time k, as illustrated in Figure 8.1 (or in other words, p_k can be seen as the time progression of the gesture). The phase is normalized in the [0,1] range (0 being the beginning and 1 the end of the gesture time).

The second dimension $\mathbf{x}_k(2)$ is set to be the *speed* v_k at k. The speed v_k is actually a speed ratio between the incoming gesture and the template gesture.

The state space can contain additional dimensions. In particular, we will extend, depending of the application, to other features such as the *scaling* (i.e. amplitude ratio, see Figure 8.1), and the *rotation* angles in 3-dimensions.

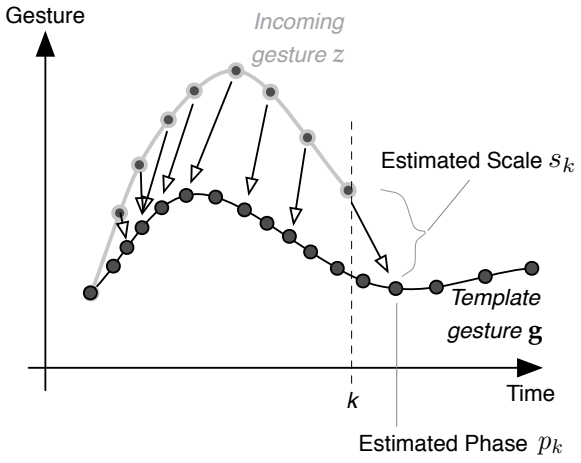

FIGURE 8.1 – Illustration of the alignment and adaptation. An incoming gesture **z** is aligned onto a template gesture **g** based on the continuous adaptation of gesture features $\mathbf{x}_k$ illustrated as p_k and s_k in the figure.

8.3.3 State transition

In the proposed model, the state transition function f_{TR} (see Equation (8.1)) is linear, given by the matrix A, and modeled probabilistically as a Gaussian distribution :

$$\begin{aligned} p(\mathbf{x}_k|\mathbf{x}_{k-1}) &= \mathcal{N}(\mathbf{x}_k|A\mathbf{x}_{k-1}, \Sigma) \\ \Sigma &= \text{diag}\left(\sigma_1 \dots \sigma_D\right) \end{aligned} \tag{8.2}$$

We choose to add a constraint, by setting a relationship between the phase and the velocity, corresponding to a first-order motion equation :

$$p_k = p_{k-1} + \frac{v_k}{T} + \mathcal{N}(0, \sigma_1) \tag{8.3}$$

where T is the template's length and σ_1 is the first element in the diagonal of Σ.

This constraint can be simply taken into account by setting the first row of the matrix A, corresponding to phase and speed, to $(1\ \frac{1}{T}\ 0 \cdots 0)$.

The other terms set to zero in the first row of the matrix A, implies that the estimation of the phase is independent of the other features $(\mathbf{x}_k(j), j > 2)$.

8.3.4 Observation function

In our model, the observation function f_{OB} (see Equation (8.1)) is chosen to be a Student's t-distribution that depends on three parameters : the mean $\boldsymbol{\mu}$, the covariance matrix Σ and the degree of freedom ν. For a K-dimensional input vector $\mathbf{z}_k$ at time k, the Student's t-distribution is as follows :

$$St(\mathbf{z}_k|f(\mathbf{x}_k, \mathbf{g}(p_k)), \Sigma, \nu) = C(\Sigma, \nu)\left(1 + \frac{d(\mathbf{z}_k, f(\mathbf{x}_k, \mathbf{g}))^2}{\nu}\right)^{-\frac{\nu+K}{2}} \tag{8.4}$$

where

$$C(\Sigma, \nu) = \frac{\Gamma(\nu/2 + K/2)}{\Gamma(\nu/2)} \frac{|\Sigma|^{-1/2}}{(\nu\pi)^{K/2}}$$

where $f(\mathbf{x}_k, \mathbf{g})$ is a function of the template **g** and the state value at k. Precisely, $f(\mathbf{x}_k, \mathbf{g})$ adapts the expected template sample $\mathbf{g}(p_k)$, given the phase p_k at k. Examples are given in the following Sections 8.4 and 8.5. The distance d between the adapted template sample and the

incoming observation is given by :

$$d(\mathbf{z}_k, f(\mathbf{x}_k, \mathbf{g})) = \sqrt{[\mathbf{z}_k - f(\mathbf{x}_k, \mathbf{g})]^T \Sigma^{-1} [\mathbf{z}_k - f(\mathbf{x}_k, \mathbf{g})]} \tag{8.5}$$

The choice of Student's t-distribution is motivated by its heavier tails compared to Gaussian distribution (i.e. the distribution is wider around the mean). This choice will be justified in the results section 8.5. In the limit $\nu \to \infty$, the t-distribution reduces to a Gaussian with mean $\boldsymbol{\mu}$ and covariance Σ.

8.3.5 Inference and algorithm for the alignment and adaptation

Sequential Montecarlo methods work by recursively approximating the current distribution of the system state using the technique of Sequential Importance Sampling : state samples are drawn from a simpler distribution and then weighted according to their importance in estimating the "true" distribution. Importance is driven by incoming samples.

The algorithm is summarized in Algorithm 1. Here we denote by N_s the number of particles used to approximate the distribution. We denote $\mathbf{x}_k^i$ the i^{th} state sample drawn and w_k^i its respective weight at time k. The weights are normalized such as $\sum_{i=1}^{N_s} w_k^i = 1$.

The set of support points and their associated weights $\{\mathbf{x}_k^i, w_k^i\}_{i=1}^{N_s}$ is a random measure used to characterize the posterior *pdf* $p(\mathbf{x}_k|\mathbf{x}_{0:k-1}, \mathbf{z}_{1:k})$. The continuous "true" state distribution can be approximated with a series of weighted Dirac's Delta functions :

$$p(\mathbf{x}_k|\mathbf{x}_{0:k-1}, \mathbf{z}_{1:k}) \approx \sum_{i=1}^{N_s} w_k^i \delta(\mathbf{x}_k - \mathbf{x}_k^i)$$

The term $\mathbf{x}_0$ represents the prior distribution (i.e., the initial state), and the posterior distribution is updated at each time step. Finally, the expected value $\hat{\mathbf{x}}_k$ of the resulting random measure is computed as :

$$\hat{\mathbf{x}}_k = \sum_{i=1}^{N_s} w_k^i \mathbf{x}_k^i$$

An optional resampling step is used to address the *degeneracy* problem, common to particle filtering approaches, as discussed in details in (Arulampalam et al., 2002; Douc and Cappé, 2005) . Resampling is introduced because after a few iterations of the inference algorithm, only a few particles have non-negligible weights (it can be shown that the variance of the importance weights can only increase over time). The resampling step corresponds to draw the particles according to the current distribution $\{w_k^i\}_i^{N_s}$. Intuitively, resampling replaces a random measure of the true distribution with an equivalent one (in the limit of $N_s \to \infty$).

In (Black and Jepson, 1998b) Black et al. choose to randomly select 5 to 10% of particles to be replaced by randomly taken initial values. This process is performed during transition and may introduce discontinuities. In our approach, the degeneracy problem is handled by defining a criterion based on effective sample size N_{eff}, as specified by Arulampalam (Arulampalam et al., 2002) :

$$N_{eff} = \frac{1}{\sum_{i=1}^{N_s} (w_k^i)^2}$$

where N_{eff} is an estimate of the effective sample size, i.e. an approximation of the number of particles that are contributing significant information to the estimation of the posterior *pdf*. The N_{eff} value is used as a criterion to operate the resampling step as shown in the Algorithm 1.

Algorithm 1: Realtime temporal alignment (step at time k with observation $\mathbf{z}_k$).

for $i = 1 \dots N_s$ **do**
- $\mathbf{x}_k^i \sim \mathcal{N}(\mathbf{x}_k | A\mathbf{x}_{k-1}, \Sigma)$
- $p_k^i := \mathbf{x}_k^i(1)$
- $p(\mathbf{z}_k|\mathbf{x}_k^i) = St(\mathbf{z}_k | f(\mathbf{x}_k^i, \mathbf{g}(p_k^i)), \Sigma, \nu)$
- $\hat{w}_k^i \leftarrow w_{k-1}^i p(\mathbf{z}_k|\mathbf{x}_k^i)$

$w_k^i \leftarrow \dfrac{\hat{w}_k^i}{\sum_j \hat{w}_k^j}, \quad \forall i = 1 \dots N_s$

$N_{eff} \leftarrow (\sum_{i=1}^{N_s} (w_k^i)^2)^{-1}$

if $N_{eff} <$ *resampling threshold* **then**
- resample $\mathbf{x}_k^1 \dots \mathbf{x}_k^{N_s}$ according to ddf $w_k^1 \dots w_k^{N_s}$ $w_k^i \leftarrow N_s^{-1} \quad \forall i = 1 \dots N_s$

return $\hat{\mathbf{x}}_k = \sum_{i=1}^{N_s} w_k^i \mathbf{x}_k^i$

8.3.6 Recognition

The inference described in the previous section can be extended for recognition as follows. Consider M templates of respective length $L_1 \dots L_M$ denoted $\mathbf{g}^1 \dots \mathbf{g}^M$. At the initialization, we assign to each state particle $\mathbf{x}_k^i$ a *gesture index* between $1 \dots M$ (denoted m_k), based on a initial distribution. Generally, a uniform distribution is chosen, that is of distributing the particles evenly across the gesture templates. This leads to extend the state configuration, applied to each particle, as follows :

$$\mathbf{x}_k^i = \begin{pmatrix} \mathbf{x}_k^i(1) \\ \vdots \\ \mathbf{x}_k^i(D) \\ m_k \end{pmatrix} \in \mathbb{R}^D \times \mathbb{N} \tag{8.6}$$

The transition probability is then adapted as follows :

$$\begin{aligned} p(\mathbf{x}_k|\mathbf{x}_{k-1}) &= \mathcal{N}(\mathbf{x}_k | A\mathbf{x}_{k-1}, \Sigma) \\ \Sigma &= \text{diag}\left(\sigma_1 \dots \sigma_D \, 0\right) \end{aligned} \tag{8.7}$$

The last element in the diagonal of Σ is set to 0 since no noise is added to the gesture index. By summing the weights w_k^i depending on the corresponding particles' gesture index, it is straightforward to compute the probability of each gesture :

$$p(\mathbf{g}_k^l | \mathbf{g}_k^m) = \sum_{j \in \mathcal{J}} w_k^j \quad , \forall l \in [1, M], \forall m \in [1, M], m \neq l$$

$$\text{where} \quad \mathcal{J} = \left\{ j \in [1, N_s] / \mathbf{x}_k^j(D+1) = l \right\} \tag{8.8}$$

8.3.7 Computational cost and precision

Due to the statistical nature of the Sequential Monte Carlo method, the inference precision increases with the number of particle N_s. Nevertheless, the computational cost is linear with the number of particles, i.e. $O(N_s)$. If we denote N_{spg} the number of particles distributed per template gesture, the cost can also be written $O(MN_{spg})$ ($N_s = MN_{spg}$), where M is the number of template gestures.

8.4 Assessment on Synthetic Data

In this section we present an evaluation of GF–PF for synthetic data. First we assess the temporal alignment between the incoming gesture and a single template, and compare it with the GF–HMM method based on a forward computation (Bevilacqua et al., 2010). We consider two different cases. In the first one, only the phase and the scaling are adapted in the inference. In the second case, the rotation angles are also adapted.

In both cases, we consider synthetic data obtained from the following curve (Viviani's curve).

$$\mathbf{C}(t) = \begin{cases} x(t) &= a(1+\cos(t)) \\ y(t) &= a\sin(t) \\ z(t) &= 2a\sin(t/2) \end{cases} \tag{8.9}$$

Figure 8.2 depicts the 3D time series $(x(t), y(t), z(t))$ with $a = 1$ and t spanning $[0, 4\pi]$.

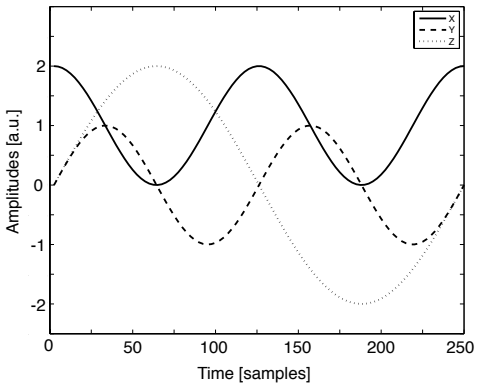

FIGURE 8.2 – The synthetic curve follows the parametric form of the so-called Viviani's curve (Equation (8.9)), for $a = 1$, $t \in [0, 4\pi]$.

8.4.1 Temporal Alignment Assessment

We define two different curves for the test and template data. The template gesture is obtained by a regular sampling of the curve described by Equation (8.9), and the input gesture is obtained by a non-linear sampling of the same function ($t \mapsto t^3$), and by adding a uniformly distributed noise. We denote with $\mathbf{C}$ and $\hat{\mathbf{C}}$ the original and the resampled curves, respectively.

$$\hat{\mathbf{C}}(t) = \mathbf{C}(t^3) + \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{C}}) \tag{8.10}$$

Model configuration

For this first case, we used state space defined as a three-dimensional vector, consisting of the phase p_k, the speed v_k and the scale s_k (this model is similar to the one presented in (Black and Jepson, 1998a)) :

$$\mathbf{x}_k = \begin{pmatrix} p_k \\ v_k \\ s_k \end{pmatrix} \in (0,1) \times \mathbb{R}^2$$

The phase feature p_k lies in the interval $[0, 1]$. The velocity v_k and scale s_k are normalized, a value of 1 corresponding to the speed (resp. scale) of the template.

The transition matrix A between states was set to :

$$A = \begin{pmatrix} 1 & 1/T & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The f function involved in the distance function for the observation likelihood (cf. Equation (8.5)) is

$$f(\mathbf{x}_k, \mathbf{g}) = \text{diag}(s_k)\mathbf{g}(p_k)$$

where $\text{diag}(s_k)$ is the diagonal matrix, of size 3×3, which elements are equal to the scaling s_k. The scaling coefficient is identical for all three input observations $x(t)$, $y(t)$, $z(t)$, (homothetic transformation).

For the comparison with the hybrid GF–HMM model we set $\nu \to \infty$ leading to a gaussian distribution whose standard deviation is σ (which is the same the distribution used by the GF–HMM model). The influence of the σ value will be discussed in Section 8.5.

Alignment Results

We report here the results concerning the estimation of the phase p_k, that describes the alignment between the test and template data. For each test, our model returns the estimated phase p_k which should ideally follow the cubic function that was used to synthesize the curve $\hat{\mathbf{C}}(t)$. From this, we calculated the mean square error between $p_k(t)$ and the ground-truth cubic function. The number of particles was set to $N_s = 1800$.

The result is illustrated in Figure 8.3, where the estimated phase p_k is plotted along the cubic function. The estimated p_k is close to the expected curve, with a average error of 0.6% ($\sigma_{\mathbf{C}} = 0.1$).

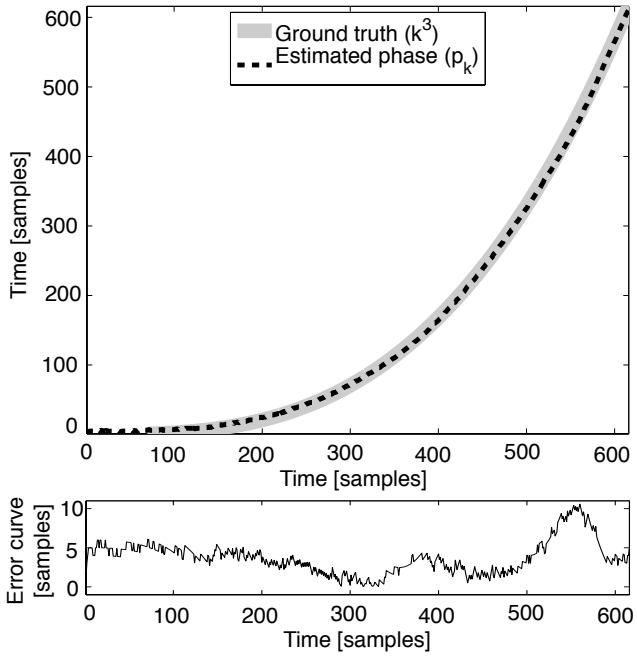

FIGURE 8.3 – Example of the estimated phase p_k (dashed black line) compared to the ground truth defined as a cubic function (gray solid line). The example is obtained with a observation likelihood with standard deviation $\sigma = 0.1$ (equals to the additive gaussian noise used for the test).

As this was obtained for a given σ value, set to the gaussian noise value ($\sigma = \sigma_{\mathbf{C}} = 0.1$), we further examined the influence of this parameter. We varied the σ value between 0.02 and 0.4 (with a step of 0.01), for both GF–PF and GF–HMM.

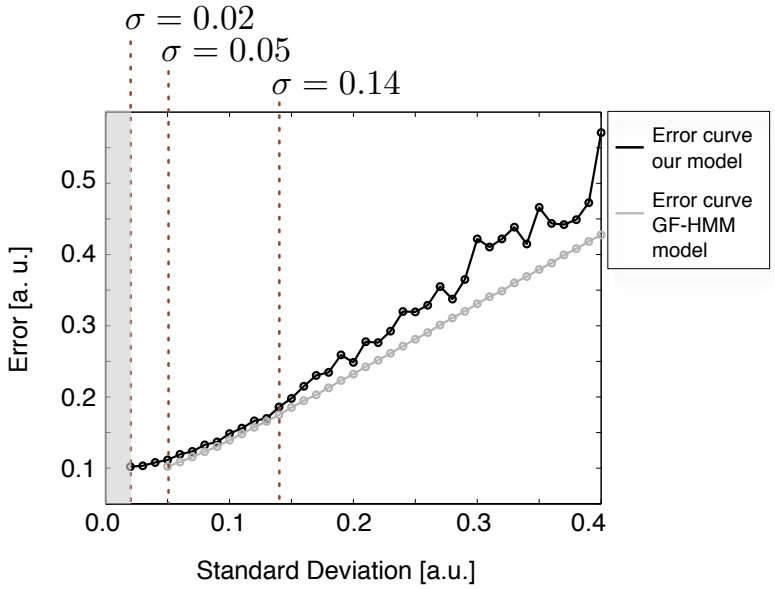

FIGURE 8.4 – Error curves obtained from both our model (black line) and GF–HMM (gray line). The error is computed from the mean square error between the ground truth cubic function and the estimated phase.

Figure 8.4 reports the results. The solid gray line is the error curve from our model and the dashed curve is from the GF–HMM model. First, we found that for σ values between 0.02 and 0.14, the errors for both models are close. This confirmed that our model, based on approximative inference (particle filters), can obtain similar results to the exact inference of GF–HMM (forward procedure).

Note that the σ values are close to the standard deviation of the noise in the test data $\sigma_{\mathbf{C}} = 0.1$. Interestingly, our model can provide equally accurate results with lower values (between 0.02 and 0.05), where the GF–HMM fails. This can be explained by the fact that the p_k and v_k values are linked through a first order motion equation, while such a constraint is not taken into account in the GF–HMM model. In other words, the phase p_k estimation is made more robust by the joint estimate of the speed v_k. This allows for the extension of the model to low p_k, where the most accurate results are obtained.

For higher σ values, the error obtained by our model increases, and is found to be higher compared to exact inference of the GF–HMM model. In this case, the estimation of probability distribution by the particle filtering is suboptimal.

8.4.2 Rotation matrix adaptation

We examine in this section the case where gesture rotation angles dynamically vary over time. Precisely, we consider the three angles ϕ, θ, ψ around x, y, z, respectively, in a Cartesian coordinate system.

The three angle time series are defined as follows :

$$\begin{cases} \phi(t) = t^2 \\ \theta(t) = t \\ \psi(t) = -t^{1/3} \end{cases} \tag{8.11}$$

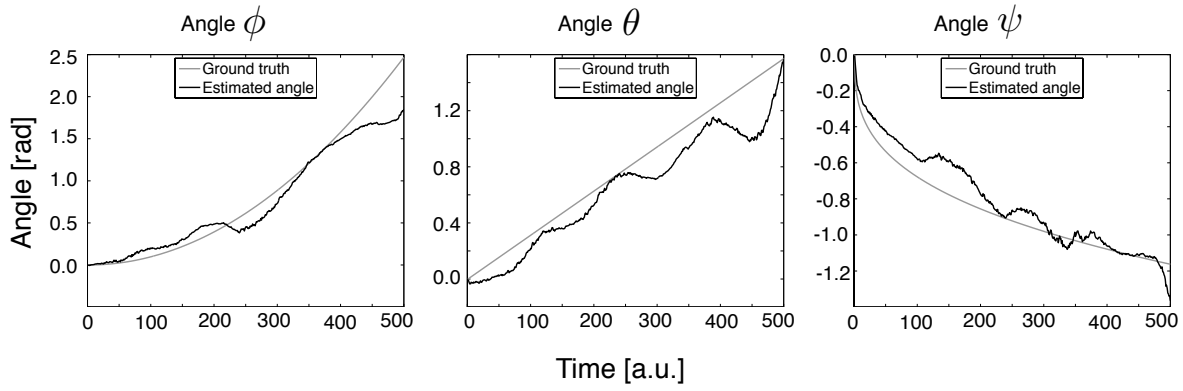

FIGURE 8.5 – Dynamic rotation estimation. The input curve is the one given by Equation 8.11 with $\sigma_{\mathbf{C}} = 0.1$. The model is configured to estimate the phase p_k, the speed v_k and the three angles ϕ_k, θ_k, ψ_k. The standard deviation used in the model is 0.1.

The 3-dimensional template curve **C** (equation (8.9)) is rotated according to this matrix in the Cartesian frame (x, y, z). The input curve is the rotated version of **C** with additive gaussian noise :

$$\hat{\mathbf{C}}(t) = R(\phi(t), \theta(t), \psi(t))\mathbf{C}(t) + \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{C}})$$

The rotation matrix $R(\phi(t), \theta(t), \psi(t))$ is computed at each time step. More details on the conventions for angles and rotation in the 3-dimension Cartesian frame are reported in Appendix 8.7.

Model configuration

The state space is defined as a 5-dimensional vector that consists of the phase p_k, velocity v_k, and the angles ϕ_k, θ_k, ψ_k. The state variable at time k is :

$$\mathbf{x}_k = \begin{pmatrix} p_k \\ v_k \\ \phi_k \\ \theta_k \\ \psi_k \end{pmatrix} \in (0, 1) \times \mathbb{R}^4$$

With the following transition matrix :

$$A = \begin{pmatrix} 1 & 1/T & 0 & 0 & \dots \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & & & \\ 0 & 0 & & I_3 & \\ \vdots & \vdots & & & \end{pmatrix}$$

Where I_3 is the identity matrix of size 3×3. Finally the observation likelihood is entirely defined by the following f function :

$$f(\mathbf{x}_k, \mathbf{g}) = R(\phi_k, \theta_k, \psi_k)\mathbf{g}(p_k)$$

Rotation Estimation Results

Figure 8.5 shows an example where the estimated angles ϕ_k, θ_k, ψ_k are plotted with the ground truth (defined by Equation (8.11)). In this example, the standard deviation σ is set to

the standard deviation of the input data ($\sigma = \sigma_{\mathbf{C}} = 0.1$). The number of particles was set to $N_s = 2500$.

We tested the effect of the standard deviation σ by varying its value between 0.01 to 0.4 (step$= 0.01$).

For each value of σ, the mean square error between estimation of the angle ϕ, θ, ψ and the ground truth angles are reported in Figure 8.6.

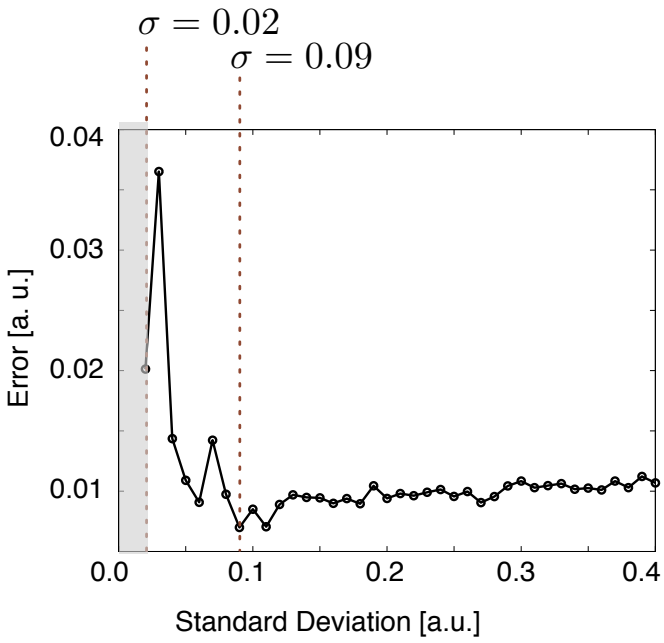

FIGURE 8.6 – Error curve obtained from our model. The error is computed from the mean square error between the ground truth angles and the estimated ones.

For $\sigma > 0.09$, the standard deviation has a very weak influence on the angle accuracy. As expected the optimal value corresponds to the standard deviation of the noise ($\sigma_{\mathbf{C}} = 0.1$). For $\sigma < 0.02$, the computation fails due to computational errors, i.e. at a certain time k all the weights w_k^i are equal to 0.

8.5 Recognition Tasks on User Data

In this section, we present two different evaluations of the GF–PF method for a recognition task performed using experimental data. Precisely we use the 2-dimensional pen gesture dataset from Wobbrock et al. (Wobbrock et al., 2007) and a 3-dimensional accelerometer-based dataset adapted from Liu et al. (Liu et al., 2009). The results are compared with various methods published recently.

8.5.1 Experiment #1 : 2D Pen gestures

In this experiment, we consider the case of two-dimensional pen gestures presented in (Wobbrock et al., 2007), and we replicate the assessment methodology.

Database

The database was introduced by Wobbrock et al. in (Wobbrock et al., 2007), and is available online [1]. It contains 16 gestures that are meant to be commands for selection, execution and entering symbols in HCI applications. Ten participants have been recruited to perform the gestures. For each one of the 16 gestures in the vocabulary (figure 8.7), "subjects entered one practice gesture before beginning three sets of 10 entries at slow, medium, and fast speeds" (Wobbrock et al., 2007). Hence, the whole database contains 4800 gesture examples.

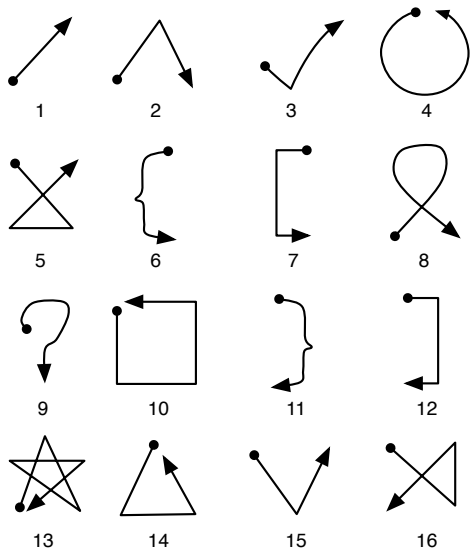

FIGURE 8.7 – Gesture vocabulary from (Wobbrock et al., 2007).

Model Configuration

In Wobbrock et al. (Wobbrock et al., 2007), the authors propose a pre-processing step that consists of rotating, scaling and translating before applying different recognition algorithms. Both the rotation angle and the scaling coefficient are considered to be invariant in the recognition process. The GF–PF method allows for taking into account these invariants by defining them as state variables, s_k and r_k respectively. The gesture features estimated are the following : position p_k, velocity v_k, scaling coefficient s_k, rotation angle r_k, and the gesture index $m_k \in [1 \dots 16]$:

$$\mathbf{x}_k = \begin{pmatrix} p_k \\ v_k \\ s_k \\ r_k \\ m_k \end{pmatrix} \in (0,1) \times \mathbb{R}^3 \times \mathbb{N}$$

The invariance by rotation and scaling leads to the following non linear function of state variables :

$$f(\mathbf{x}_k, \mathbf{g}(p_k)) = \text{diag}(s_k) \begin{pmatrix} \cos(r_k) & -\sin(r_k) \\ \sin(r_k) & \cos(r_k) \end{pmatrix} \mathbf{g}(p_k)$$

1. Database available at : http://depts.washington.edu/aimgroup/proj/dollar/

	$1 **recognizer**	**DTW**	**GF–HMM**	**GF–PF**
	offline operated after scaling and rotation estimation		**online** no adaptation of scaling neither rotation	**online** incremental adaptation of scaling and rotation
Mean	97,27 %	97,86 %	95,78 %	**98.11** %
Std	2,38 %	1,76 %	2,06 %	2,35 %

TABLE 8.1 – Results obtained on a unistroke gesture database presented in (Wobbrock et al., 2007). Our model has the following parameterization : $\sigma = 130$, $\nu = 0.1$.

The state transition matrix A_l, for the template gesture index $l \in [1, M]$ is given by :

$$A_l = \begin{pmatrix} 1 & 1/T_l & 0 & 0 & \dots \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & & & \\ 0 & 0 & & I_3 & \\ \vdots & \vdots & & & \end{pmatrix}$$

where T_l is the length of the l-th gesture template.

Evaluation Procedure

The evaluation procedure is directly taken from Wobbrock et al. (Wobbrock et al., 2007), essentially based on a statistical "leave-one-out" approach. One template per gesture is randomly chosen from the 10 trials, and one test example is chosen randomly from the remaining trials. This process is repeated 100 times.

The GF–PF and GF–HMM methods were also evaluated using this same procedure, and compared with the results of the methods reported in Wobbrock et al. (Wobbrock et al., 2007), namely the $1 recognizer and DTW. The number of particles used in the evaluation is $N_s = 2000$.

Recognition Results

The results are reported in Table 8.1, where the GF–PF method is compared with three other methods (using mean and standard deviations for the recognition rate).

Two are offline methods, $1 recognizer and DTW, both operated after a pre-processing step correcting for the variations in scaling and rotation. The other two methods, GF–PF and GF–HMM are online methods, reporting results while the gestures are performed. Contrary to GF–HMM, GF–PF can adapt incrementally the dynamic scaling and rotation variations.

Comparing first on-line methods, the results show that the GF–PF gives better results than the GF–HMM method ($98,11\%$ vs $95,78\%$), which is mainly due to the fact that GF–PF can adapt the scaling and rotation. Compared to the off-line methods, the GF–PF method gives slightly better results to the $1 recognizer and DTW. These results are thus consistent with what was expected, confirming that the incremental adaptive correction in GF–PF is effective, and that the results we obtain with this on-line method can be at least equivalent to standard off-line methods that include similar corrections.

These results were obtained for a fixed set of the standard deviation σ and the Student's ν parameters (used in the observation function, section 8.3.4). We detail here how these parameters actually influence the recognition accuracy. Precisely, we report the recognition rate for a large set of σ values (from 10 to 150 with step$= 10$) and ν values (0.5, 1.0, 1.5 and ∞ = Gaussian

distribution). The variability of the recognition rate is plotted in Figure 8.8, superimposed to the results obtained with the $1 recognizer and DTW (which are methods that do not depend on such parameters)

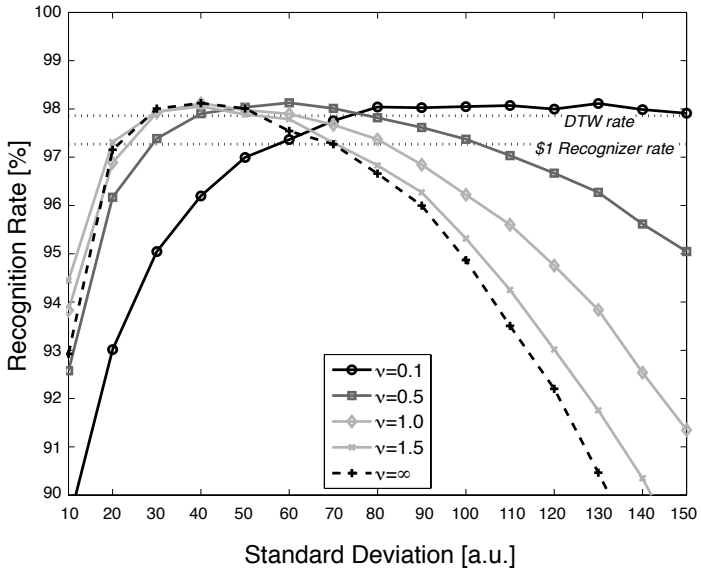

FIGURE 8.8 – Recognition rates obtained from the 2-dimensional pen gesture database by evolving observation distribution (defined as a Student's t-distribution) parameters σ and ν. While σ evolves from 10 to 150 with a step= 10, ν takes 4 values : $0.5, 1.0, 1.5$ and ∞.

Two important points must be noted. First, as expected, the best recognition rate is obtained for a restricted range of σ and ν values. Nevertheless, the recognition variability varies smoothly with these parameters, with no local maxima. Second, the Student's t-distribution is advantageous to the Gaussian distribution ($\nu = \infty$), since it significantly reduces the sensitivity of the recognition rate to the σ values. Globally, these results show that data specific training procedure might not be required, since the recognition remains optimal over a large range of the Student's t-distribution parameters.

8.5.2 Experiment #2 : 3D gestures sensed using accelerometers

This section presents a second experiment using 3D hand gestures, captured with handheld accelerometers. One of the frequent difficulties of using gesture recognition with such systems resides in the user variations in handling the devices, which introduces offset, scaling and rotation of the 3D accelerometer signals. Since the GF–PF model can potentially handle these variations, such use cases are important for the evaluation of our approach.

Database

The vocabulary considered contains eight different gestures reported in Figure 8.9. This gesture vocabulary was previously proposed for mobile systems (Kela et al., 2006; Liu et al., 2009).

We collected data from four participants with the following procedure. We used the 3-D embedded accelerometers in a mobile phone. Each participant was asked to perform the eight gestures with five different orientations. Precisely, each orientation corresponded to handle the interface with different angles, approximately $0, \pi/6, \pi/4, \pi/3, \pi/2$, as shown in Figure 8.10 for the circle (gesture 3). Each orientation was perfomed twice, leading to a database with a total of 320 gestures. Each start and end of each gesture was set by user, pushing a foot pedal.

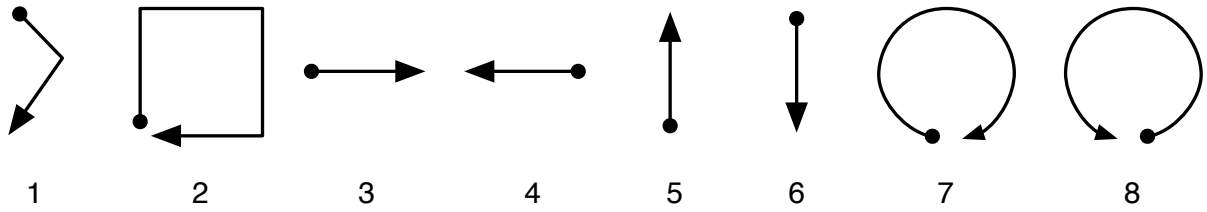

FIGURE 8.9 – Gesture vocabulary taken from (Liu et al., 2009).

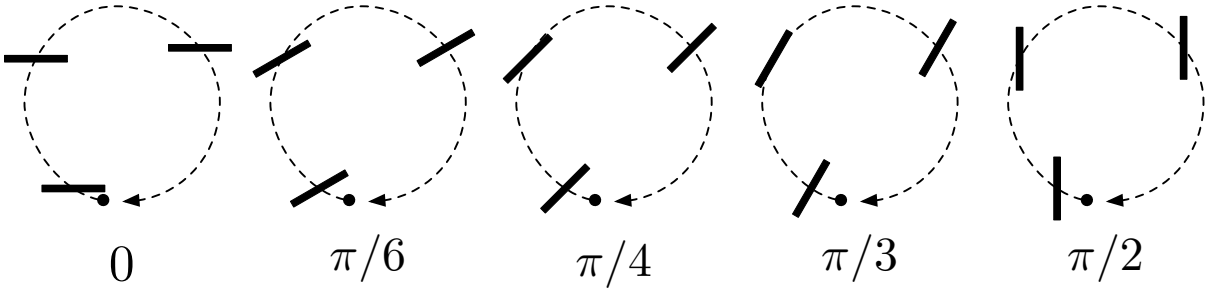

FIGURE 8.10 – Five different orientations of the interface. Example on gesture 7 drawing a circle.

An example of 3D accelerometer data is shown in Figure 8.11. In this figure, we report gesture 1, performed with two different orientations : angle 0 and $\pi/2$. One can see that the dashed curve (z-axis) in the upper plot (angle 0) is the opposite of the solid black line (x-axis) in the lower plot ($\pi/2$). In this case, the change of orientation leads to a complete change of the data along the three axis of the accelerometer data.

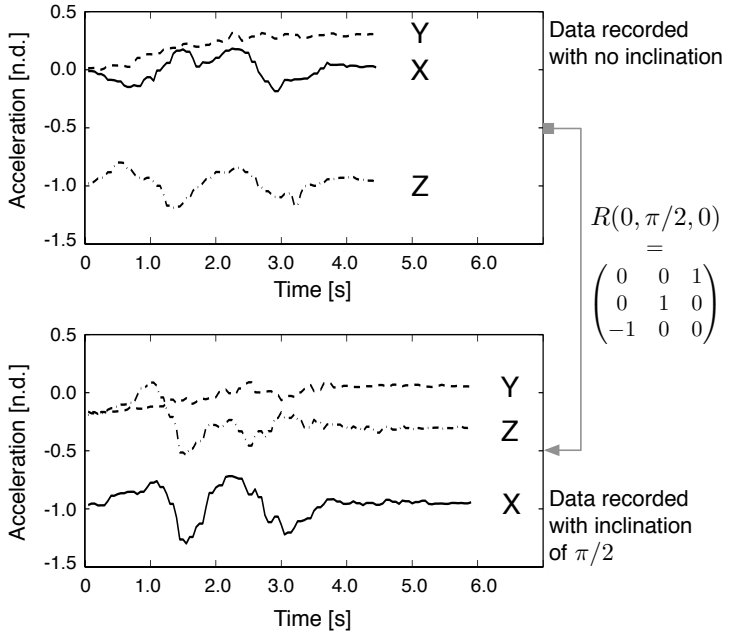

FIGURE 8.11 – Example of gesture data from the database : gesture 1 performed with two orientations. Upper plot : orientation $\phi = \theta = \psi = 0$. Lower plot : orientation $\phi = 0, \theta = \pi/2, \psi = 0$.

Model configuration

In this case, both the scaling coefficients (three scaling coefficients s^x, s^y, s^z) and the rotation angles (three angles ϕ, θ, ψ) must be taken into account in the recognition process. As in the previous experiment, the GF–PF model allows for taking into account these invariants by incoporating them as state variables : $\mathbf{x}_k = \big(p_k, v_k, s_k^x, s_k^y, s_k^z, \phi_k, \theta_k, \psi_k, m_k\big)$.

The invariance by rotation and scaling leads to the following non linear function of state variables :

$$f(\mathbf{x}_k, \mathbf{g}(p_k)) = \begin{pmatrix} s_k^x & 0 & 0 \\ 0 & s_k^y & 0 \\ 0 & 0 & s_k^z \end{pmatrix} R(\phi_k, \theta_k, \psi_k)\mathbf{g}(p_k)$$

where $R(\phi_k, \theta_k, \psi_k)$ is the rotation matrix in three dimensions given by the Euler angles. As previously, we refer the reader to the Appendix 8.7 for the rotation conventions.

The state transition matrix A_l depends on the gesture template l and is written :

$$A_l = \begin{pmatrix} 1 & 1/T_l & 0 & 0 & \dots \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & & & \\ 0 & 0 & & I_7 & \\ \vdots & \vdots & & & \end{pmatrix}$$

where I_7 is the identity matrix of size 7×7 and T_l is the length of the l-th template gesture.

Test Procedure

The evaluation presented here consists in the comparison of the recognition accuracy between different methods : the uWave algorithm defined in (Liu et al., 2009) based on a Dynamic Time Warping, the GF–HMM method, the GF–PF method and a particle filtering method (based on condensation) that takes into account only scaling invariance, described in (Black and Jepson, 1998a) and referred here to PF–condensation.

The test procedure is identical with the one proposed by Liu et al. for testing their uWave algorithm ((Liu et al., 2009)), corresponding to a "leave-one-out" cross-validation design : At the i-th test, we use the i-th trial from each gesture as reference and the remaining examples belong to the testing set.

The total number of particles is $N_s = 4000$.

Recognition Results

The results are reported in Table 8.2, where the recognition results can be compared between all methods. The means show that the GF–PF method obtains a significantly higher success rate than the other methods, with a mean rate of 81.5%. This result can be explained by the fact that GF–PF is the only method that takes into account scaling and rotation invariance. Such an advantage is therefore critical in accelerometer data where the device orientations are not specified and leading to large scale and angle variations.

Comparing the other methods, we note that as expected, the offline method uWave, based on DTW obtains better results (61.5%) than GF–HMM (51.3%) and the PF–condensation method (55.6%) ((Black and Jepson, 1998a)) which are online methods.

The difference between offline and online methods are even more striking when inspecting the confusion matrix. For example, uWave method (offline) obtained a higher classification rate for gesture 5 than with all the other online methods because of a confusion occurs between gesture 5 and 2 which both start identically. Another problematic gesture for the online

	uWave	GF–HMM	PF-Condensation	GF–PF
	Offline	Online	Online Scale adaptation	Online Offset/Rotation adaptation $\sigma = 1.7, \nu = 0.1$
Gesture 1	54.4 %	46.1 %	50.8 %	86.7 %
Gesture 2	51.9 %	35.6 %	28.6 %	91.1 %
Gesture 3	46.9 %	43.6 %	55.0 %	70.0 %
Gesture 4	46.9 %	56.7 %	56.9 %	83.6 %
Gesture 5	78.6 %	40.3 %	53.1 %	61.7 %
Gesture 6	53.9 %	42.5 %	53.9 %	78.9 %
Gesture 7	71.1 %	76.9 %	73.3 %	89.4 %
Gesture 8	87.8 %	68.6 %	72.8 %	90.8 %
Mean	**61.5** %	**51.3** %	**55.6** %	**81.5** %
Std	15.6 %	14.7 %	14.0 %	10.7 %

TABLE 8.2 – Results of the recongition process for different methods : uWave algorithm defined in (Liu et al., 2009) based on a Dynamic Time Warping, the GF–HMM method, the GF–PF method and the PF-based model described in (Black and Jepson, 1998a). The GF–PF method was used with $\sigma = 1.7, \nu = 0.1$ (for more details see Figure 8.12).

methods GF–HMM and PF–Condensation is gesture 2 that is misclassified with gestures 3, 4, 7 and 8 for GF–HMM and with gestures 3, 4 or 5 for the PF-Condensation.

The most problematic gestures with GF–PF are gestures 3, 4, 5 and 6 that can be obtained through a rotation of one of the other. This explains most of the errors encountered with the GF–PF.

As for previous evaluation on 2D pen gesture, we also report the influence of the standard deviation σ and the Student's ν on the recognition rate of the GF–PF method. Figure 8.12 shows the recognition rates with the following parameters ranges : σ value from 0.5 to 7.5 and ν value from 0.5 to ∞ (Gaussian distribution).

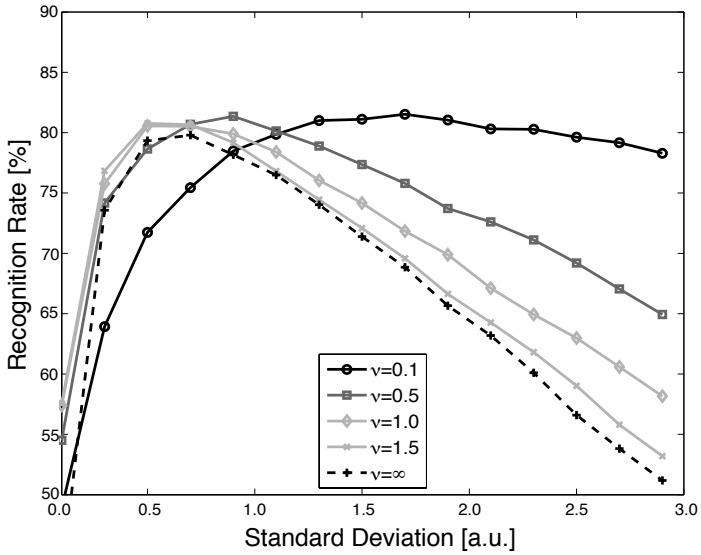

FIGURE 8.12 – Recognition rates using GF–PF on 3-dimensional accelerometers database with different observation distributions. The σ parameter varies from 0.5 to 7.5 and the ν parameter varies from 0.5, to ∞.

The results are very similar to the case of the 2-dimensional gestures : the optimal recognition rate varies smoothly with these parameters, with no local maxima. Large ν tends to lead to slightly higher recognition rate, and less sensitivity to the σ values. Overall, these results confirmed that finding the range of optimal values for the recognition is large.

8.6 Discussion and Conclusion

We described a method, called GF–PF, for gesture recognition that can adapt, in real-time, to variations occurring in the performance. We essentially demonstrated in this paper the case of phase, scaling and rotation adaptations. Nevertheless, the extension to other type of invariance is straightforward using the same formalism. Also note that the feature adaptation can be defined in a flexible manner : the feature initial values can be chosen arbitrary in an interval or even several distinct intervals.

By design, GF–PF belongs to the template-based methods. Therefore, we compared it with similar approaches such as DTW or more recent methods such as the $1 recognizer (Wobbrock et al., 2007) or the GF–HMM (Bevilacqua et al., 2010) (and thus we chose not to perform comparisons with other methods that require training with a large number of examples).

Globally, we found that our method always gave either equal or better recognition results than all the other methods. This can be explained by the fact that the GF–PF algorithm adapts dynamically to large differences between the gesture performance and the templates. This dynamic adaptation is particularly important since it is aimed to be used where the gesture classes are defined using single templates. In such cases, the expected gesture variations can be modeled explicitly, as it is the case with standard methods relying on large training database (e.g. HMM).

This was supported by both evaluations. In the first one, we compared the GF–PF with the $1 recognizer and DTW implementation of Wobbrock et al. (Wobbrock et al., 2007), that also correct for scale and rotation invariance. In this case, the GF–PF performs very closely to these methods, while GF–HMM, not taking into account such invariance, performs worse.

The difference is even more striking in the second evaluation with 3D gestures, where the GF–PF performed significantly better than all the other methods (DTW, GF-HHM, PF), being the only one taking into account both scale and rotation invariances. The PF-condensation method, with only scale invariance, gave intermediate results between the GF–PF method and the DTW or GF–HMM that do not take into account neither scale nor rotation invariances.

Compared to the PF-condensation and GF-HMM methods that use a Gaussian distribution for observation likelihood function, the GF–PF method uses a Student's t-distribution. This choice significantly reduces the sensitivity of the standard deviation parameter σ. Since this parameter which might *a priori* be difficult to estimate with limited training data, the use of Student's t-distribution can broaden the applicability of the method.

The results obtained with GF–PF are remarkable considering the fact that it operates in a causal manner : the recognition results and the parameters adaptation are updated each time a new sample is received. On the contrary, standard recognition schemes compute the results only once the gesture is finished (as DTW, Rubine or $1 recognizer), which generally allows for more robust decoding algorithm.

Thus, the fact that the GF–PF performance can equal or outperform non-causal methods demonstrates that the causal inference is robust, thanks to the coupling imposed between the phase and velocity estimation. Precisely, the phase and velocity are coupled through a kinematic model (similarly to a Kalman filter, see 8.3.3). This forces the tracking to be continuous along the state sequence, avoiding thus possible instability generally occurring in standard causal inference method (such as in the forward computation of hidden Markov model).

Note that a causal inference represents a clear advantage for applications in Human-Machine Interaction, since partial results are available during the gestures (and not only once

the gestures are finished). This would allow for example for the use of the current estimation of the scaling as a control parameter during the gesture, or to anticipate which gesture is currently performed (we call "early" gesture recognition in our applications) as described by Bau et al. (Bau and Mackay, 2008) and Appert et al. (Appert and Bau, 2010).

In summary, the GF–PF method we proposed represents a clear improvement over the GF–HMM algorithm we developed previously. Our contribution concerns the use of a general formalism based on a particle filtering inference, allowing for the online adaptation of gesture features. This formalism is more general than the work of Black et al. (Black and Jepson, 1998a), using a different scheme to avoid particles degeneracy, and a different observation likelihood function.

Our current applications for the expressive control of digital media (sound and visuals), with a particular emphasis on early recognition capabilities, will directly benefit from this work. Moreover, the adaptive gesture features estimation can be used to estimate parameters describing *how* a gesture is performed. This should represent a step forward for the design of human-machine systems taking into account the gesture expressivity.

8.7 Annex

Rotation matrix convention

Let us consider the Cartesian frame (x, y, z), the three Euler angles ϕ, θ, ψ rotating vectors about respectively x, y and z induce the three following rotation matrices :

$$R_\phi = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\phi) & -\sin(\phi) \\ 0 & \sin(\phi) & \cos(\phi) \end{pmatrix}$$

$$R_\theta = \begin{pmatrix} \cos(\theta) & 0 & \sin(\theta) \\ 0 & 1 & 0 \\ -\sin(\theta) & 0 & \cos(\theta) \end{pmatrix}$$

$$R_\psi = \begin{pmatrix} \cos(\psi) & -\sin(\psi) & 0 \\ \sin(\psi) & \cos(\psi) & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The rotation matrix in 3-dimension considered in the paper the clockwise rotation defined as : $R = R_\phi R_\theta R_\psi$

Chapitre 9

Segmenting and Parsing Instrumentalists' Gestures

B. Caramiaux[1], **M.M. Wanderley**[2], **F. Bevilacqua**[1]

[1] UMR IRCAM-CNRS, Paris, France
[2] IDMIL, CIRMMT, McGill University, Montreal, Canada

Abstract : This article presents a segmentation model applied to musician movements, taking into account different time structures. In particular we report on ancillary gestures that are not directly linked to sound production, whilst still being entirely part of the global instrumental gesture. Precisely, we study movements of the clarinet captured with an optical 3D motion capture system, analyzing ancillary movements assuming that they can be considered as a sequence of primitive actions regarded as base shapes. A stochastic model called segmental hidden Markov model is used. It allows for the representation of a continuous trajectory as a sequence of primitive temporal profiles taken from a given dictionary. We evaluate the model using two criteria : the Euclidean norm and the log-likelihood, then show that the size of the dictionary does not influence the fitting accuracy and propose a method for building a dictionary based on the log-likelihood criterion. Finally, we show that the sequence of primitive shapes can also be considered as a sequence of symbols enabling us to interpret the data as symbolic patterns and motifs. Based on this representation, we show that circular patterns occur in all players' performances. This symbolic step produces a different layer of interpretation, linked to a larger time scale, which might not be obvious from a direct signal representation.

Keywords : Segmental Model, Gesture Analysis, Ancillary Gestures

9.1 Introduction

Musicians make several movements while performing music. Movements intended to generate sounds are commonly called *instrumental gestures* (Cadoz and Wandereley, 2000; Wanderley and Depalle, 2005; Jensenius et al., 2009) and movements that are not directly involved in sound production (or music production) are usually called *accompanying gestures* (Cadoz and Wandereley, 2000) or *ancillary gestures* (Wanderley and Depalle, 2005). In this article, we propose a methodology for clarinetist's ancillary gestures segmentation highlighting their inherent multi-level *information* content. This work is related to other recent studies on musical gestures, in particular instrumental gesture modeling ((Engel et al., 1997; Loehr and Palmer, 2007; Dahl, 2000; Maestre, 2009; Rasamimanana and Bevilacqua, 2008)) and computational models for investigating expressive music performance (Widmer et al., 2003; Widmer and Goebl, 2004). This can give important perceptive insights for the design of virtual instruments, sound installations and sound design applications.

Ancillary gestures

Ancillary gestures are musical gestures (Jensenius et al., 2009) that are not related to sound production but convey relevant information about the player's expressivity during a performance. In (Davidson, 1993), the author shows that the expressive intentions of musical performers are carried most accurately by their movements. This was later shown for the particular case of clarinet performance (Vines et al., 2004). Vines et al. explore how expressive gestures of a professional clarinetist contribute to the perception of structural and emotional information in musical performance. A main result is that the visual component carries much of the same structural information as the audio. Previously, Wanderley in (Wanderley, 2002) has investigated clarinetists' ancillary gestures providing findings that can be summarized as follows : the same player performing a musical piece several times tends to use the same gestures ; some gestures related to structural characteristics of the piece tend to be similar across the performances of different musicians whilst others remain subjective. These results are specified in (Wanderley et al., 2005). The authors show that clarinetists' subjective interpretation can be clustered according to which parts of the body are the most involved in ancillary gestures : some expressed in their knee and others used waist-bending gestures. Moreover, from the clarinetists' point of view, they show that the players feel uncomfortable when they try to consciously restrain their gestures whereas most of them seem to be aware of their expressive movements but not conscious of the gesture details. A recent study by Teixeira et al. (Teixeira et al., 2010) has highlighted movement patterns in the clarinetists' head by an empirical analysis and a qualitative segmentation. Results from these works highlight importance of ancillary gestures in communicating intention to the audience as well as understanding their expressive and spontaneous nature.

However, most of these works remain qualitative and do not propose quantitative methods for characterizing subjective gesture expressivity. One of the reasons is the problem of retrieving which part of the body (or which movement feature) is relevant for the analysis from high dimensional captured data often provided by a 3D full body motion capture system. Two main approaches can be used.

1. The *top-down approach* considers all the data and tries to find a subset of relevant features explaining the gesture. Usual techniques are PCA (Glardon et al., 2004), Neural Networks (Moeslund et al., 2006), or CCA for cross-modal dimension reduction (Caramiaux et al., 2010c).
2. The *bottom-up approach* considers a restricted part of the movement selected by prior knowledge and shows that it can be used to make suitable assessments on gesture expressivity ((Dahl, 2000; Maestre, 2009; Rasamimanana and Bevilacqua, 2008)) .

In the scope of this paper, we follow the second approach in selecting a specific part of the captured elements, namely the instrument, that has been shown to be pertinent to characterize instrumentalists' ancillary gestures (Wanderley, 2002; Wanderley et al., 2005).

Gesture as a sequence of primitive actions

Our basic hypothesis is that musical gestures can be considered as a sequence of primitive actions understood as primitive shapes. Previous works in cognitive sciences stated that people "make sense of continuous streams of observed behavior [like actions, music, ...] in part by segmenting them into events" (Kurby and Zacks, 2008). Events can occur simultaneously at different time scales and according to a hierarchical structure.

In (Guerra-Filho and Aloimonos, 2007), the authors propose to adapt the linguistic framework to model human activity. The proposed human activity language consists of a three-level architecture : kinetology, morphology and syntax. Interestingly, kinetology "provides a symbolic representation for human movement that [...] has applications for compressing, decompres-

sing and indexing motion data". Similarly, in activity recognition literature, a usual technique is to recognize actions defined as human motion units and activities defined as sequences of actions (see (Turaga et al., 2008) for a survey).

In a recent paper, Godøy et al. (Godøy et al., 2010) explore the theory that "perceived [music related] actions and sounds are broken down into a series of chunks in peoples' minds when they perceive or imagine music". In other words we holistically perceive series of both gesture and sound units : gesture and sound are cut into smaller units and the fusion and transformation of respective units lead to larger and more solid units.

We consider that music-related gestures, like ancillary gestures, can be described according to different timescales, meaning different segmentation levels (or *chunking* levels) like for the *human activity language* defined in (Guerra-Filho and Aloimonos, 2007). Our aim is to provide a robust quantitative analysis technique, first, to represent the gesture signal into sequences of symbolic units (segmentation and indexing) and, second, to allow for further analysis of ancillary gestures taken as sequences of symbols (parsing). In this study, we will show that a trade-off has to be made between these two goals. This work is complementary to the previous work by Widmer et al. (Widmer et al., 2003), showing that quantitative methods from machine learning, data mining or pattern recognition are suitable for the analysis of music expression and allow for retrieving the various structural scales in music. An important difference resides in the data used. Widmer et al. used MIDI like data while we use continuous data from a full-body motion capture system.

This paper is organized as follows. We first report previous work on human motion segmentation in the next section. Then we propose an overview of the general architecture of our methodology in section 9.3. The system is based on two main parts : first, the definition of a suitable dictionary for expressive gesture representation (in section 9.4) ; second, the stochastic modeling by SHMM that is formally presented in section 9.5. Section 9.6 details our experiments on a chosen database. First, we evaluate the pertinency of the model for representing the data using a geometric and a probabilistic criterion. Then, we show that it is suitable for motion pattern analysis of clarinetists' interpretation of a music piece. In section 9.7, we conclude and propose short-term prospective works.

9.2 Related Work

Motion segmentation methods can be roughly categorized into either unsupervised or supervised techniques. Unsupervised segmentation algorithms do not need any prior knowledge of incoming signals. Barbič et al. (Barbič et al., 2004) have shown that human behaviors can be segmented using simple methods like Principal Component Analysis (PCA), probabilistic PCA (PPCA) and Gaussian Mixture Model (GMM) that are only based on information available in the motion sequence. Changes in intrinsic data dimension (PCA, PPCA methods) or changes in distribution (GMM method) define segments' limits. Other methods use velocity properties of the joint angles (Fod et al., 2002) or perform implicit segmentation as a learning process (Brand and Hertzmann, 2000). In this last paper, Brand et al. use an unsupervised learning process on human motion to build *stylistic HMM* defined as a generic HMM (for instance describing bipedal human motion dynamics) changing according to a style parameter (for instance describing walking or strutting). This method allows complex human motions to be segmented and re-synthesized. However the internal states defining motion units are difficult to interpret and the method gives access to solely one timescale for movement description. More sophisticated methods like the nonparametric Bayesian process are used to model learning of action segmentation and its causal nature but are specific to goal-oriented actions (Buchsbaum et al., 2009).

The second set of algorithms are supervised techniques, where *primitives* (in a wide sense) attributed to the signal segments are known. Arikan et al. in (Arikan et al., 2003) have proposed

a motion synthesis process based on a sequence of primitive actions (e.g. `walking–jumping–walking`) given by the user. It allows for higher-level control on motion synthesis but requires an annotated gesture database. Annotations are usually provided by the user and make sense for either action-oriented movement or activity synthesis. Our solution defines motion primitives as temporal profiles or shapes rather than words. An interesting model previously used for shape modeling and segmentation is the segmental hidden Markov model (SHMM). This model has been studied in different research fields : speech recognition (Ostendorf et al., 1996), handwriting recognition (Artières et al., 2007) and time profile recognition of pitch and intensity evolution in (Bloit et al., 2010). SHMM allows continuous signals to be segmented and indexed at the same time. The system first needs a base shape dictionary used to describe input gestures as a sequence of basic shapes. Then a model defined as a sequence of shapes can be learned (or fixed) for recognition process. Hence, a character is a sequence of strokes (Artières et al., 2007) or a violin sketch like *tremolo* is a sequence of pitch shapes (Bloit et al., 2010). Our contribution is to show that the SHMM-based approach can efficiently represent clarinetists' ancillary gestures as a sequence of primitive shapes useful for the analysis of gesture patterns characterizing idiosyncratic player interpretations.

9.3 System overview

Figure 9.1 illustrates the general architecture of the methodology. It is specified for clarinetists' ancillary gestures but the methodology can be used for other kinds of gesture inputs like action-oriented human motion. Here, we focus on data captured by a 3D motion capture system that gives access to marker positions along the Cartesian axis, allowing the skeleton and the instrument movement to be reconstructed.

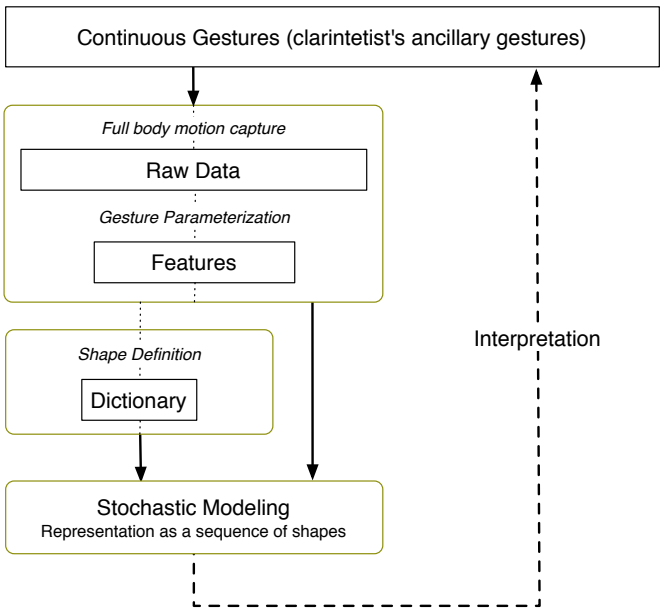

FIGURE 9.1 – Architecture of our system. Continuous gestures are captured by a specified motion capture system. From the raw data, we define the gesture features (in this paper we consider 2 features) and build a base shape dictionary. A segmental HMM is applied on captured gesture based on the defined dictionary. Resulting gesture representation can be interpreted in terms of clarinetist's expressive movements.

We assume a *prior knowledge* on the dataset that consists of a selection of relevant features for gesture representation and the definition of a set of primitive shapes (namely the base shape dictionary). This prior knowledge is based on previous work in the literature. It corresponds

to the first two blocks in figure 9.1 and will be further detailed in the next section. Even if using a supervised segmentation technique, the methodology is modular and these blocks could be replaced by a learning process either supervised or unsupervised. For instance, from a suitable gesture parameterization, we could learn primitive shapes of the dictionary by specifying the number of shapes we require or by using previously annotated clarinetists' gestures.

The stochastic model is based on a segmental hidden Markov model that represents the input gesture signal from the database by the likeliest sequence of continuous and time-warped shapes taken in the dictionary. SHMM requires that both dictionary elements and input signal have the same representation. Interpretation allows the initial hypothesis to be validated, i.e. that clarinetist's expressive gestures can be represented as a sequence of meaningful primitive shapes. Interpretation consists of verification with recorded video and observation.

9.4 Gesture parameterization and dictionary

In this section we present the chosen gesture parameterization and the set of shapes composing a dictionary. This prior knowledge is based on previous work (Wanderley, 2002; Wanderley et al., 2005; Palmer et al., 2009) on clarinetists' ancillary gesture analysis. We first select the gesture features ; then we propose four base shape dictionaries that will be used in this paper. These dictionaries are defined to be flexible enough to handle expressive gesture representation.

9.4.1 Features for clarinetist's gestures

Full body motion capture

As mentioned in the introduction, we suppose that we have access to marker positions from a 3D motion capture system. An absolute cartesian frame (x, y, z) is defined by the motion capture system calibration. From previous work (see (Wanderley, 2002), (Wanderley et al., 2005) and (Palmer et al., 2009)), the bell's movements have been shown to convey relevant information about clarinetists' gestural expressivity. A local frame is defined in which we describe the movements of the bell. The origin of the local coordinate system is set to the reed marker, and we consider the vector drawn by the clarinet $(bell - reed)$ in the Cartesian frame.

Gesture parameterization

Because both the origin and the clarinet's dimensions are fixed, the clarinet's movements can be entirely defined by its angles in the relative spherical coordinates. Let $\mathbf{C}$ be a vector representing the clarinet in the cartesian coordinates system : $\mathbf{C}(x, y, z) = (bell - reed)(x, y, z)$. Transformation to a spherical system as depicted in figure 9.2 is denoted by : $\mathbf{C}(\rho, \theta, \phi)$. In the spherical coordinates, ρ is the radius, θ the azimuth angle and ϕ the inclination angle. Here ϕ is preferably called elevation angle. Since ρ is constant, we consider only θ, ϕ.

9.4.2 Dictionary

A dictionary is a set of primitive shapes defining the basis on which an input gesture is decomposed. As mentioned above, each shape is parameterized by the azimuth and elevation angles θ, ϕ.

Defining segments

In this paper we consider four dictionaries of different sizes, as depicted in figure 9.3. The figure details all the shapes used to build each dictionary. Each shape in the figure describes

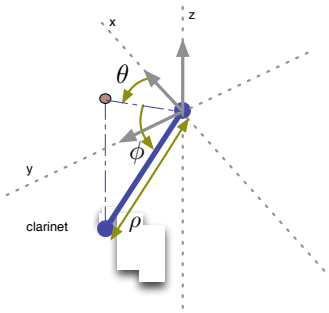

FIGURE 9.2 – The clarinet's movement from cartesian to spherical coordinates system. The radius ρ is constant so we choose azimuth θ and elevation ϕ to describe the clarinet's bell movements.

the evolution of the bell's movements described by the evolution of the spherical coordinates (θ, ϕ). The first dictionary contains 8 primitive shapes that describe four equal parts of two circles : clockwise and counterclockwise directions. The second dictionary contains 44 primitive shapes. It generalizes the previous one and contains it. It takes into account the diagonals and intermediate trajectories between quarter-circles and diagonals. The third dictionary is an intermediate between dictionary 1 and 2. It contains 12 primitive shapes that include with quarter-circles and diagonals. The fourth dictionary also contains 12 primitive shapes : 8 from dictionary 1 plus vertical and horizontal trajectories.

Indexing primitive shapes

Each shape from the dictionaries is associated with an (integer) index. Figure 9.3 illustrates the four considered dictionaries and the shape indices. Globally, all the shapes shared by several dictionaries have the same index within these dictionaries. This simplifies the comparison of sequences of indices, obtained from distinct dictionaries. In the same way, we associate intermediate integer indices to intermediate shapes (e.g. between quarter-circle and diagonal in dictionary 2). Finally, dictionary 4 has four shapes that are not included in the other dictionaries and not intermediate of previously defined shapes, so we chose to index them by negative values.

9.5 Stochastic modeling

We first present the theory of Segment Hidden Markov Models and then further discuss learning, preprocessing and inference.

9.5.1 Segment hidden Markov model (SHMM)

SHMM is a generative model used for shape-modeling. It generalizes classical HMM in the sense that emission probability density functions are defined at a segment level instead of at the sample level. Therefore, a segment generated by SHMM is allowed to follow a chosen regression form (Kim and Smyth, 2006). Figure 9.4 illustrates SHMM technique applied to an

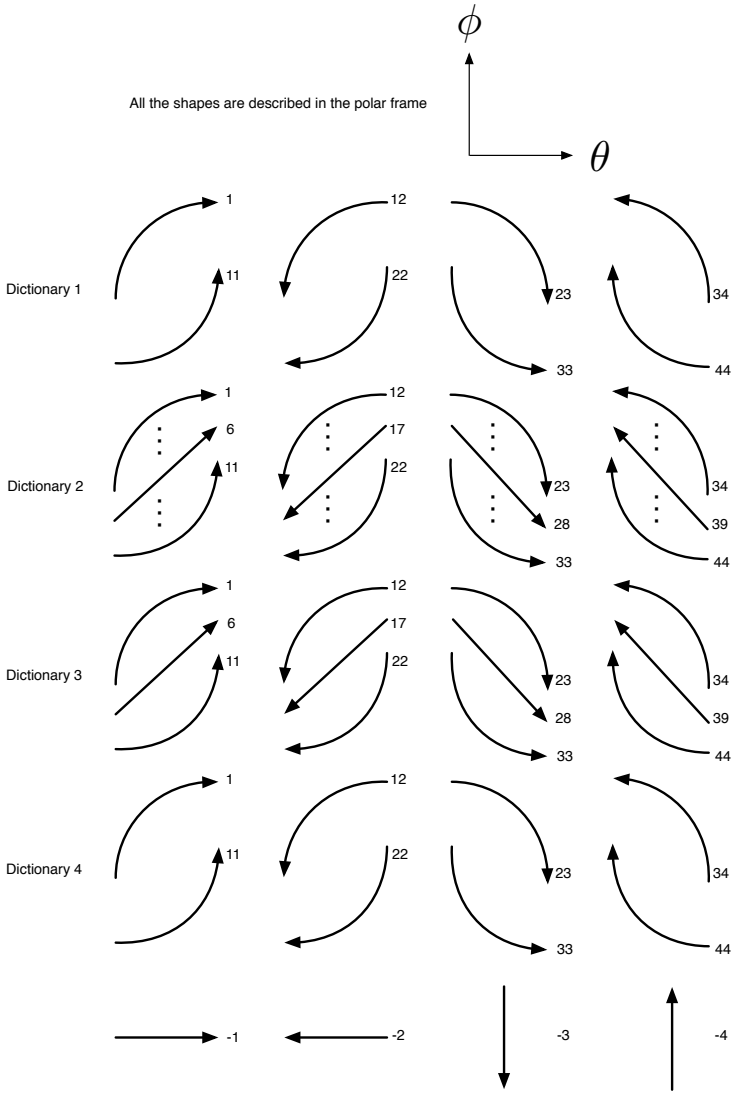

FIGURE 9.3 – The four dictionaries defined for our study. From top to bottom : the first dictionary contains 8 elements corresponding to quarter-circles ; the second dictionary also has the diagonal and shapes in-between quarter-circle and diagonal leading to 44 elements ; the third one contains 12 elements which are quarter-circles and diagonals ; finally the last one has 12 elements that are 8 elements from dictionary 1 plus horizontals and verticals.

input signal curve. A uniformly sampled signal is taken as input of the segment model. An inference process returns the likeliest sequence of states (which correspond to elements of the dictionary) that fits the input signal and their respective durations.

Formally, we denote $\mathbf{y} = [\mathbf{y}_1 \dots \mathbf{y}_T]$ the whole incoming feature vector sequence. A subsequence of $\mathbf{y}$ from t_1 to t_2 (with $t_1 < t_2$) is denoted $\mathbf{y}_{t_1}^{t_2} = [\mathbf{y}_{t_1} \dots \mathbf{y}_{t_2}]$ and its length is written $l = t_2 - t_1 + 1$. SHMM allows for representing $\mathbf{y}$ as a sequence of segments :

$$\mathbf{y} = \left[\mathbf{y}_1^{l_1} \quad \mathbf{y}_{l_1+1}^{l_1+l_2} \quad \dots \quad \mathbf{y}_{\sum_i l_i}^{T}\right]$$

Each segment is of length l_i and we have $\sum_{i=1}^{\tau} l_i = T$ where τ is the number of segments inferred by SHMM to represent $\mathbf{y}$. Hence each SHMM state q emits a sequence $\mathbf{y}_{t_1}^{t_2}$ and a length $l = t_2 - t_1 + 1$ according to a density function $p(\mathbf{y}_{t_1}^{t_2}, l|q) = p(\mathbf{y}_{t_1}^{t_2}|l, q) \times p(l|q)$. The distribution of segment durations is $p(l|q)$, and the likelihood of the sequence is $p(\mathbf{y}_{t_1}^{t_2}|l, q)$. If we write $q_1^{\tau} = [q_1 \dots q_\tau]$ the sequence of states, taking values in a finite set $\mathcal{S}$, associated to the input

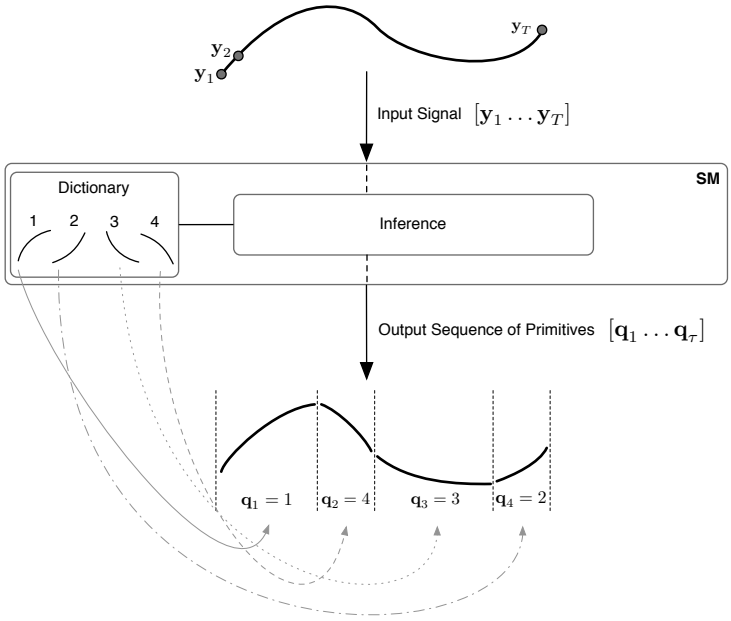

FIGURE 9.4 – Illustration of the application of the segment model on a uniformly sampled input continuous curve. The inference process finds the likeliest sequence of states and their durations that generates the input signal. States are elements of the dictionary that can be time stretched to fit the input curve.

sequence $\mathbf{y}$, and $l_1^\tau = [l_1 \dots l_\tau]$ the sequence of lengths, taking values in a finite set $\mathcal{L}$, the probability that the model generates the input sequence $\mathbf{y}_1^T$ is :

$$p(\mathbf{y}_1^T|q_1^\tau) = \sum_{l_1^\tau} p(\mathbf{y}_1^T|l_1^\tau, q_1^\tau) p(l_1^\tau|q_1^\tau) \tag{9.1}$$

Where the sum is over all possible duration sequences. Considering that the segments are conditionally independent given the state and the duration and considering that the pairs $(state, duration)$ are themselves independent, we can rewrite the probabilities as :

$$\begin{aligned} p(\mathbf{y}_1^T|l_1^\tau, q_1^\tau) &= \prod_{i=1}^{\tau} p\left(\mathbf{y}_{l_{i-1}+1}^{l_i}|l_i, q_i\right) \\ p(l_1^\tau|q_1^\tau) &= \prod_{i=1}^{\tau} p\left(l_i|q_i\right) \end{aligned} \tag{9.2}$$

Figure 9.5 represents an unrolled SHMM as a graphical model where each arrow represents a conditional dependency. At the bottom is the generated sequence $\mathbf{y}_1^T$. Hidden layer states are : q_t (a segment index) ; l_t (the segment's duration) ; and X_t (a state from a segment emitting the observation $\mathbf{y}_t$).

9.5.2 Learning and inference of SHMM

Learning SHMM

From the previous description, the hidden layer dynamics of SHMM can be modeled by three probability distribution functions :

1. State prior distribution : $\pi(i) = p(q_1 = s_i), \forall s_i \in \mathcal{S}$ or how the first shape of the sequence is chosen.

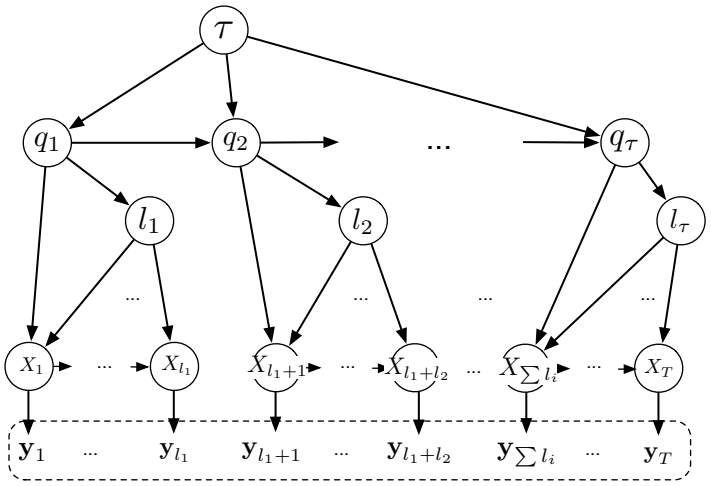

FIGURE 9.5 – Graphical model of a Segmental Hidden Markov Model (Murphy, 2002). Arrows are conditional dependencies between variables. Inference is finding the likeliest q_1^τ and l_1^τ that generate $\mathbf{y}_1^T$. Hence duration l_i is dependent on state q_i and generated segment is dependent on the duration and the current state.

2. State duration distribution : $p(l_n|q_n = s_i)$, $l_n \in \mathcal{L}$ or how segment durations are weighted during inference.
3. State transition distribution : $p(q_{n+1} = s_j|q_n = s_i)$ denoted a_{ij}, $\forall(i,j) \in [1 \dots \tau]^2$ or how shapes in a dictionary are weighted during the inference.

Typical tools used in HMM framework for training can be used to learn SHMM parameters (e.g. Expectation–Maximization algorithm (Ostendorf et al., 1996)). As mentioned in section 9.2, no well-defined ancillary gesture vocabulary can be used for training. Thus, the methodology adopted is to define prior dictionaries (section 9.4.2), then show that SHMM is relevant for ancillary gestures representation (section 9.6.2) and discuss the construction of *inter-* and *intra-* players gestures classes (section 9.6.4). For that purpose we use a generic configuration of SHMM based on uniform distributions :

1. $\pi(i)$ *uniform* means that any shape in the considered dictionary can be used as the first shape in the sequence.
2. $p(l|q_n = s_i)$ *uniform* means that shapes placed in the sequence can have any duration, each possible duration having the same weight. This choice is discussed later.
3. $p(q_{n+1} = s_j|q_n = s_i)$ *uniform* means that for each shape, any shape placed afterwards has the same weight (the so-called *ergodic* model).

Inference

Inference is the estimation of the likeliest state sequence that has emitted the input gesture data. It means estimating the number of segments $\hat{\tau}$, the segment sequence $\hat{q_1}^{\hat{\tau}}$ and the corresponding length sequence $\hat{l_1}^{\hat{\tau}}$. This can be done by finding the arguments maximizing the likelihood function defined by equation (9.1), that is :

$$(\hat{\tau}, \hat{q_1}^{\hat{\tau}}, \hat{l_1}^{\hat{\tau}}) = arg \max_{\tau, q_1^\tau, l_1^\tau} \sum_{l_1^\tau} p(\mathbf{y}_1^T|l_1^\tau, q_1^\tau) p(l_1^\tau|q_1^\tau) \tag{9.3}$$

As previously mentioned, transition probability distribution and duration probability distribution are uniform. Here, we define the observation (or emission) probability distribution. An input signal is represented as a bi-dimensional time series, uniformly sampled, representing

the evolution of azimuth and elevation angles θ, ϕ. The incoming sequence of observations $\mathbf{y}_1^T = [\mathbf{y}_1 \dots \mathbf{y}_T]$ is defined such that :

$$\mathbf{y}_t = \begin{bmatrix} \theta_t \\ \phi_t \end{bmatrix}$$

Emission probability is defined as :

$$p\left(\begin{bmatrix} \theta_{l_{i-1}+1}^{l_i} \\ \phi_{l_{i-1}+1}^{l_i} \end{bmatrix} | l_i, q = s\right) \propto \exp\left(-\sum_{j=l_{i-1}+1}^{l_i} \frac{\left\{[\theta_j - \theta(s)_j]^2 + [\phi_j - \phi(s)_j]^2\right\}}{2\sigma^2}\right) \tag{9.4}$$

Where $\theta(s)_j$ (respectively $\phi(s)_j$) is the value of the first coordinate (respectively the second) of shape s at time j ; and σ is the gaussian standard deviation. Exact inference is made using the forward-backward Viterbi algorithm. For an observation sequence of length T, it takes $O(MDT^2)$ where M is the number of primitives in dictionary, D is the maximum length of possible durations. Hence, doubling the number of elements in a dictionary implies doubling the computation time. It can be a criterion for selecting a dictionary among the others.

Preprocessing and resynthesis

While processing the inference, each segment to be compared to shapes in a dictionary is normalized to $[0, 1]$ meaning that both azimuth and elevation signals are translated by their minimum and scaled by their maximum. Let us consider the i-th segment ; we define an offset coefficient

$$\min_{l_{i-1}+1 \leq j \leq l_i} (\begin{bmatrix} \theta_j \\ \phi_j \end{bmatrix})$$

as well as a scaling coefficient

$$\max_{l_{i-1}+1 \leq j \leq l_i} (\begin{bmatrix} \theta_j \\ \phi_j \end{bmatrix})$$

These coefficients are stored for the resynthesis process. During resynthesis, for each shape taken sequentially in the inferred sequence by SHMM, we scale the whole shape by the scaling coefficient and translate the scaled shape by the offset coefficient.

9.6 Results

In this section we first present the material used for the analysis. We then inspect the accuracy of the model with respect to the considered dictionary, and we finally show how the resulting sequence of indexes can be parsed and what kind of information it gives us for characterizing and analyzing ancillary gestures.

9.6.1 Database

The database used for experiments was recorded at the Input Devices and Music Interaction Laboratory (IDMIL) at McGill University, Montreal, Canada. From the whole database we have selected four clarinetists playing the first movement of the Brahms *First Clarinet Sonata Opus 120, number 1*. The task was to interpret the piece four times in a neutral way. The set-up was as follows : The clarinet sound was recorded using an external microphone, and a video camera was used to record all the players' performances. Clarinet movements were recorded using a 3D passive infra-red motion capture system. Two of the markers were placed on the clarinet, one the reed and one the bell.

In order to cross-analyze players' performances, we need to align each performance to a single reference. The common reference is the score of the first movement of the Brahms

sonata. First we build a synthetic interpretation : the score is translated into a pitch time series with a fixed sample rate (see figure 9.6). Synthetic pitch evolution corresponds to the piece played regularly following a tempo of 100. The pitch evolutions of the subjects' performances are pre-processed to be discrete and aligned to the synthesized signal using the Dynamic Time Warping (DTW) technique. Since audio and gesture streams are recorded synchronously, the resulting alignment is also applied to gesture data.

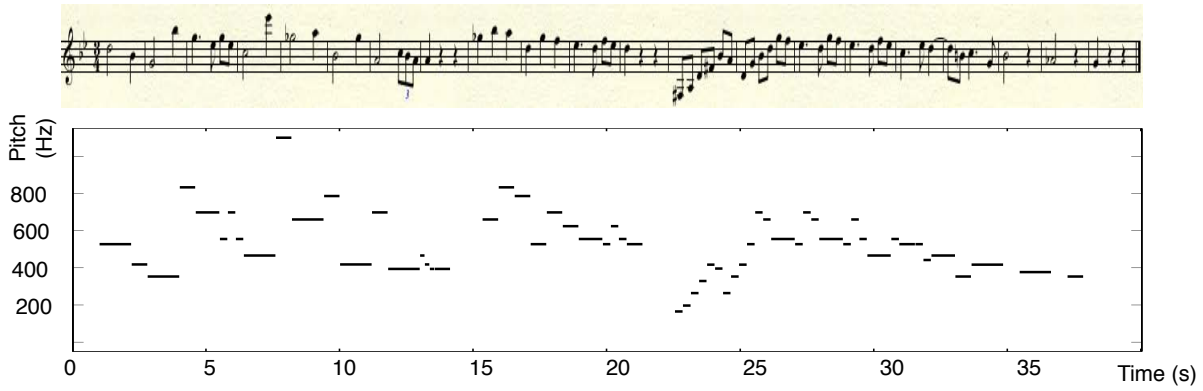

FIGURE 9.6 – Synthetic interpretation. The figure is the pitch signal (piano roll like) of an interpretation of Brahms' *First Clarinet Sonata Opus 120, number 1* played at tempo 100.

An example of one performance by each clarinetist is depicted in figure 9.7. Solid black lines represent azimuth signals, and dashed black lines represent elevation signals. Each signal has been centered to have a zero mean and aligned on the reference interpretation.

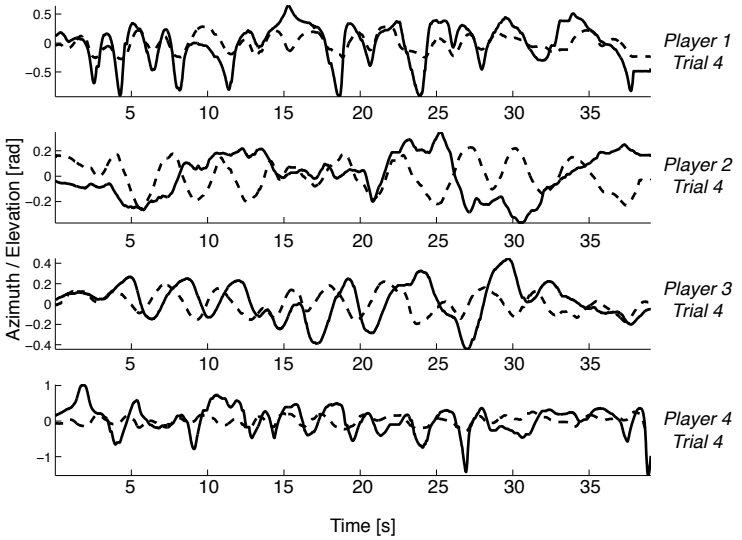

FIGURE 9.7 – Examples of signals for one interpretation by players 1, 2, 3, 4. Solid lines are azimuth signals θ_t, and dashed lines represent elevation signals ϕ_t

9.6.2 Ancillary gesture as a sequence of base shapes

To clarify the reading, let us take the example of a part of player 4's fourth performance from 17 seconds to 22 seconds (the whole signal is plotted at the bottom of figure 9.7). We

denote $\mathbf{y}_{t_1}^{t_2}$ where $t_1 = 17$ and $t_2 = 22$, the sequence of observations taken as input for SHMM. Considering the dictionaries 1 and 2, the model inferred a sequence of shapes per dictionary. Figure 9.8 shows the results : on the upper part are the results obtained with dictionary 1 ; on the lower part are those obtained with dictionary 2. For each, two plots are depicted : at the top is the azimuth angle ($\theta_{t_1}^{t_2}$) ; and at the bottom, the elevation angle ($\phi_{t_1}^{t_2}$). Dashed lines are the original angle time series and gray solid lines are the sequences of shapes inferred by SHMM for both dictionaries. Bracketed integers are the shape indices from the considered dictionary.

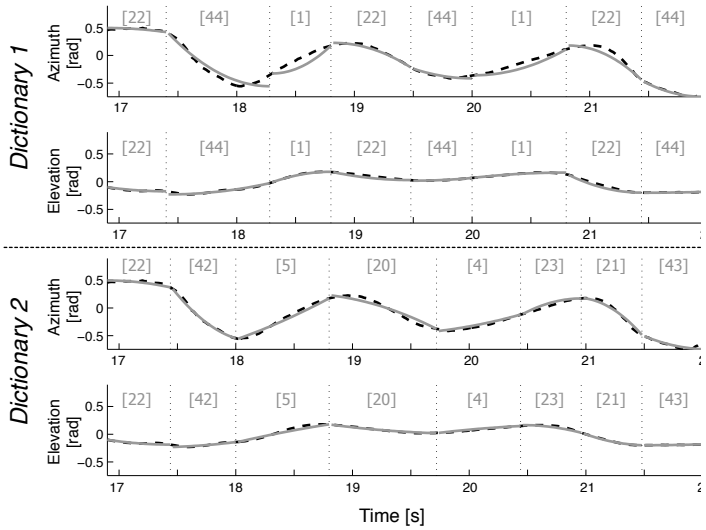

FIGURE 9.8 – Examples of resynthesized curves. Upper part presents the input signal and the resynthesized signal using dictionary 1. Dashed lines are original curves (azimuth at the top, elevation at the bottom), and piecewise gray lines are the shapes inferred by SHMM. Similarly, the lower part reports the result using the second dictionary

Intuitively, a more exhaustive dictionary should better represent a given continuous multidimensional curve. This can be observed in figure 9.8 around 18.5sec. Let us consider dictionary 1, the likeliest shape for the input signal around 18.5sec is the shape indexed by 1, and it better matches the elevation signal than the azimuth signal. In dictionary 2, the result reveals that the likeliest shape to model this part of the signal is shape 5 that does not belong to dictionary 1. This shape is an intermediate shape between 1 and the diagonal 6. On the other hand, this simple example shows a more regular sequence of indices (bracketed integers) on the upper part of figure 9.8 than on the lower part. Hence, a more exhaustive dictionary seems to provide a more varying sequence of symbols.

This example illustrates how a gesture input is represented by a sequence of shapes taken from a dictionary and shows that differences appear according the choice of the dictionary. Firstly, we want to generalize the approach by systematically evaluating the accuracy of such a representation through other input gestures from the database and the available dictionaries. Secondly, we want to qualitatively analyze the sequences of symbols provided by the model according to the different dictionaries.

9.6.3 Evaluation of the model

SHMM infers the likeliest sequence of shapes together with the likeliest sequence of the shapes' durations. The concatenation of the inferred time warped shapes offers a new representation of the input signal (as presented in the schematic view figure 9.4). Here, we inspect how accurate is the re-synthesis in representing ancillary gestures from real performances of

clarinetists. To do so, we propose two evaluation criteria : the Euclidean norm and the log-likelihood. While the first criterion tests the fitting accuracy, the second tests how likely is the model to generate the data (also called *prediction power* (Kim and Smyth, 2006)). As mentioned before, distributions involved in the model are uniform. The possible durations are from 0.1sec to 1.3sec (step of 0.1sec) and the standard deviation used is $\sigma = 0.1$ radian.

Evaluation using the Euclidean norm $\|.\|_2$

As mentioned in section 9.6.1, the database contains four natural interpretations of the Brahms *sonata* by four players. We evaluate how these gestures are represented according to each of the four dictionaries defined earlier. An Euclidean distance is computed between the resynthesized signal and the original one. Therefore, one distance value per interpretation is returned. We average over the interpretations so that one value remained per player. Results are reported in table 9.2 showing means and standard deviations.

	Dictionaries (#elements)			
$\times 10^{-3}$	1 (8)	2 (44)	3 (12)	4 (12)
Player 1	1.41 ± 0.18	0.84 ± 0.12	0.84 ± 0.12	1.40 ± 0.19
Player 2	0.37 ± 0.07	0.19 ± 0.02	0.21 ± 0.01	0.37 ± 0.07
Player 3	0.39 ± 0.07	0.18 ± 0.03	0.24 ± 0.01	0.39 ± 0.07
Player 4	1.28 ± 0.37	1.07 ± 0.51	1.04 ± 0.38	1.33 ± 0.31

TABLE 9.1 – Averaged Euclidean distance with standard deviation. Values are reported in 10^{-3} radians. Lowest values correspond to better fitting between the resynthesized signal and the original one.

In this table, lowest values mean better fitting. Globally, the results show that the model efficiently fits the incoming data : the maximum mean value is 1.41×10^{-3} radians (player 1, dictionary 1) corresponding to the mean quantity of variation between the two curves. Moreover standard deviations across interpretations are very low, meaning that there are not important variations intra-participant (Wanderley, 2002). For players 1, 2 and 3, lowest scores are obtained for dictionary 2 although they are very close to the scores obtained with dictionary 3. Dictionary 1 and 4 return the same scores and a close look at inferred sequences reveals that SHMM returns the same sequence of shapes for both dictionaries, meaning that the vertical and horizontal shapes in dictionary 4 are not involved in the inferred sequence. The case of player 4 is singular because standard deviations are high and the dictionaries can not be statistically discriminated : the Student's t-test shows that the mean obtained for dictionaries 1 to 4 are not significantly different at the 5% significance level. To conclude, according to the Euclidean norm, dictionary 2 and 3 are equivalent as well as dictionaries 1 and 4. Hence the number of elements in a dictionary is not necessarily linked to a better fitting.

Evaluation of the temporal prediction score

We compute the log-likelihood $\log p(\mathbf{y}_1^T | l_1^\tau, s_1^\tau)$ for the observation time series $\mathbf{y}_1^T$ (or test data). This score refers to the likelihood assigned to the incoming observation by the model (Kim and Smyth, 2006). Higher scores mean that the model is likely to generate the test data. In other words, the model has a better predictive power. In order to be length independent, we normalize by the length of the observation sequence. The average log-likelihood values are computed for each subject and the results are given in table 9.2.

The results show that the highest log-likelihood scores are obtained with dictionary 2, meaning that the probability to generate data from the Brahms database is higher with dictionary 2 than either with dictionary 1, 3 or 4. Dictionary 2 is more exhaustive than the other three,

	Dictionaries (#elements)			
	1 (8)	2 (44)	3 (12)	4 (12)
Player 1	-1.335 ± 0.023	-0.643 ± 0.058	-0.743 ± 0.063	-1.353 ± 0.023
Player 2	-1.294 ± 0.041	-0.641 ± 0.028	-0.712 ± 0.047	-1.309 ± 0.042
Player 3	-0.907 ± 0.148	-0.481 ± 0.033	-0.584 ± 0.037	-0.920 ± 0.148
Player 4	-1.163 ± 0.087	-0.662 ± 0.056	-0.748 ± 0.056	-1.179 ± 0.085

TABLE 9.2 – Cumulative Log-likelihood ($\sum_{t=1}^{T} \log(p(\mathbf{y}_1, ..., \mathbf{y}_t | q_1^\tau))$) averaged over the interpretations and the standard deviations. The criterion returns the likelihood that the model generates the observation data. Highest values correspond to better perdiction.

and table 9.2 shows that the scores obtained with dictionary 3 are significantly better than with dictionary 1. An interpretation is that the more exhaustive a dictionary is, the better it is to generate (and consequently to predict) the observed data. This is in contrast with the results based on Euclidean distance : we add *information* [1] from dictionary 3 to dictionary 2 by adding new primitive shapes whereas it does not affect how the reconstructed curve fits the input curve. This will be refined in section 9.6.4. However, considering the scores obtained with dictionary 4, a t-test (using a 5% significance level) shows that they are not significantly different from the scores with dictionary 1 (similar to the evaluation using the Euclidean norm) even if dictionary 4 contains more elements than dictionary 1.

This can be explained as follows. As mentioned in section 9.5, the log-likelihood is based on the observation likelihood and the transition probability. Therefore, adding elements in a dictionary : increases the observation likelihood as long as they better fit the observations (dictionary 2, 3) ; decreases their transition probability (uniform over the elements of the dictionary) (dictionary 2, 3 and 4)). Hence, a dictionary can be tested using this criterion as follows : starting from a simple dictionary (e.g. dictionary 1), one can add element by element and inspect the resulting log-likelihood. If the score is decreasing, it means that the the fitting criterion is stronger than the decreasing weight. Otherwise, the added element is not relevant for the input observations.

9.6.4 Parsing the sequences of indices

The sequence of shapes can be analyzed as a sequence of symbols (integer indices). The motivation is to observe the real-world data set of ancillary gesture signals at a higher level than the *shape level* presented in the previous section. A symbolic representation of continuous signal allows the retrieval of patterns and motifs based on *parsing* processes : the continuous signal can be considered as a string. There exists a large set of methods for *string* analysis (from machine learning, pattern recognition, theoretic computer science and so forth). Here we propose a qualitative interpretation by considering ancillary gestures as strings that could be useful for future research in this field. Thus, we discuss how the design of a dictionary can determine the pattern analysis. We start by considering the first dictionary as an example, and we compare to the results obtained with dictionaries 2 and 3 (in this section dictionary 4 is discarded since it does not add relevant information).

From shape level to pattern level

In this section, we inspect how the shapes and their durations are temporally distributed in the signal. Figure 9.9 presents the data for player 3's second interpretation. At the top, the temporal evolution of the pitch is represented as a continuous signal and zero values mean

1. *Information* is to be understood as the accuracy of the model to generate the input data

silences. Below, we report both the azimuth signal (solid line) and elevation signal (dashed line). Finally, at the bottom the sequence of shapes' index inferred by the stochastic model is plotted as a piecewise constant signal with circles indicating the starting and ending points of the shape. It appears that short-duration shapes are concentrated around 2 seconds (62.5% of shapes have durations lower than 500ms), 13 seconds (80% lower than 500ms), 24 seconds (75% lower than 500ms) and 33 seconds (66.7% lower than 500ms). In-between these instants, sequences of longer duration define a specific recurrent pattern (so-called *motif*) 1–23–22–44 that is highlighted by gray lines. A more detailed study of motifs will be given afterwards.

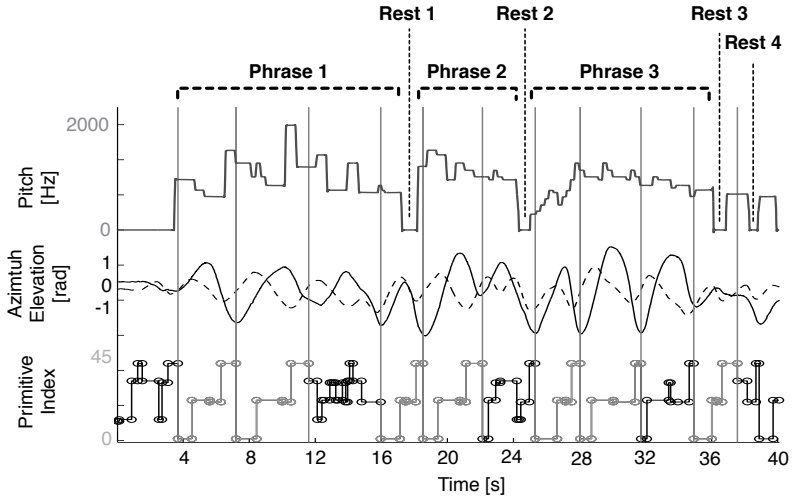

FIGURE 9.9 – Player 3's second interpretation. At the top we report the pitch evolution corresponding to this specific interpretation. In the middle, the solid line represents the bell's azimuth evolution while the dashed line draws the bell's elevation. At the bottom, we report the segmentation obtained by SHMM using the first dictionary. We added indications about the Brahms sonata : phrases 1, 2 and 3 as well as the various rests.

Three parts are of particular interest : $[12, 16]$ seconds (part 1), $[22, 25]$ seconds (part 2) and $[32, 35]$ seconds (part 3). Part 1 is in-between two identical patterns defined by the sequence 1–23–22–44. Our hypothesis is that it defines an articulation between two phrases separated by a rest. This is reinforced by the second part ($[22, 25]$ seconds) : it corresponds to the articulation between the second phrase and the third one which starts by the same pattern 1–23–22–44. Finally, part 3 seems to be of different nature. Since the global pattern is very similar to the recurrent pattern 1–23–22–44, the short-duration shapes detected in the middle of the pattern seem to be due to "noise" (i.e. unwanted perturbation that is not specific to the interpretation). Let us analyze further this assumption by discussing figure 9.10. The figure depicts the same performance as in figure 9.9. Pitch signal has been omitted for readability. On the upper part, we report the beginning and the end of the gesture as well as the three parts described above. On the lower part, we depict each occurrence of the pattern 1–23–22–44.

The seven occurrences of the pattern 1–23–22–44 correspond to a circle performed in counterclockwise direction and are mostly occurring during musical phrases. The second and the last circles clearly include the beginning of the next shape : it gives a trace of the articulation between shapes at a fine temporal scale. On the other hand, if considering the three parts of interest described above, part 1 (fig. 9.10, $[12, 16]$sec) and part 2 (fig. 9.10, $[22, 25]$sec) reflect a different behavior : an abrupt change of direction and the articulation between this change and the beginning of the next circle. Part 3 (fig. 9.10, $[32, 35]$sec) corresponds to a circle performed counterclockwise supporting our previous guess that short-duration shapes in this specific part are due to noise and do not reflect articulations between different behaviors. This study reveals that occurring patterns can be identified using the symbolic representation and

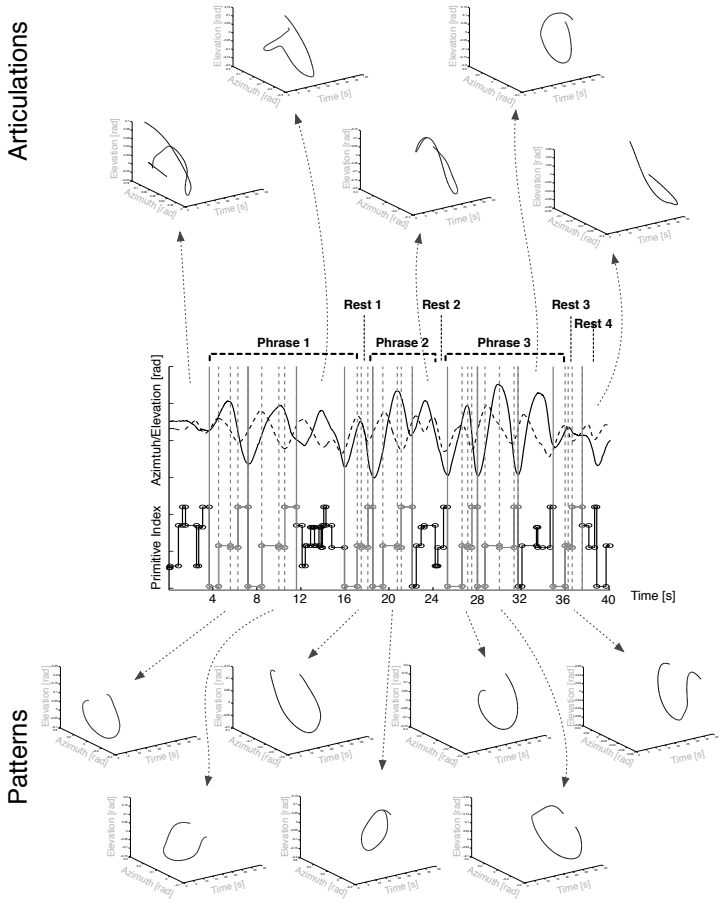

FIGURE 9.10 – Player 3's second interpretation. This figure shows the azimuth and elevation signals (respectively solid and dashed lines) and the segmentation inferred by SHMM using the first dictionary. Patterns are in gray, and articulations are in black. We explicitly plot the bell's original movement for these particular parts. Patterns are clockwise circles, with idiosyncratic movements in between them. Patterns can be retrieved using the symbolic representation though it is not clear from only the original signals.

highlights higher temporal structure in ancillary gestures (patterns – articulations). Interestingly, this could be handled in the SHMM by defining higher level models constituted by the sequences defining the patterns. Consequently, a non-ergodic model would optimize the recognition of such patterns.

Identifiable motifs

In the previous example, a clockwise circular pattern is repeated seven times and is described as the sub-sequence of indices 1–23–22–44 in the first dictionary. Here we show that a symbolic representation gives a powerful tool for exact motif retrieval in the continuous gesture signal. A quick enumeration of recurrent patterns in the data from players' interpretations shows that the previously introduced sequence 1–23–22–44 occurs 38 times (20% over all the interpretations in the database), mostly in interpretations by player 3 (25 occurrences) and 4 (12 occurrences) than player 1 (1 occurrence) and player 2 (0 occurrences). The pattern begins by shape 1 drawing the circle from the leftmost point. A circular permutation of the pattern index means starting the circle at a different position. A quick study shows that starting from the

left point creates the pattern 22–44–1–23 occurring 42 times (22%) : 32 occurrences for player 3 and 9 occurrences for player 4.

Figure 9.11 shows the two symbolic sequences corresponding to the inferred shapes for the fourth interpretations by both players 3 and 4. The co-occurring pattern 1–23–22–44 is emphasized by gray frames : 8 occurrences for player 3 and 5 occurrences for player 4. Note that the movement patterns do not necessarily occur at the same moment in the piece.

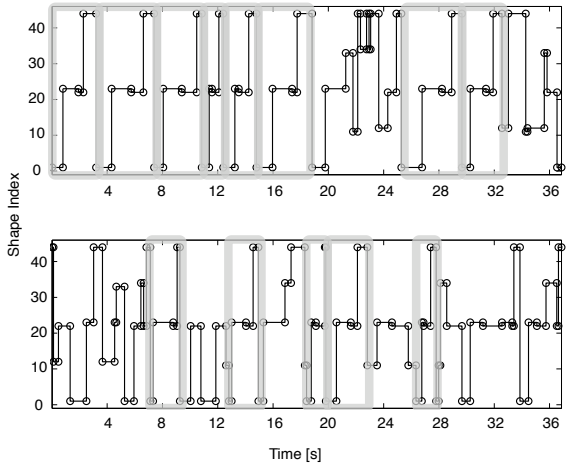

FIGURE 9.11 – At the top we can see the segmentation of player 3's fourth interpretation. Patterns 1–23–22–44 are highlighted by gray frames. At the bottom, we can see the segmentation of player 4's fourth interpretation. Same patterns are similarly highlighted. 1–23–22–44 pattern is particularly found in the interpretations by these two players.

A similar analysis shows that the pattern 11–34–12 occurs 19 times (10%) across all the interpretations and specifically in player 1's interpretations (12 occurrences) and players 2's interpretations (7 occurrences). This pattern consists in three quarter-circles in counterclockwise direction, that is, the opposite strategy from players 3 and 4. Hence, the four clarinetists make use of circles in their interpretations.

Motif-based dictionary information

In previous sections we have shown that a dictionary can hold different information contents : number of elements and/or accuracy of these elements for the specific input. In this section, we analyze the influence of a chosen dictionary on ancillary gesture motif recognition. Let us focus on pattern 1–23–22–44 in player 3's interpretations. Figure 9.12 reports resynthesized patterns in (θ, ϕ) coordinates for dictionaries 1, 2 and 3 together with the original input trajectories. The first line illustrates the patterns inferred by SHMM with dictionary 1. Above each reconstructed pattern, we report the corresponding sequence of indices. The second line (resp. the third) illustrates patterns inferred with dictionary 2 (resp. 3). Circle patterns corresponding to the index sequence 1–23–22–44 are marked by dashed rectangular boxes.

At first sight, the recurring pattern 1–23–22–44 retrieved using dictionary 1 has various shapes : closed circle, open circle, continuous and noncontinuous circles. Even so, all these shapes are labeled 1–23–22–44, that is a circle performed in clockwise direction. Using a more exhaustive dictionary allows for a better fitting of the resynthesized signal from the original signal (see section 9.6.2 and lines 2 and 3 in figure 9.12) : SHMM inferred more accurate shapes when using dictionary 2 and 3. For instance, let us consider the first pattern at each line in figure 9.12. The sequence of symbols given by dictionary 1 (first line) is 1–23–22–44. If we use dictionary 2 (second line), inference induces a novel sequence of indices that corresponds to

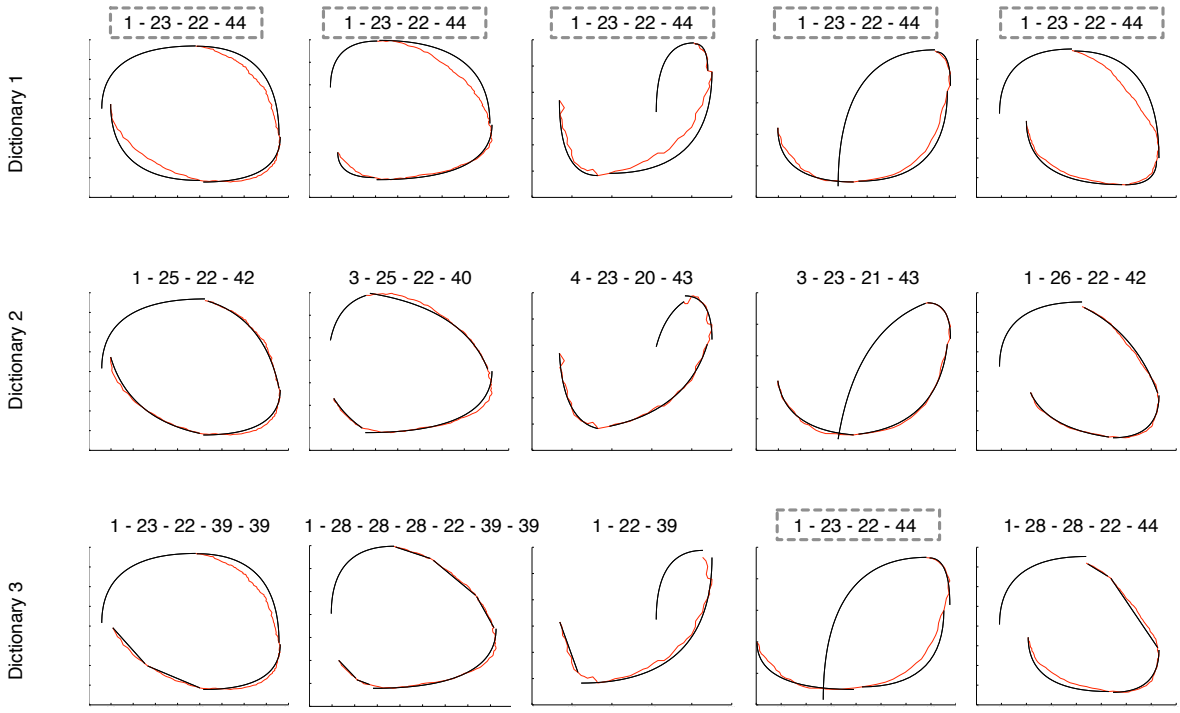

FIGURE 9.12 – Player 3's fourth interpretation. We can see 1–23–22–44 pattern occurrences in the interpretation inferred using three dictionaries : dictionary 1 at the top, dictionary 2 in the middle and dictionary 3 at the bottom. Columns refer to 5 moments in the piece. Shapes approach original signal (thin red lines) while the dictionary gets more exhaustive but symbolic representations become highly sensitive to the variations across shapes.

the following shape index transformations :

$$\begin{aligned} 1 &\rightarrow 1; \\ 23 &\rightarrow 25; \\ 22 &\rightarrow 22; \\ 44 &\rightarrow 42; \end{aligned}$$

Similarly, inferred sequence using dictionary 3 (third line) shows other novel sequence of index that corresponds to the following transformations :

$$\begin{aligned} 1 &\rightarrow 1; \\ 23 &\rightarrow 23; \\ 22 &\rightarrow 22; \\ 44 &\rightarrow 39 - 39; \end{aligned}$$

Hence, proposing a large set of shapes (e.g. dictionary 2) for gesture shape modeling implies a larger set of symbolic representations. In figure 9.12, all plotted circle patterns have a distinct symbolic representation (distinct sequence of indexes). It means that SHMM inferred sequences of shapes that tend to adapt to variations in the observations and, consequently, the model is more discriminative. On the contrary, the use of dictionary 1 allows similar continuous patterns (e.g., circles) to be retrieved even if they are not exactly the same at each occurrence (the so-called *motifs* as defined in (Yankov et al., 2007)). The inferred symbolic sequence

constitutes a general *stereotype* in the shape of a clockwise circle. Finally, sequences of indices inferred by SHMM with dictionary 3 vary in length : some indices are either repeated, omitted or changed. Even if this dictionary provided a good trade-off between a sample-by-sample fitting criterion (Euclidean norm criterion) and a temporal predictive criterion (log-likelihood criterion), it seems to be less adapted for symbolic representation of the patterns considered.

9.7 Conclusion and perspectives

In this paper, we have presented a segmentation algorithm applied to a real-world data set of clarinetists' ancillary gestures. The hypothesis is that ancillary gestures can be suitably analyzed by multi-level modeling techniques. The general system is based on the definition of a dictionary of primitive shapes and a stochastic temporal modeling of the sequence of shapes that best represents the input gesture signal. The model is tested on a database containing four interpretations of the first movement of the Brahms *First Clarinet Sonata Opus 120, number 1* by four clarinetists.

Firstly, we have shown that the proposed model infers sequences of shapes that accurately fit the input signal. A detailed study of the Euclidean distance between both the re-synthesized and the input signals has shown that the fitting does not depend on the number of elements in a dictionary. We have also shown that the log-likelihood of the temporal sequence increases (i.e better predictive power) if we add relevant elements in the dictionary (e.g dictionary 1 to dictionaries 2 and 3) but might decreases while the number of elements increases if the added elements are not pertinent (e.g, dictionary 4). A trade-off has to be made between the number of elements and how representative of the input observations these elements are. We proposed an incremental process to build a dictionary based on these criteria.

Secondly, we have shown that a greater concentration of short duration shapes occurs between recurrent regular sequences of longer shapes that we called patterns. Short duration shapes model articulations between phrases in the piece while patterns occur within phrases. Such high level structure seems not to be trivially retrievable if we consider only the initial signal.

Finally, we have shown that some patterns occur in all clarinetists' interpretations studied here. Precisely, we have shown that the choice of the dictionary makes a difference in retrieving the patterns. That is, a large set of shapes leads to variations in the sequence of symbols that represents the pattern. On the contrary, a smaller set of shapes allows for pattern stereotype definition. Hence, symbolic representation using semantically relevant primitive shapes highlights higher time structures in gesture signals that can be otherwise hidden.

A future improvement of our system will take into account a hierarchical structure in the dictionary. A first dictionary with stereotypical shapes will be defined, accompanied by refinements of each one these elements. We will also pursue a real-time implementation of the model.

Quatrième partie

Applications et Conclusions

Chapitre 10

Applications

10.1 Introduction

L'introduction de ce manuscrit de thèse a présenté un *instrument* conceptuel qui envisageait la possibilité pour un musicien d'écouter un son enregistré puis d'effectuer des gestes à l'écoute de ce son pour ensuite réinterpréter le son enregistré au travers de sa propre performance gestuelle. Dans ce chapitre, nous reportons des applications concrètes et implémentées formant des éléments constitutifs de cet *instrument* conceptuel. Pour cela, et en accord avec les contributions issues des chapitres précédents, nous allons porté notre intérêt sur le type particulier de relation geste–son mise en évidence dans le cas de sons abstraits, à savoir une corrélation entre les morphologies de descripteurs gestuels (notamment la vitesse) et les morphologies de descripteurs sonores (l'énergie).

Dans ce chapitre nous commençons par présenter deux applications. Ces applications ne constituent pas un objectif fixé dans cette thèse mais ce sont des **validations de principe** des parties précédentes que nous avons implémentées et présentées dans des conférences. La première est un système de sélection d'un son (décrit par sa morphologie) dans une base de données à partir du geste de l'utilisateur (section 10.2). La seconde est une méthode de re-synthèse de sons enregistrés liée à la morphologie du geste de contrôle (section 10.3). Nous proposons ensuite des perspectives pour l'évolution de ces applications (section 10.4).

Au delà des applications des sections 10.2 et 10.3, nous présentons aussi des travaux annexes **??**. Ces travaux reprennent des idées développées pour la modélisation du geste (partie III) mais en vue d'une application plus étendue, sortant du cadre musical. L'extension porte sur l'analyse des qualités de mouvement dans un contexte d'installation artistique interactive.

10.2 Sélection de sons par les gestes

Dans cette section nous présentons une application des travaux précédents, à savoir l'utilisation de la CCA et d'un modèle d'alignement, pour la sélection de sons dans une base de données à partir d'un geste effectué par un utilisateur. L'idée sous-jacente est de considérer ces méthodes comme des mesures de similarité entre geste et son. La sélection de son par le geste est un cas particulier des requêtes par contenu dans des bases multimedia. Un tel outil peut être ensuite implanté dans des applications de *design* sonore ou des installations interactives. Cette section reporte les travaux de l'article (Caramiaux et al., 2011). Notons qu'un tel système de requête par le geste avait déjà été imaginé par Leman et reporté dans (Leman, 2007) (voir chapitre 7, p.194). Des exemples sont disponibles en ligne à l'adresse `http://baptistecaramiaux.com/blog/?page_id=14#soundselection`.

10.2.1 Concept

Considérons le scénario où un utilisateur imagine un son qu'il aimerait retrouver dans une base de données. Ce son est, par ailleurs, trop abstrait pour être décrit avec des mots (ou *tags*). Cet utilisateur doit trouver un autre moyen par lequel effectuer sa requête. L'idée est que si ce son abstrait possède une morphologie temporelle clairement perceptible, l'utilisateur peut vouloir tracer le son dans l'air ou sur une surface, comme nous l'avons montré dans la partie II. Ainsi, le but de l'outil est de choisir le son dont le profil de descripteurs correspond le plus au profil dessiné par le geste de l'utilisateur. Le problème dans ce cas est une requête par contenu dans une base multimedia. C'est un problème général bien connu dans la littérature. Le geste d'entrée est usuellement appelé *la requête* et le son obtenu *la cible*.

Le problème général des requêtes par contenu dans une base de données multimedia a été beaucoup étudié dans la littérature. Dans la communauté musicale (*Music Information Retrieval* (MIR)), un cas particulier de requête par contenu est la requête par fredonnement (Dannenberg et al., 2007). Les systèmes proposés pour ce problème permettent à l'utilisateur de trouver un morceau de musique dans une base de données en fredonnant une partie de la mélodie. La plupart des travaux dans ce domaine utilisent la notion de *contours* qui est la séquence des différences relatives de hauteurs entre les notes successives de la mélodie. Un autre exemple musical est le problème de requête en tapant le rythme (Jang et al., 2001). Le système est basé sur la détection d'impacts et un alignement temporel entre le rythme détecté dans la requête et celui de chaque son de la base.

Au niveau gestuel, il y a peu de littérature sur des systèmes de requêtes par geste dans une base de données audio, que ce soit dans le domaine des nouvelles interfaces pour l'expression musicale ou MIR. Certaines applications se basent sur des requêtes à partir d'une trajectoire décrivant celle d'un descripteur voulu, par exemple le profil de hauteur d'une flûte. Ces applications utilisent un alignement temporel type DTW comme distance. Nous référons le lecteur au site de l'équipe Représentations Musicales de l'Ircam (`http://repmus.ircam.fr/`) où une telle application est utilisée dans le cadre de l'orchestration automatique.

10.2.2 Prototype

Dans cette section nous présentons le système proposé, les algorithmes, puis l'implémentation, disponible dans le logiciel de programmation graphique Max/MSP.

Le système proposé

La figure 10.1 illustre le système de sélection proposé. Un utilisateur exécute un geste qui correspond, dans la perspective de l'utilisateur, à un son abstrait imaginé. Après un module de pré-traitement, le système rassemble plusieurs algorithmes qui calculent la correspondance multimodale entre séries temporelles gestuelles et sonores. Chaque algorithme extrait une partie spécifique de l'information dans la relation entre geste et son. L'algorithme est un choix de l'utilisateur. Un index correspondant au son trouvé dans la base de données est retourné avec un score qui indique la pertinence de la cible. Finalement, le son le plus pertinent est joué à l'utilisateur. Notons que le son n'est joué qu'après la fin de la requête.

Algorithmes

Nous présentons ici deux algorithmes qui permettent de mesurer la similarité entre la requête gestuelle et les sons de la base de données.

1. **Sélection basée sur la corrélation**. La méthode est basée sur l'analyse canonique présentée dans l'étude exploratoire du chapitre 3 (cf. l'article complet (Caramiaux et al., 2010c)).

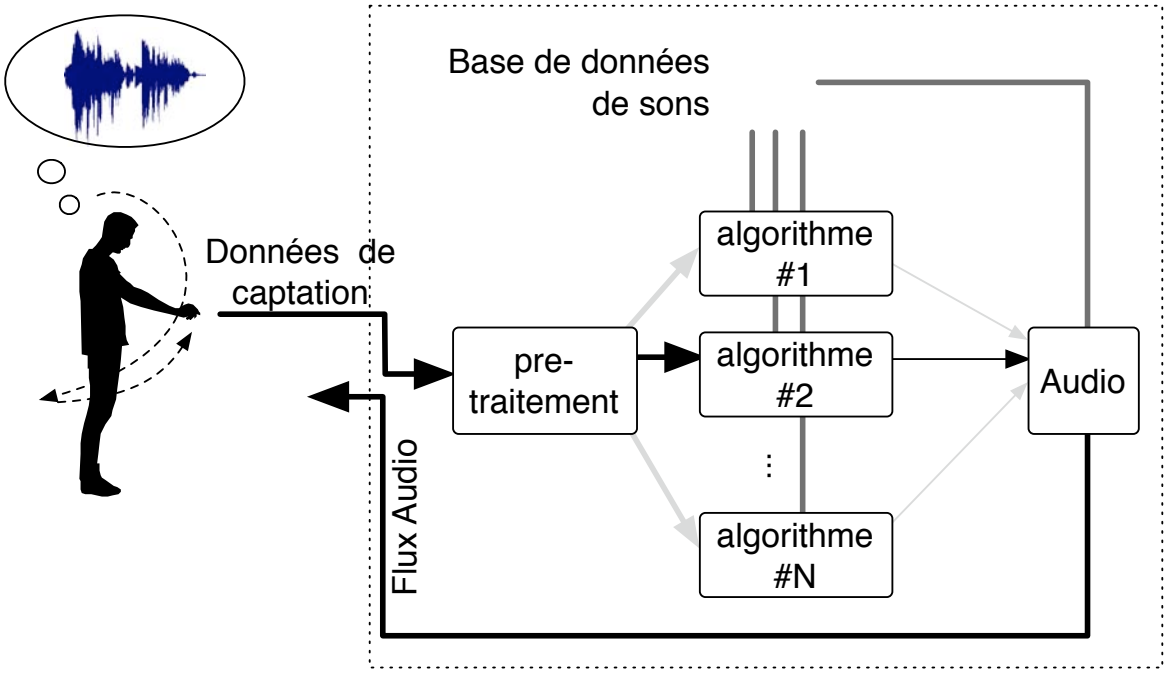

FIGURE 10.1 – Le système de sélection proposé. Un son qui correspond au mieux au geste d'entrée est sélectionné dans une base de données. La correspondance geste–son dépend de l'algorithme utilisé qui est un choix de l'utilisateur.

Un outil de sélection de son basé sur CCA permet en premier lieu l'extraction des descripteurs gestuels et sonores les plus corrélés. Le premier coefficient de corrélation (qui est le plus élevé) est utilisé comme score pour tester la pertinence de la correspondance. Un son est sélectionné si la variation d'une combinaison linéaire de ses descripteurs est similaire à la variation d'une combinaison linéaire des descripteurs gestuels. Comme la corrélation est calculée échantillon-par-échantillon, un score élevé indique aussi que le geste est synchrone au son. Enfin, le son est sélectionné à la fin de l'exécution du geste impliquant le besoin de marquer le début et la fin de celui-ci.

2. **Sélection basée sur l'alignement temporel**. La méthode est basée sur l'alignement temporel en prenant en compte les variations spatiales (cf. algorithme présenté dans le chapitre 8). La méthode permet de préserver la variabilité locale inhérente dans le geste et choisir comme critère de similarité la correspondance des profils temporels dans sa globalité et la cohérence dans les amplitudes. La méthode est efficace d'un point de vue computationnel, elle est multidimensionnelle, temps réel, adaptative et ne requiert pas un apprentissage important. Dans le contexte de l'application, un son est sélectionné si l'utilisateur exécute un geste qui évolue de manière similaire aux descripteurs sonores considérés, mais peut être dilaté localement dans le temps. Cette méthode nécessite de choisir au préalable les descripteurs à comparer ainsi que de normaliser ces descripteurs afin de pouvoir les comparer.

Implementation

Les algorithmes sont implémentés dans l'environnement de programmation temps réel Max/MSP et utilisent la librairie développée dans l'équipe Interactions Musicales en Temps Réel à l'Ircam, appelée MnM (Bevilacqua et al., 2005) (le lecteur intéressé peut trouver cette librairie à l'adresse `http://ftm.ircam.fr/index.php/MnM`). La base de sons est chargée dans

une mémoire tampon multi-pistes (utilisant la technologie MuBu [1] (Schnell et al., 2009)) qui contient N sons avec leurs descripteurs audio. Ces descripteurs sont directement calculés dans Max/MSP grâce à un module d'analyse. Les données de mouvement sont reçus par OSC (*Open Sound Control* [2]) permettant l'utilisation d'un panel large d'interfaces. Quand l'analyse est terminée, le programme retourne l'index du meilleur son trouvé (en fonction de l'algorithme choisi) dans la base de données et ce son est visualisé dans l'éditeur de MuBu. Une capture d'écran reportée dans la figure 10.2 montre l'application avec l'éditeur, la forme d'onde et certains contrôles.

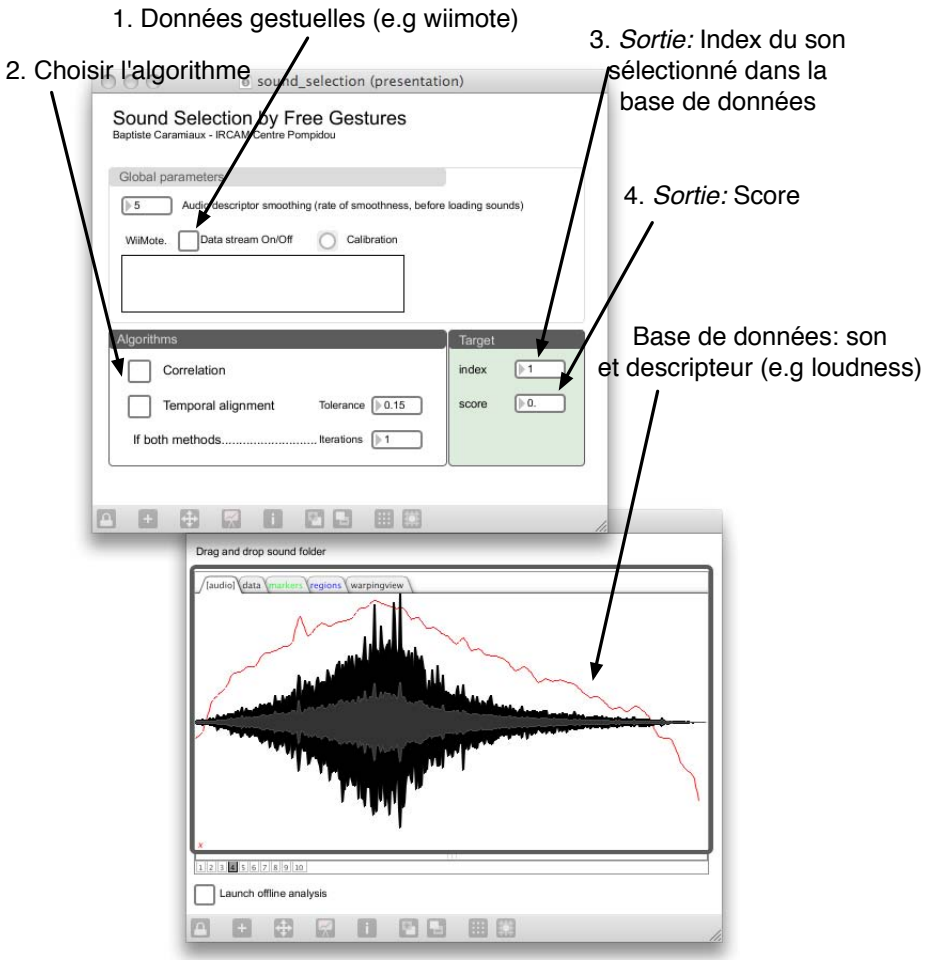

FIGURE 10.2 – Programme Max/MSP pour la sélection de son basée sur une requête gestuelle. L'exemple est donné pour un descripteur audio (l'intensité) et utilise un contrôleur WiiMote. L'utilisateur peut choisir quel algorithme fait la sélection.

10.2.3 Conclusion

L'application présenté est une validation de principe des études sur la relation geste–son et des méthodes de modélisation. Ce travail a été montré en démonstration à NIME 2011 (Caramiaux et al., 2011).

Nous avons présenté une application permettant la sélection de son à partir des gestes d'un utilisateur. L'application calcule la similarité entre l'entrée gestuelle et des sons pris dans une base de données. La similarité est calculée entre descripteurs gestuels et sonores et selon

1. Voir http://imtr.ircam.fr/imtr/MuBu
2. http://www.opensoundcontrol.org

plusieurs algorithmes, permettant une plus grande souplesse pour l'utilisateur. Une version a été développée dans le programme Max/MSP en utilisant la librairie MnM.

Les stratégies mises à jour dans l'association de gestes à des sons environnementaux prennent en compte si une action causale (source du son) peut être ou non reconnue. Ces stratégies devront être intégrées à un tel système de sélection par requête. Cette extension n'a pas été effectuée dans ce travail de thèse mais est une perspective à court-terme. D'autres perspectives sont à envisager pour le développement d'une réelle application, potentiellement commerciale, et particulièrement une validation utilisateur. Ceci sera précisé dans la section 11.2.

10.3 Sélection temps réel et adaptation

La deuxième application présentée dans ce chapitre est une méthode de re-synthèse de sons enregistrés liée à la morphologie du geste d'un utilisateur. L'idée principale de cette application est inspirée de la deuxième méthode de sélection présentée précédemment. La méthode d'alignement peut être temps réel, comme celle présentée dans le chapitre 8. Ainsi, la sélection aussi est temps réel et le son sélectionné peut être joué à mesure que le geste est effectué. Les variations estimées par la méthode servent à moduler la synthèse. Cette section reporte les travaux de l'article (Caramiaux et al., 2010a) reporté dans l'annexe D. Des exemples sont disponibles en ligne à l'adresse `http://baptistecaramiaux.com/blog/?page_id=14#morphresynth`.

10.3.1 Concept

Comme précédemment, l'application s'inspire d'un scénario imaginé. Un utilisateur a, à sa disposition, un ensemble de sons abstraits décrits par leurs morphologies acoustiques, dans une base de données. Cet utilisateur pense à un son qu'il aimerait obtenir à partir de sa morphologie. Il peut alors imiter cette morphologie en effectuant un geste et le système joue le son de la base de données dont la morphologie correspond le plus au geste et à mesure que celui-ci est effectué. Les variations temporelles et spatiales du geste, comme des ralentissements ou accélérations, sont retranscrites dans la synthèse. Cette application permet le respect de la structure temporelle à la fois du geste et du son.

Le problème de re-synthèse de sons enregistrés à partir du geste n'a pas été beaucoup traité dans la littérature. Les systèmes utilisant des technologies similaires sont plutôt issus du domaine artistique. Par exemple, Bevilacqua et al. utilisent un système d'alignement temps réel geste–geste (Bevilacqua et al., 2010) pour la resynthèse d'un son basée sur la position du suivi dans le geste de référence.

10.3.2 Prototype

Le système proposé

La figure 10.3 illustre le système de re-synthèse de sons. Un utilisateur pense à un geste et l'effectue. Un système de captation envoie les données échantillonnées au système. Celui-ci calcule les descripteurs utilisés dans l'alignement temporel entre le geste et le son. La méthode d'alignement temps réel est celle présentée dans le chapitre 8. Chaque template est un ensemble de profils de descripteurs sonores d'un son dans la base de données. Ainsi, il y a autant de références que de sons dans la base. À chaque nouvelle observation gestuelle, le système estime la position dans le son le plus probable : c'est le processus de sélection effectué en temps réel. Cette position spécifie une trame sonore qui est jouée. De plus, le modèle permet l'adaptation d'autres caractéristiques aux variations gestuelles qui peuvent être utilisées pour la synthèse comme la vitesse ou l'amplitude. Le moteur de synthèse utilisé est la synthèse par vocodeur de phase avec préservation des transitoires (Röbel, 2003).

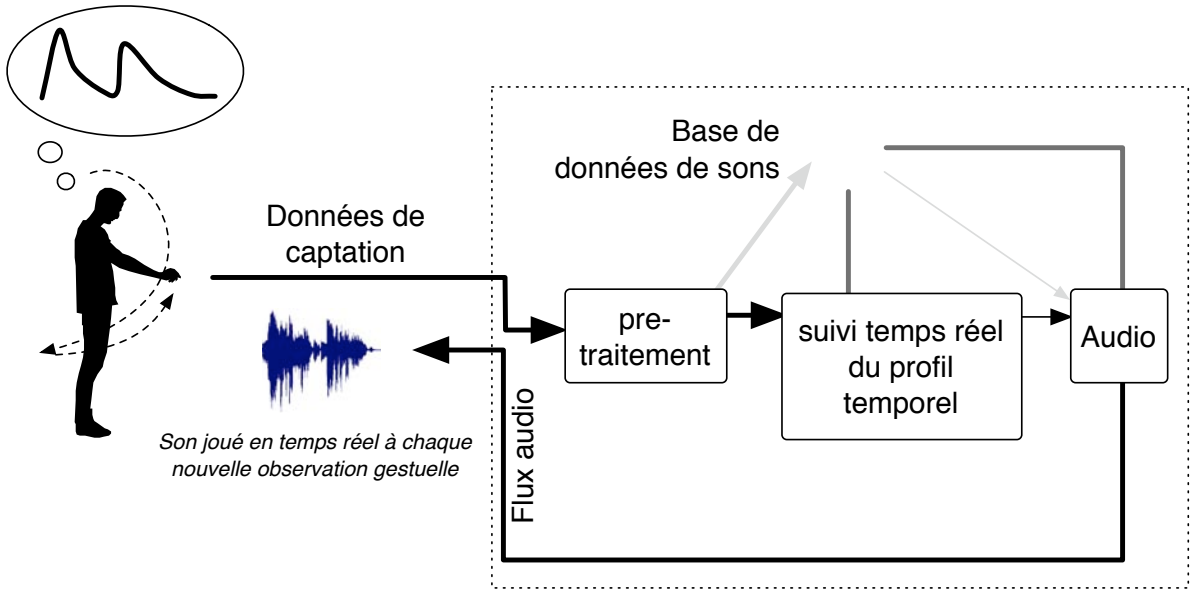

FIGURE 10.3 – Le système de re-synthèse d'un son enregistré. Lorsque le geste est effectué, un son est joué en temps réel dont la morphologie correspond au mieux à la morphologie du geste (de ces paramètres).

Exemple

Afin d'éclaircir le discours, on présente ici un exemple concret de son aligné sur le geste en explicitant l'alignement. On prend l'exemple d'un geste effectué à l'écoute d'un son provenant de l'expérience présentée dans le chapitre 5. Dans le corpus de sons, nous prenons celui correspondant aux grains de riz versés dans un bol mais après transformation [3]. Parmi les performances des 11 candidats, nous prenons comme exemple la norme de la vitesse du candidat 4. Ce descripteur sera mis en correspondance avec l'intensité sonore du son écouté (en accord avec les conclusions des chapitres 5 et 6). La figure 10.4 illustre les deux signaux pris pour l'exemple. Sur cette figure, la courbe tracée en noir est l'intensité sonore du son considéré et celle en gris est la norme de la vitesse de la main du participant.

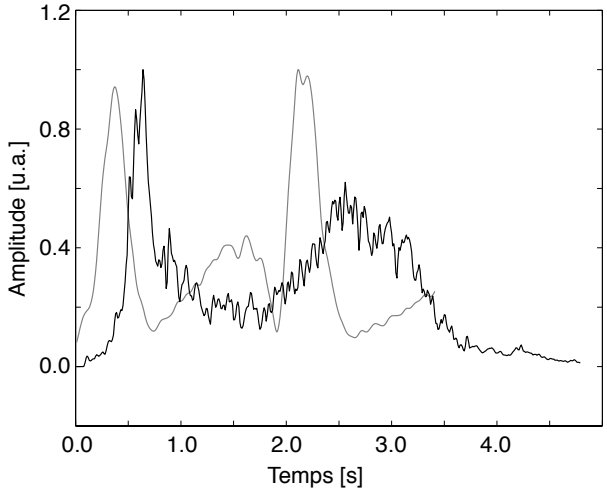

FIGURE 10.4 – Données de référence et de test. En noir est l'intensité sonore prise comme référence dans le modèle. En gris est représentée la courbe de vitesse du geste effectué à l'écoute du son. Cette courbe est la donnée d'observation.

3. Pour rappel, cette transformation consiste en une analyse en bandes Mel, du son original, par lesquelles nous faisons passer un signal audio de bruit blanc.

L'intensité sonore est la référence et la vitesse absolue correspond au test. Les séries temporelles sont alignées de manière causale par l'algorithme présenté dans le chapitre 8. Celui-ci permet l'alignement temporel de deux séries temporelles toute en adaptant certaines caractéristiques des signaux. Ici, trois paramètres sont adaptés : la position, la vitesse et l'amplitude du signal. Le résultat est visualisé sur la figure 10.5. A gauche, nous avons reporté les données initiales. Au centre, deux graphiques représentent l'adaptation : en haut est la position (donc l'alignement entre les séries temporelles) ; en bas est la courbe d'adaptation de l'amplitude. Enfin à droite, le graphique montre la vitesse absolue inchangée (en gris) et l'intensité sonore alignée, dont l'amplitude varie. L'alignement temporel donne les

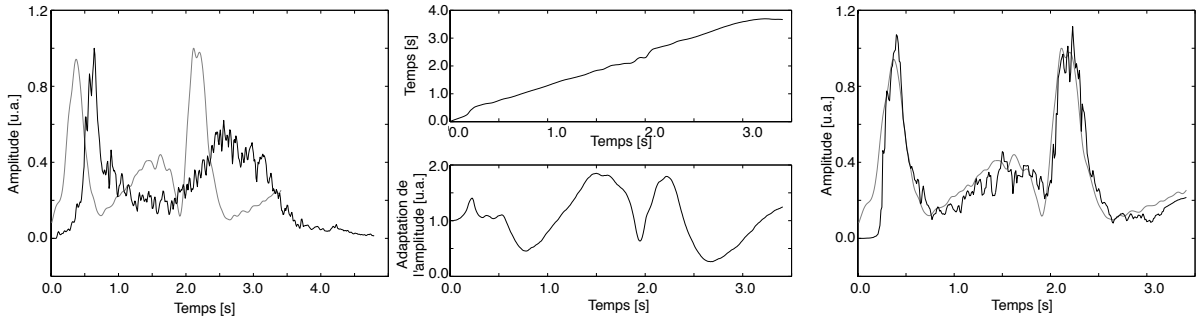

FIGURE 10.5 – Résultat de l'alignement pour la resynthèse. À gauche, sont illustrées les courbes d'intensité et de vitesse initiales. Au milieu en haut, nous avons la courbe d'alignement et en dessous l'adaptation de l'amplitude inférée par l'algorithme. À droite, nous avons la courbe d'intensité sonore *adaptée* à la vitesse.

positions des grains à lire dans le fichier son original et l'amplitude peut être liée au volume de sortie du lecteur. Un exemple sonore peut être écouté en ligne à l'adresse suivante http://baptistecaramiaux.com/blog/?page_id=14#morphresynth.

10.3.3 Conclusion

L'application présentée est une validation de principe de la méthode l'alignement temporel en temps réel dans un contexte multimodal geste–son. Cette application est issue d'un travail publié à la conférence SMC en 2010 (Caramiaux et al., 2010a) où un tel modèle est étudié en tant que mesure de divergence entre ces deux modalités.

Nous avons proposé une méthode de resynthèse utilisant l'algorithme présenté dans le chapitre 8. Premièrement, l'algorithme utilisé est basé sur un alignement temporel et adaptatif entre les données gestuelles et les descripteurs sonores. Deuxièmement, la synthèse par vocodeur de phase rend possible le pilotage de la tête de lecture dans un fichier de son en évitant des artefacts indésirables dûs à des problèmes de phase. Enfin l'algorithme d'alignement adapte d'autres caractéristiques des signaux qui peuvent être utilisées comme contrôle expressif de la synthèse, comme par exemple l'amplitude.

10.4 Perspectives pour les applications présentées

Pour l'application de sélection, un travail plus approfondi sur une méthode hybride entre extraction de descripteurs et alignement temporel, doit être effectué. En effet, les deux modèles proposés (CCA et alignement) sont « orthogonaux » : l'analyse canonique se base sur la synchronisation qui peut être obtenue avec l'alignement temporel des signaux, mais l'alignement repose sur un choix pertinent des descripteurs gestuels et sonores à aligner, ce qu'apporte l'analyse canonique. Ainsi, une solution serait de rendre dynamique, c'est à dire d'intégrer au modèle temporel, la réduction de dimension multimodale. Cela peut être fait par un modèle stochastique où les probabilités d'observation utilisées dans les modèles temporels

dépendraient des matrices de réduction de dimension. Les matrices de projection sont ainsi adaptées. Un modèle s'en approchant a été proposé récemment par Wang et al. dans (Wang et al., 2008).

Pour la resynthèse, une perspective envisagée est l'ajout d'un modèle de structure à l'échelle du segment ajouté au modèle précédent. Ceci permettrait d'avoir accès à un niveau *macro* de la performance musicale (comme indiqué par Jordà dans (Jordà, 2008)). La structure plus haut niveau pourrait être définie a priori, comme par exemple issue d'une composition gestuelle écrite, ou alors elle pourrait être apprise à la volée par des méthodes d'apprentissage automatique.

Enfin, un système mettant en lien les morphologies du geste et du son s'insère dans l'instrument conceptuel présenté. En effet, on peut imaginer que le mouvement effectué à l'écoute de la musique (supposée multipistes) peut prendre en main certaines pistes plus texturelles. Dans ce cas, le contrôle n'est pas sur tout le phénomène musical perçu mais certains de ses éléments.

Chapitre 11

Conclusion

« A kind of synthesis, but with some elements that perhaps you wouldn't have expected in advance. I always like that when that happens, when something comes that is more than the sum of the parts. »

– Evan Parker

11.1 Synthèse générale

Le travail de thèse présenté dans ce manuscrit s'est basé sur un *instrument* de musique conceptuel qui nous a permis de définir un cadre général pour les problématiques traitées. Ces problématiques sont centrées autour de la relation geste–son en performance musicale amenant à l'étude des thématiques principales suivantes : le geste en réaction au son, la modélisation du geste, le contrôle gestuel de la synthèse sonore.

11.1.1 Partie I

Cette thèse commence par proposer un état de l'art composite autour des thématiques mises en jeu dans la performance musicale avec l'outil informatique, à savoir une revue de la littérature autour de la conception des instruments de musique numériques, du geste musical (point d'entrée de l'instrument) et du lien entre action et perception (lien fondamental dans la pratique musicale). Ce qui nous intéresse est de mieux comprendre les relations entre geste et son. Notre démarche a donc été d'examiner quels gestes sont associés aux sons que nous voudrions contrôler avec un instrument de musique numérique. Nous avons mené une étude exploratoire ayant comme objectif d'étudier les gestes en réponse à des sons variés (musicaux, environnementaux) pris dans un corpus. Cette méthode était une analyse multimodale non-supervisée, appelée analyse canonique, entre les descriptions numériques du geste et des sons enregistrés. La méthode a permis de mettre en avant plusieurs questions touchant à la fois aux stratégies cognitives pour l'association d'un geste à un son et à la fois à la modélisation des structures temporelles du geste. Dans les parties suivantes nous avons apporté des éléments de réponse à ces questions par le biais d'expérimentations, de modélisations et d'applications.

11.1.2 Partie II

Dans cette partie, nous présentons les études expérimentales de la thèse. Elles visent à étudier les gestes effectués en réponse aux sons en spécifiant le stimulus afin de faire apparaître différentes stratégies de relations entre geste et son.

Nous avons tout d'abord construit un corpus de sons comprenant à la fois des sons causaux (dont l'action est identifiable) et non-causaux (dont l'action n'es pas identifiable). Ce deuxième corpus a été obtenu par une transformation spectrale sur le signal audio. Les corpus ont été

validés par un test d'écoute. Ces corpus ont ensuite été utilisés dans l'expérience où des participants devaient effectuer des gestes en réponse aux sons du corpus. Nous avons montré que des stimuli, dont l'action causale était identifiable, amenaient les participants à imiter cette action. Ainsi l'action est encodée dans les propriétés acoustiques du son et décodée par les participants. Le décodage est cependant idiosyncratique. Dans le cas où l'action n'est pas identifiable, la relation entre geste et son est plus directe, c'est à dire que les participants tracent le son : leur mouvement suit l'évolution des descripteurs acoustiques. Dans ce cas, les gestes effectués (en considérant un son) par les participants sont consistants, ce qui va dans le sens de l'étude exploratoire de la première partie. Cette deuxième stratégie a ensuite été examinée de manière plus détaillée.

Pour cela une deuxième expérience a été présentée. Les sons utilisés dans cette expérience étaient synthétisés à partir de trajectoires, données a priori, pour la hauteur, l'intensité et la brillance. Ces trajectoires sont par exemple *monte*, *descend*. L'expérience a révélé des stratégies qui avaient déjà été imaginées (mais non validées) dans l'étude exploratoire. Outre la consistance des gestes entre les participants, nous avons montré que dans le cas de stimuli avec une évolution de la hauteur perceptible, le contrôle est plus généralement basé sur la position alors que dans le cas où la hauteur n'est pas pertinente, le contrôle est plus complexe, notamment la vitesse joue un rôle important et est souvent liée à l'intensité sonore ou à la brillance. Dans l'espace sonore, trois classes de contrôle apparaissent : contrôle de la brillance, de l'intensité ou des deux. En revanche, le contrôle gestuel est basé soit sur la position, soit sur la vitesse.

11.1.3 Partie III

Dans cette partie nous nous focalisons sur l'analyse du geste et sa modélisation. Les études visent à proposer des modèles pour les structures temporelles à différentes échelles.

Tout d'abord nous avons présenté une méthode de reconnaissance de gestes représentés comme des séries temporelles multidimensionnelles. La méthode repose sur la définition de classes gestuelles utilisant un seul template. Notre méthode aligne le geste d'entrée sur les templates utilisant le filtrage particulaire. Contrairement aux méthodes standards basées sur la programmation dynamique (DTW, HMM), l'algorithme permet d'adapter et de suivre en temps réel les variations spatiales et temporelles du geste. Nous avons mené différentes évaluations en utilisant : des données synthétiques, des gestes bidimensionnels sur une surface tactile, et des gestes tridimensionnels provenant d'accéléromètres. Les résultats montrent que la méthode proposée obtient de meilleures performances pour la reconnaissance que les méthodes standards et ceci grâce à l'adaptation aux variations. De plus, la méthode met à jour de manière continue les résultats de l'adaptation et de la reconnaissance ce qui offre de grandes possibilités d'utilisation pour des systèmes interactifs tels que ceux en Interaction Homme-Machine.

La deuxième contribution concerne l'utilisation du modèle segmental pour la segmentation et l'analyse syntaxique des gestes ancillaires du clarinettiste. Ces gestes sont liés à la musique, conformément à l'état de l'art, nous faisons l'hypothèse qu'ils ont une structure liée à la structure musicale. Nous avons montré que le geste ancillaire pouvait se décomposer en une séquence de trajectoires prises dans un dictionnaire. La construction de ce dictionnaire peut être aidée de critères telles que la distance euclidienne entre la séquence resynthétisée et le geste original ainsi que la log-vraisemblance de la séquence obtenue. Nous avons ainsi montré qu'un dictionnaire exhaustif n'est pas gage d'une meilleure modélisation. La séquence de primitives obtenue fait apparaître des motifs récurrents chez l'instrumentiste et pour plusieurs instrumentistes : nous examinons particulièrement des cercles réalisés avec le pavillon de la clarinette. Ces motifs sont consistants et entre-coupés de motifs de coarticulations entre les phrases de la partition. De même, pour un clarinettiste les cercles couvrent une mesure alors que pour un autre clarinettiste ils couvrent deux mesures. Ainsi, le modèle segmental

permet l'analyse de gestes expressifs, tels que les gestes ancillaires, d'un niveau signal à un niveau symbolique, ouvrant des perspectives pour l'analyse syntaxique et pour le contrôle haut niveau de la musique.

11.1.4 Partie IV

Deux applications ont été présentées. La première regroupe l'utilisation de l'analyse canonique et de l'alignement temporel pour une application de sélection de son dans une base de données à partir d'une requête gestuelle. La seconde propose l'utilisation de l'alignement en temps réel pour la synthèse. Cette deuxième application permet la synthèse morphologique d'un son. L'outil de suivi conduit une tête de lecture dans le fichier audio au fur et à mesure que le geste est effectué. Ces applications ne sont pas figées mais sont construites de manière à accepter différentes méthodes, telles des boîtes à outils pour la conception d'interactions.

11.1.5 Schéma de synthèse

Le schéma de la figure 11.1 réunit les contributions synthétisées précédemment de manière graphique. Ce schéma se lit de haut (le geste) en bas (le son). Nous avons reporté les stratégies de représentation par un geste physique du geste sonore. Ces stratégies sont liées à différents types de contrôle qui peuvent ensuite être pris en compte dans la modélisation, celle-ci pilotant la synthèse.

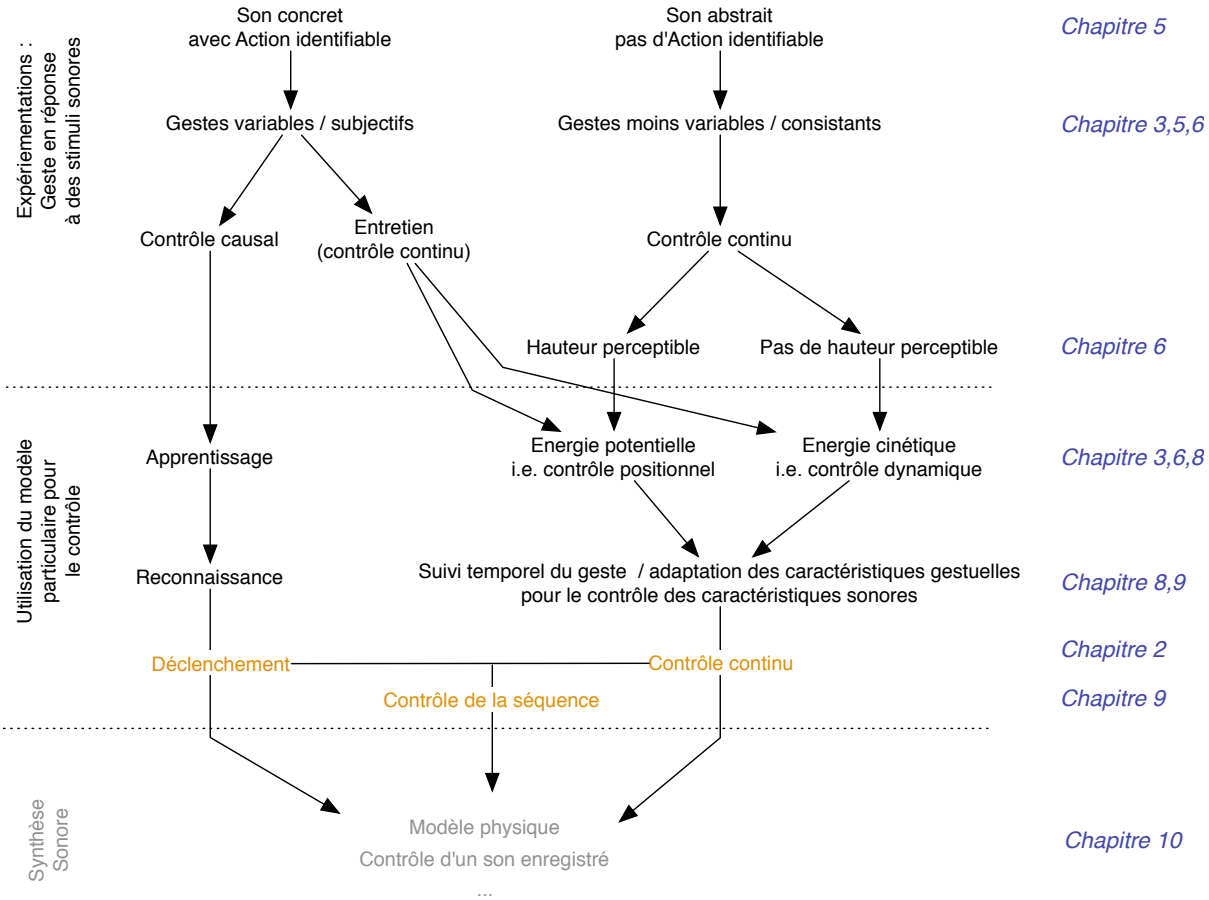

FIGURE 11.1 – Schéma réunifiant les résultats expérimentaux et la modélisation.

11.2 Perspectives

De ce travail de thèse, nous pouvons dégager un certain nombre de perspectives concrètes pour des recherches futures. Ces perspectives se structurent autour des relations entre geste et son suivant trois grands axes clairement identifiés dans le manuscrit, à savoir : l'étude expérimentale, la modélisation, et les applications.

Le premier axe porte sur des aspects en sciences cognitives et reprend les travaux sur les stratégies gestuelles en réponse à des stimuli sonores. Il serait intéressant de continuer ces travaux en s'orientant vers le problème de la coarticulation gestuelle. Ce problème examine comment des gestes, pris de manière isolée, sont déformés lorsqu'ils sont effectués en séquence. On peut s'interroger sur le lien qu'il peut exister entre des articulations au sein du stimulus sonore et son influence sur le geste. Une méthodologie similaire à celle présentée dans ce manuscrit est un bon point de départ.

Indépendamment, il serait pertinent d'examiner en détails les liens avec les neurosciences, notamment dans le cadre d'un projet en collaboration avec des acteurs du domaine. Nous avons présenté dans les états de l'art, des travaux liés aux neurosciences pour le lien entre action et perception auditive. Nous pensons que le cadre défini dans cette thèse peut induire des avancées plus fondamentales dans l'étude de la relation entre geste et son pour la performance musicale. Les travaux en neuroimagerie portant sur les liens entre mouvements et musique sont peu nombreux, et les implications peuvent être intéressantes comme l'étude de l'apprentissage sensori-moteur dans la performance musicale.

Le second axe porte sur la modélisation du geste et notamment l'investigation de modèles hiérarchiques souples pour l'analyse du geste et le contrôle temps réel du son et des structures. En première approche, nous pensons que le modèle dynamique avec filtrage particulaire peut être lié avec une structure plus haut niveau modélisant les transitions entre gestes primitifs (à l'instar du modèle segmental). Nous avons commencé l'étude du modèle hiérarchique pour l'analyse du geste (Françoise et al., 2011). Dans ce travail, nous avons pu constater que lorsque les variations temporelles sont grandes (gestes effectués par différentes personnes, à différents tempi), le modèle hiérarchique donne de meilleurs résultats que le modèle segmental dans une tâche de segmentation et pour des données d'accéléromètres. Une première démonstration a aussi été réalisée (Françoise et al., 2012). De manière théorique, il s'agit d'effectuer une comparaison formelle pour la segmentation entre le filtrage particulaire les modèles segmentaux ou hiérarchiques ou des autres méthodes non basées sur des modèles comme le *cusum* (pour la détection de changement dans un processus stochastique) ou les processus de Dirichlet pour la segmentation (Buchsbaum et al., 2009).

Dans le même axe, et de manière complémentaire aux modèles cités dans le paragraphe précédent, il serait pertinent d'envisager des méthodes d'apprentissage pour l'étude des relations entre geste et son. De récentes avancées dans l'apprentissage actif, faisant intervenir l'utilisateur, sont une direction indiquée pour envisager ce problème (e.g. apprentissage par renforcement (Sutton, 1984) avec des applications musicales (Cont et al., 2007; Le Groux and Verschure, 2010)).

Le dernier axe porte sur les applications. Les validations de principe présentées dans cette thèse sont prometteuses. Il s'agit dans la suite de développer une application plus complète et modulaire (comme celle présentée par le schéma 11.1). La validation utilisateur des applications présentées dans ce manuscrit était en dehors du cadre de nos recherches mais est à envisager. Dans cette optique, il sera intéressant d'étudier la pertinence des modèles précédemment proposés, et intégrés à l'application, d'un point de vue utilisateur. Ce type d'études a eu un grand intérêt récemment (Fiebrink et al., 2011). Elles interrogent sur la façon dont l'utilisateur peut agir sur des systèmes interactifs qui mettent en jeu des méthodes d'apprentissage automatique où certains paramètres doivent être manipulés afin d'obtenir le comportement escompté.

En outre, ces applications peuvent se décliner pour différents publics et utilisations : des interfaces portables et ubiquitaires à des applications en spectacle vivant. Différentes facettes de la même application pourront être envisagées suivant si elle se dédie au grand public ou à des experts.

Cinquième partie

Annexes et Articles Complémentaires

Annexe A

Modèles Bayésiens

A.1 Méthodes statiques

Les méthodes d'analyse de données numériques sont dites statiques lorsqu'elles ne prennent pas en compte l'ordre temporel au sein des données. En d'autres termes, n'importe quelle permutation appliquée sur les données n'influe pas le résultat de la méthode. L'hypothèse est que ces données sont les réalisations d'une ou plusieurs variables aléatoires. Les méthodes statiques sont utilisées pour des applications telles que la fouille de données, la réduction de dimension, la classification, l'extraction de caractéristiques ou encore la sélection de caractéristiques. Dans cette thèse, nous les présenterons dans le cadre d'une tâche de réduction de dimension permettant l'extraction de caractéristiques.

A.1.1 Réduire la dimensionalité

Les moyens de captation offerts pour la transcription de grandeurs physiques gestuelles en données numériques donnent lieu à une grande quantité de données souvent redondantes. Un exemple est l'ensemble de données provenant d'un système de captation de mouvement 3D par caméra infra-rouge. Dans ce type de système, chaque marqueur placé sur le corps du participant et/ou son instrument de musique donne lieu à trois valeurs (une pour chaque dimension) décrivant la position à un certain instant t. Lors d'une séance de collecte de données de n minutes, le nombre de données revient à 3 (le nombre de dimension) par le nombre de marqueurs M par la fréquence d'échantillonnage f_s par le nombre de secondes de captation $60 \times n$. Parmi cet ensemble, beaucoup de marqueurs sont dépendant. Ceci est dû, entres autres, aux contraintes biomécaniques du corps. Il est ainsi primordial de pouvoir réduire le nombre de paramètres pour l'analyse.

Parmi les méthodes statiques, certaines permettent la réduction de dimension d'un ensemble de données initiales. Dans le formalisme des méthodes de réduction de dimension, les paramètres (gestuels) sont aussi appelés variables aléatoires ou caractéristiques (ang. *features*) (Fodor, 2002). Ces méthodes se divisent en deux familles distinctes : la sélection de caractéristiques (ang. *feature selection*) ou l'extraction de caractéristiques (ang. *feature extraction*). La première trouve un sous-ensemble des variables initiales. La seconde transforme les données d'un espace de dimension supérieure en une représentation dans un espace de dimension inférieure.

Une méthode statique simple et très utilisée pour la réduction de la dimension de données par extraction de caractéristiques est l'analyse par composantes principales (dans la littérature *Principal Component Analysis*, appellation abrégée en PCA).

A.1.2 Analyse en composantes principales

Description

La PCA permet la représentation d'un ensemble de variables corrélées en un ensemble de variables non corrélées. Ainsi, cette méthode se base sur la covariance et les corrélations au sein d'un ensemble de variables aléatoires. Notons $\mathbf{Y}$ l'ensemble des réalisations de ces variables aléatoires initiales. La matrice $\mathbf{Y}$ est de dimension $n \times m$ où n est le nombre de réalisations et m le nombre de variables. Les variables $\mathbf{y}_i$ (colonnes de $\mathbf{Y}$), $1 \leq i \leq m$ sont corrélées entre elles et certaines varient plus que d'autres (i.e. véhiculent plus d'information). L'idée est qu'au lieu d'inspecter toutes les corrélations ($1/2m(m-1)$ combinaisons) ou covariances, il est plus intéressant d'analyser un sous-ensemble de p variables conservant la plupart de l'information donnée par les corrélations et covariances dans les variables initiales. la PCA est une méthode linéaire c'est à dire que les variables extraites sont des combinaisons linéaires des variables initiales. Ainsi, la PCA trouve une première composante $\mathbf{Ya}_1 = \sum_{i=1}^{m} a_{i1}\mathbf{y}_i$ qui maximise la variance var($\mathbf{Ya}_1$) puis une deuxième $\mathbf{Ya}_2$ non corrélée à la première qui a la deuxième plus grande variance, etc... Si on note $\mathbf{A}$ la matrice des coefficients des combinaisons linéaires, on peut écrire :

$$\mathbf{X} = \mathbf{YA} \tag{A.1}$$

Où les composantes principales $\mathbf{x}_i$ (colonnes de $\mathbf{X}$) pour i allant de 1 à N sont non corrélées entre elles et ordonnées de telle façon à avoir les variances des composantes rangées par ordre décroissant.

État de l'art

Dans la littérature, la PCA a été utilisée pour représenter des données (ici de mouvements) de dimension élevée comme la forme, des trajectoires dynamiques de mouvements, l'apparence, etc. Les données de mouvement de dimension élevée se retrouvent naturellement dans le domaine de la vision. L'analyse de scènes vidéo induit une description dans un espace de dimension élevée (e.g. l'image elle-même) dans lequel seul un sous-ensemble local a un intérêt pour l'application (par exemple la reconnaissance). Il ne s'agit pas ici de faire une revue des travaux du domaine de la vision ayant utilisés la PCA mais seulement de donner des clés de compréhensions des utilisations possibles. On réfère le lecteur à l'article de De La Torre et al. (De La Torre and Black, 2003) qui reprend des travaux utilisant la PCA et certaines extensions pour des données vidéos. L'idée principale étant l'apprentissage d'un sous-espace de représentation pour la reconnaissance, le suivi, la détection ou la modélisation d'arrière-plan. Une application intéressante pour notre propos est l'utilisation de la PCA pour la segmentation. Barbič et al. (Barbič et al., 2004) ont proposé deux approches *en-ligne* (i.e. que le système parcourt le geste du début à la fin et effectue la segmentation pendant ce parcours) basée sur la PCA. La première utilise la version de la PCA présentée dans cette thèse. L'idée est d'inspecter la dimension du sous-espace généré par la méthode. Lorsque cette dimension change, il définit la fin du segment précédent et le début d'un nouveau. La seconde méthode utilise une version probabiliste de la PCA (Roweis, 1998; Tipping and Bishop, 1999). L'idée est similaire : la distribution du mouvement est estimée comme étant une gaussienne centrée sur la moyenne du mouvement sur un certain nombre de trames vidéo et la covariance est calculée à partir des composantes principales données par PCA. Dès que le mouvement semble ne plus suivre cette gaussienne, un nouveau segment commence avec une nouvelle estimation de la moyenne et de la covariance.

La PCA a été utilisée dans le domaine de l'informatique graphique et plus particulièrement pour la synthèse de mouvement. En effet, les composantes principales calculées à partir de données de captation 3D forment des vecteurs de mouvement qui peuvent être linéairement

combinés pour former de nouveaux mouvements valides (Urtasun et al., 2004). Alexa et al. (Alexa and Müller, 2000) proposent une méthode d'animation basée sur la concaténation de composantes principales. Glardon et al. (Glardon et al., 2004) utilisent cette méthode pour la synthèse de la marche. De même Urtasun et al. (Urtasun et al., 2004) étendent la PCA pour la modélisation et la synthèse de style dans la marche, la course ou les sauts. Cependant la PCA dans ces cas d'utilisation n'a pas de composantes interprétables en termes de marche ou course.

Comme nous l'avons mentionné précédemment, la captation de mouvement 3D par caméra infra-rouge et marqueurs réfléchissant placés sur le corps retourne un grand nombre de données redondantes. La PCA aide a trouver quelle partie du mouvement est la plus informative. Cette méthode a été utilisée pour l'analyse clinique biomécanique (Daffertshofer et al., 2004), pour l'analyse cinématique de mouvements dansés (Hollands et al., 2004). De même Toivainen et al. (Toiviainen et al., 2010) étudie les composantes principales de danseurs afin d'extraire la métrique dans le mouvement et la lier à la musique. La représentation réduite et compacte donnée par la PCA peut ainsi être utilisée à des fins de contrôle. Par exemple dans (Bevilacqua et al., 2003), les auteurs proposent l'analyse de mouvements dansés en composantes principales utilisées pour piloter un moteur de synthèse. Dans la recherche musicale, où les systèmes de captation de mouvement 3D sont très utilisés, certains auteurs utilisent la PCA pour extraire les composantes du mouvements prédominants lors d'une performance musicale et lier ces composantes (du moins leurs variations) à la partition (MacRitchie et al., 2009).

A.1.3 Autres méthodes

Une extension classique de la PCA est l'utilisation de noyaux. Une autre extension est le changement d'espace de projection. Les données ne sont plus projetées dans un espace euclidien mais sur une variété Riemanienne. Cette extension de la PCA est appelée Analyse en Géodesique Principale (Fletcher et al., 2004) et a été utilisée pour la compression numérique de données de mouvements 3D (Tournier et al., 2009). D'autres méthodes de réduction de dimension basée sur une géometrie non-euclidienne existent : *Local Linear Embedded* (Roweis and Saul, 2000) et Isomap (Tenenbaum et al., 2000). Cette dernière méthode a par ailleurs été utilisée pour le mouvement humain et la robotique dans (Jenkins and Matarić, 2004).

A.2 Modèles de Markov à états cachés

Les modèles de Markov à états cachés (ou *Hidden Markov Models* (HMMs)) ont eu beaucoup de succès pour la modélisation des séries temporelles pour l'analyse, la synthèse et la reconnaissance. Ces modèles ont des avantages qui seront rappelés dans cette partie, notamment la représentation compacte et souple de données temporelles dans lesquelles il y a des incertitudes. Parmi ses domaines d'application, les HMMs ont été utilisés dans des systèmes pour la reconnaissance et notamment dans la parole (Rabiner, 1989), le geste (Yamato et al., 1992), la biologie (Durbin, 1998), etc.

A.2.1 Formalisation

Un HMM est un modèle génératif pour des séries temporelles. Si on se donne une séquence d'observations $\mathbf{y}_{1:T}$ (observations à la sortie du système à étudier), on fait l'hypothèse qu'elles ont été générées par un processus Markovien caché. On peut alors étudier la probabilité que la séquence soit générée par le HMM et retrouver la séquence optimale d'états cachés la générant. Dans le cas présent, les observations sont des données continues et sont supposées suivre

une distribution de probabilité Gaussienne. Une série temporelle discrète $q_{1:T}$ de T états est inférée par le HMM. Cette série est un processus Markovien d'ordre 1 qui suit une distribution multinomiale sur l'ensemble des états. L'ensemble des états est fini, noté $\mathcal{Q}$ et de cardinal N. Pour rappel un processus de Markov d'ordre 1 respecte la loi de probabilité conditionnelle suivante :

$$p(q_t = k|q_{t-1} = j_{t-1}, \ldots, q_0 = j_0) = p(q_t = k|q_{t-1} = j_{t-1})$$

Où $k, j_0, \ldots, j_{t-1} \in \mathcal{Q}$. En d'autres termes, l'information à l'instant t ne dépend que de l'information à l'instant $t-1$ ($t-n$ dans le cas général d'un processus de Markov d'ordre n). Le modèle définit la probabilité jointe sur les états et les observations $p(\mathbf{y}_{1:T}, q_{1:T})$ par :

$$p(\mathbf{y}_{1:T}, q_{1:T}) = p(q_1)p(\mathbf{y}_1|q_1) \sum_{t=2}^{T} p(q_t|q_{t-1})p(\mathbf{y}_t|q_t) \tag{A.2}$$

La formulation de l'équation (A.2) est "réduite" car les probabilités sur les états comme $p(q_t)$ s'écrit normalement $p(q_t = i)$ pour $i = 1 : N$. Un HMM se caractérise par

1. $p(q_1)$: la probabilité a priori sur les états (notée plus précisément par $\pi_i = p(q_1 = i)$)
2. $p(q_t|q_{t-1})$: la probabilité de transition d'un état à un autre. On notera $a_{ij} = p(q_t = j|q_{t-1} = i)$ se résumant à une matrice carrée $A = (a_{ij})_{1 \leq i,j \leq N}$.
3. $p(\mathbf{y}_t|q_t = i)$: la probabilité d'observation, notée $b_j(\mathbf{y}_t)$, qui se résume à une matrice $B = (b_j(\mathbf{y}_t))_{1 \leq t \leq T, 1 \leq i \leq N}$.

Un HMM sera noté $\lambda = (A, B, \pi)$, en accord avec l'article de Rabiner (Rabiner, 1989). Un état discret génère des observations suivant une loi Gaussienne (pouvant être multidimensionnelle).

A.2.2 Inférence et apprentissage

Nous rappelons dans cette partie les principaux algorithmes d'inférence pour les HMMs. Une description plus complète est donnée dans le l'article de Rabiner et al. (Rabiner, 1989) ou sous forme de réseaux bayésiens dynamiques dans la thèse de Murphy (Murphy, 2002) et modèles graphiques (Bilmes, 2002).

Algorithme *Forward-Backward*

Etant donnée une séquence d'obervation $\mathbf{y}_{1:T}$, et un HMM $\lambda = (A, B, \pi)$, on aimerait calculer la probabilité $p(\mathbf{y}_{1:T}|\lambda)$. Pour cela, on peut utiliser l'algorithme *Forward-Backward* qui consiste à définir deux variables : *forward* et *backward*. Dans la suite les probabilités sont conditionnées par le modèle λ que l'on omet par soucis de lisibilité.

- Variable *forward* (ou filtrage). La propriété de Markov étant vérifiée, cette variable peut être calculée de manière incrémentale :

$$\begin{aligned} \alpha_t(i) &= p(\mathbf{y}_{1:t}, q_t = i) \\ &= p(\mathbf{y}_t|q_t = i) \sum_{j=1}^{N} p(q_t = i|q_{t-1} = j)p(\mathbf{y}_{1:t-1}, q_{t-1} = j) \\ &= p(\mathbf{y}_t|q_t = i) \sum_{j=1}^{N} p(q_t = i|q_{t-1} = j)\alpha_{t-1}(j) \\ &= b_i(\mathbf{y}_t) \sum_{j=1}^{N} a_{ji}\alpha_{t-1}(j) \end{aligned} \tag{A.3}$$

- Variable *backward* (ou lissage)

$$\begin{aligned} \beta_t(i) &= p(\mathbf{y}_{t+1:T}, q_t = i) \\ &= \sum_{j=1}^{N} p(\mathbf{y}_{t+1}|q_{t+1} = j)p(q_{t+1} = j|q_t = i)p(\mathbf{y}_{t+2:T}, q_{t+1} = j) \\ &= \sum_{j=1}^{N} p(\mathbf{y}_{t+1}|q_{t+1} = j)p(q_{t+1} = j|q_t = i)\beta_{t+1}(j) \\ &= \sum_{j=1}^{N} b_j(\mathbf{y}_{t+1})a_{ij}\beta_{t+1}(j) \end{aligned} \tag{A.4}$$

Algorithme de Viterbi

Etant donnée une séquence d'obervations $\mathbf{y}_{1:T}$, et un HMM $\lambda = (A, B, \pi)$, on aimerait avoir la séquence d'états $q_{1:T}$ optimale dans le sens où elle maximise la vraissemblance $p(\mathbf{y}_{1:T}, q_{1:T})$. Pour cela on utilise l'algorithme de Viterbi qui définit une variable $\delta_t(i)$ telle que :

$$\begin{aligned} \delta_t(i) &= \max_{q_{1:t-1}} p(q_1, \ldots, q_{t-1}, q_t = i, \mathbf{y}_{1:t}) \\ &= \max_{q_{1:t-1}} \left[p(\mathbf{y}_t | q_t = i) \sum_{q_{t-1}} p(q_t = i | q_{t-1}) p(\mathbf{y}_{1:t-1}, q_{t-1}) \right] \\ &= p(\mathbf{y}_t | q_t = i) \max_{q_{t-1}} \left[p(q_t = i | q_{t-1}) p(\mathbf{y}_{1:t-1}, q_{t-1}) \right] \\ &= p(\mathbf{y}_t | q_t = i) \max_j \left[p(q_t = i | q_{t-1} = j) \delta_{t-1}(j) \right] \end{aligned} \tag{A.5}$$

L'algorithme nécessite une passe en arrière pour récupérer les index des états qui optimisent le chemin.

A.3 Structure temporelle à plusieurs niveaux

Dans cette partie, nous nous focalisons sur deux modèles qui nous intéressent particulièrement : le modèle segmental et le modèle hiérarchique.

A.3.1 Modèles segmentaux

Les modèles de Markov à états cachés présentés précédemment ont trois limitations auxquelles répondent les modèles segmentaux. Premièrement, la modélisation de durée est faible car la distribution n'est pas explicite dans le modèle. Deuxièment, les observations sont indépendantes sachant l'état émettant l'observation. Finalement, les caractéristiques extraites d'une séquence d'obervations sont limitées par le fait de considérer les obervations une à une (*frame-based*) (Ostendorf et al., 1996). Ainsi les modèles segmentaux sont un cas particulier des modèles semi-Markoviens c'est à dire, qu'à la différence des modèles Markoviens, chaque état a une durée variable qui est le nombre d'observations émis tout en restant dans ce même état (Yu, 2010).

Description

Un modèle segmental à M états est décrit comme un HMM par une matrice A des probabilités de transition entre états, une matrice B des probabilités d'émission, une probabilité initiale Π auxquelles s'ajoutent une distribution sur les durées pour chaque état et un modèle de segment pour chaque état. Par exemple, la distribution sur les durées peut être uniforme (Artières et al., 2007) ou encore suivre une loi de Poisson (Kim and Smyth, 2006). De même, le modèle de segment pour chaque état peut être appris (Ostendorf et al., 1996) ; ou décrit de manière paramétrique, les paramètres étant appris sur une base d'entraînement (Artières et al., 2007; Kim and Smyth, 2006) ; ou encore donné a priori (Bloit et al., 2010).

Dans cette section, nous présentons le modèle segmental tel qu'il est présenté dans (Ostendorf et al., 1996). Ce modèle se base sur un vocabulaire de formes pour décrire une séquence d'observations. Ces formes peuvent être dilatées dans le temps exclusivement de manière uniforme. Formellement, si on note $\mathbf{y}_{1:T}$ une séquence d'observations, une sous-séquence de $\mathbf{y}$ allant de t_1 à t_2 (avec $t_1 < t_2$) sera notée $\mathbf{y}_{t_1:t_2}$ et sa longueur sera notée $l = t_2 - t_1 + 1$. Le modèle segmental permet de représenter la séquence $\mathbf{y}$ en sous-séquences concaténées :

$$\mathbf{y} = \left[\mathbf{y}_{1:l_1} \quad \mathbf{y}_{l_1+1:l_1+l_2} \quad \cdots \quad \mathbf{y}_{\sum_i l_i:T} \right]$$

Chaque segment est de longueur l_i et on a $\sum_{i=1}^{\tau} l_i = T$ où τ est le nombre de total de segments utilisés pour représenter $\mathbf{y}$. Ainsi, le modèle segmental a deux caractéristiques importantes :

1. Segmentation. Le processus des durées dans chaque état renseigne sur le début et la fin des segments, segmentant la séquence d'observations.
2. Régression. Pour un état s de durée l les observations $y_{1:l}$ sont décrites par un modèle de régression donné, ce qui permet l'accès à des segments interprétables (plus haut niveau que l'interprétation signal).

La probabilité qu'un état émette la séquence d'observations $\mathbf{y}_{t_1:t_2}$ peut alors s'écrire de la manière suivante :

$$p(\mathbf{y}_{t_1:t_2}, l_t = l|q_t = j) = p(\mathbf{y}_{t_1:t_2}|l_t = l, q_t = j)p(l_t = l|q_t = j)$$

Où $p(l_t = l|q_t = j)$ suit la distribution sur les durées définie a priori et $p(\mathbf{y}_{t_1:t_2}|l_t = l, q_t = j)$ est la probabilité d'observer $\mathbf{y}_{t_1:t_2}$ sachant le modèle de régression correspondant à l'état j et dont la durée est l.

Inférence et apprentissage

Les algorithmes présentés dans la section A.2, à savoir l'algorithme *forward-backward* et l'algorithme de Viterbi peuvent être utilisés dans le cas segmental. Seulement, dans le cas segmental s'ajoute la variable des durées. L'inférence est donc faite sur les états cachés (les segments) et les variables de durée. Pour les cas envisagés, nous nous intéressons essentiellement au problème de trouver la séquence optimale d'états avec leurs durées, c'est à dire au problème de segmentation. De ce fait, nous ne présentons ici que l'algorithme de Viterbi (les variables *forward* et *backward* peuvent aisément être déduites à l'aide de la règle de Bayes pour les probabilités conditionnelles).

Comme pour le cas des HMMs, dans l'algorithme de Viterbi il faut exprimer la variable $\delta_t(i)$ utilisée dans la première passe de l'algorithme (la seconde passe étant seulement une lecture de tableau). Cette variable s'écrit de la manière suivante :

$$\begin{aligned} \delta_t(i) &= \max_{n, l_{1:n}, q_{1:n-1}} p(\mathbf{y}_{1:t}, l_{1:n}, q_{1:n-1}, q_n = i) \\ &= \max_{n,j} p(\mathbf{y}_{t'+1:t}|l_n, q_n)p(l_n|q_n)p(q_n|q_{n-1} = j)\delta_{t'}(j) \end{aligned} \tag{A.6}$$

Un modèle segmental a une structure à deux niveaux dans lesquels, les états sont discrets. Il s'agit donc d'inférer ces deux niveaux temporels imbriqués.

A.3.2 Modèles hiérarchiques

Le modèle de Markov à états cachés hiérarchique (Fine et al., 1998) est un modèle dans lequel chaque état émet un modèle autonome. Ceci permet une hiérarchie (temporelle) plus complexe, c'est à dire à plusieurs niveaux (potentiellement supérieur à 2). En revanche, contrairement au modèle segmental où les durées dans les états pouvaient être explicitement modélisées par une distribution de probabilité, dans le modèle hiérarchique, présenté ici, chaque chaîne à chaque niveau respecte la condition de Markov. À l'instar du modèle segmental, le modèle hiérarchique a été d'abord utilisé pour la reconnaissance de la parole pour la modélisation des hiérarchies temporelles du langage : des phonèmes aux mots, des mots à la phrase.

Description

Soit un modèle hiérarchique à D niveaux. Le niveau 1 est la racine et le niveau D comprend les états de production. Chaque état est un modèle, donc si on note q^d un état au niveau d, il génère un modèle noté λ^{q^d}. Ainsi un modèle hiérarchique est

$$\lambda = \left\{\lambda^{q^d}\right\}_{d=1\ldots D}$$

Seul le niveau D produit des observations, donc aura une probabilité d'émission. En revanche les autres niveaux ont une probabilité de transition vers des états au même niveau ainsi que des transitions verticales qui permettent d'entrer dans un sous-modèle via une distribution de probabilités initiales. On peut écrire :

$$\lambda = \left\{ \left\{ A^{q^d} \right\}_{d=1\ldots D}, \left\{ \Pi^{q^d} \right\}_{d=1\ldots D-1}, \left\{ B^{q^D} \right\} \right\}$$

La figure A.1 est un exemple arbitraire de topologie pour un modèle hiérarchique. Le modèle de la figure génère l'expression régulière définit par : $a^+|a+(xy)^+b|c(xy)^+d$. Dans cette figure, les flèches pleines représentent les transitions au sein d'un niveau donné d : A^{q^d}. Il y a donc des transitions à tous les niveaux. Les flèches en pointillés sont les transitions aux niveaux inférieurs. L'état dans lequel commence le niveau inférieur est donné par la distribution Π^{q^d}. Les flèches avec tirets sont les émissions de symboles provenant des états de production. Par construction, tous les symboles sont émis au niveau le plus bas, c'est à dire D. Enfin les cercles doubles sont les états terminaux : ces états n'émettent pas mais indiquent la fin d'un sous-modèle à l'état parent afin de faire une transition au niveau supérieur.

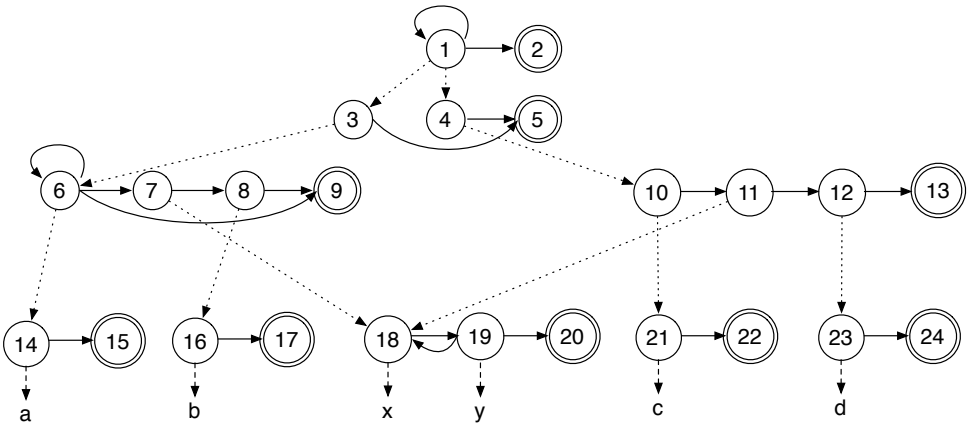

FIGURE A.1 – Exemple de topologie pour un modèle hiérarchique

Inférence

Il est possible de définir les variables *forward* et *backward* pour l'inférence de même que l'algorithme de Viterbi peut être adapté. Nous donnons la forme de la variable *forward* α et nous laissons le lecteur intéressé par une formulation générale des différentes variables se référer à l'article de Fine et al. (Fine et al., 1998). La variable α est définit par :

$$\alpha(t, t+k, q_i^d, q^{d-1}) = p(\mathbf{y}_t, \ldots, \mathbf{y}_{t+k}, q_i^d \text{ finit à } t+k | q^{d-1} \text{ commence à } t+k)$$

En d'autres termes, étant dans l'état de plus haut niveau q^{d-1} à l'instant t, on parcourt les états inférieurs q_1^d, q_2^d jusque q_i^d (on rappelle que eux mêmes peuvent être des modèles, donc peuvent générer aussi des séquences, c'est pourquoi $i \neq k$). À l'état q_i^d nous entrons dans un état terminal est retournons au niveau supérieur $d-1$. La variable $\alpha(t, t+k, q_i^d, q^{d-1})$ correspond à la probabilité de faire ce parcours au vu des observations $(\mathbf{y}_t, \ldots, \mathbf{y}_{t+k})$. Ainsi l'inférence sur la séquence globale d'observations réside en trois parcours imbriqués. Le premier parcours allant de 1 à T est le début de la séquence (la première variable de α). Ensuite, un deuxième parcours est effectué sur les valeurs de k, c'est à dire la taille de la sous-séquence (la deuxième variable de α). Enfin un troisième parcours s'effectue sur les instant t' compris entre t et $t+k$ utilisés pour calculer les probabilités à propager. L'algorithme a une complexité cubique

en temps et linéaire en espace (i.e. la dimension de l'espace d'états). Cette complexité peut être réduite à linéaire en temps et quadratique en nombre d'états et de niveaux de hiérarchie (Murphy and Paskin, 2001).

A.4 Modèles continus

Les modèles de Markov permettent de "résumer" les régularités temporelles d'un signal de manière compacte et adaptée pour la reconnaissance. Seulement, ils sont basés sur l'hypothèse que tout le passé d'un signal est contenu dans K entités formant l'espace d'états et suivant une distribution multinomiale. Cette hypothèse permet l'utilisation d'une méthode d'apprentissage très efficace (l'algorithme EM), ce qui a contribué au succès de tels modèles. Ainsi, modéliser N bits d'information sur le passé nécessite 2^N états. Les modèles à états continus ont une meilleure capacité de représentation mais seulement deviennent vite non-résoluble dans le cas où le système dynamique devient non-linéaire. Dans cette partie nous présentons le cas simple des systèmes dynamiques linéaires pour lesquels il existe un algorithme efficace d'inférence. Ensuite nous présentons un cas particulier d'algorithme d'inférence pour des systèmes non-linéaires.

A.4.1 Systèmes linéaires dynamiques (SLDs)

Les systèmes linéaires dynamiques se définissent par le système d'équations linéaires suivant :

$$\begin{aligned} \mathbf{x}_t &= \mathbf{A}\mathbf{x}_{t-1} + \mathbf{v}_t \\ \mathbf{y}_t &= \mathbf{B}\mathbf{x}_t + \mathbf{w}_t \end{aligned} \tag{A.7}$$

Où $\mathbf{A}$ est la matrice du système dynamique linéaire sous-jacent et $\mathbf{B}$ la matrice d'emission. Les bruits additifs $\mathbf{v}_t$ et $\mathbf{w}_t$ sont gaussiens, de moyenne nulle et de matrices de covariance respectives $\mathbf{Q}, \mathbf{R}$. Ils représentent le bruit de transition ($\mathbf{v}_t$) et le bruit de mesure ($\mathbf{w}_t$). Les HMMs peuvent aussi être présentés sous forme d'un système dynamique linéaire (cf. (Roweis and Ghahramani, 1999) pour plus de détails) avec la différence principale que les variables latentes dans le système (A.7) sont continues.

La probabilité de transition entre états cachés (passage de $\mathbf{x}_{t-1}$ à $\mathbf{x}_t$) ne suit plus une loi multinomiale mais gaussienne :

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{A}\mathbf{x}_{t-1}; \mathbf{Q}) \tag{A.8}$$

Ainsi, le système dynamique sous-jacent impose que l'état suivant soit "proche" de l'état précédent (relativement à une certaine variance) et le "déplacement" vers l'état suivant se fait par $\mathbf{A}$. À l'instar des HMMs, la probabilité d'émission suit une loi gaussienne.

$$p(\mathbf{y}_t|\mathbf{x}_t) = \mathcal{N}(\mathbf{B}\mathbf{x}_t; \mathbf{R}) \tag{A.9}$$

L'inférence dans des modèles à états cachés continus n'est résoluble que dans le cas où les relations sont linéaires.

Inférence

Le cas linéaire permet l'écriture explicite de la probabilité jointe. En effet, le processus latent est un processus Markovien d'ordre 1 et les observations sont indépendantes ce qui conduit à l'égalité suivante (similaire pour les HMMs) :

$$p(\mathbf{y}_{1:T}, \mathbf{x}_{1:T}) = p(\mathbf{x}_1) \prod_{t=2}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}) \prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{x}_t) \tag{A.10}$$

En reportant les équations (A.8) et (A.9) on peut écrire la probabilité jointe :

$$\begin{aligned} p(\mathbf{y}_{1:T}, \mathbf{x}_{1:T}) &= C \times |\mathbf{V}_1|^{-1/2} \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} [\mathbf{x}_t - \pi_1]' \mathbf{V}_1^{-1} [\mathbf{x}_t - \pi_1] \right\} \times \\ & |\mathbf{Q}|^{-T/2} \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} [\mathbf{x}_t - \mathbf{A}\mathbf{x}_{t-1}]' \mathbf{Q}^{-1} [\mathbf{x}_t - \mathbf{A}\mathbf{x}_{t-1}] \right\} \times \\ & |\mathbf{R}|^{-T/2} \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} [\mathbf{y}_t - \mathbf{B}\mathbf{x}_t]' \mathbf{R}^{-1} [\mathbf{y}_t - \mathbf{B}\mathbf{x}_t] \right\} \end{aligned} \tag{A.11}$$

Où C est une constante et où on assume que la probabilité initiale $p(\mathbf{x}_1)$ suit une loi gaussienne de moyenne π_1 et de matrice de covariance V_1. Ainsi, le problème de trouver $\mathbf{x}_t$ le plus vraisemblable revient à trouver les zéros du gradient du logarithme de la probabilité définie par (A.11). Comme le logarithme transforme les produits de (A.11) en somme, il est possible de trouver la valeur de $\mathbf{x}_t$ optimal au sens qui minimise l'erreur des moindres carré. C'est l'algorithme du filtrage de Kalman.

Limitations

Dans le cadre de l'analyse du geste, les systèmes dynamiques linéaires ne peuvent pas modéliser les effets non-linéaires dus aux articulations, au squelette, aux divers frottements et forces, etc... De manière générale, dans le nature, les systèmes dynamiques sont non-linéaires. L'approximation linéaire est élégante et efficace dans certains cas. De la même manière que pour les HMMs, les SDLs peuvent être appris sur des données (adaptation de l'algorithme EM avec le filtrage de Kalman), ce qui leur donne une importance toute particulière. Le cas non-linéaire a attisé beaucoup de réflexions et a donné lieu à beaucoup de publications. Dans le cadre de cette thèse nous nous focalisons sur deux aspects utiles pour des applications gestuelles : l'approximation du cas non-linéaire par une séquence de systèmes linéaires ; et la résolution d'un système non-linéaire simple par une méthode de Monte Carlo séquentielle (ou *filtres particulaires*).

Une autre limitation des filtres de Kalman est l'hypothèse de distributions gaussiennes. Ainsi, dans un contexte de suivi, il sera impossible de suivre plusieurs objets à la fois. Ceci est très limitant notamment dans le cadre de la reconnaissance de gestes où il s'agit de reconnaître un geste d'entrée comme étant un template parmi plusieurs. La méthode par filtrage particulaire (qui sera détaillée dans la section A.4) a l'avantage de pouvoir estimer une distribution non-gaussienne et ainsi de pouvoir suivre plusieurs gestes en propageant la vraisemblance de chacun.

A.4.2 Approcher la non-linéarité : commuter les SLDs

Etant donnée l'existence d'un algorithme d'inférence optimal et d'une méthode d'apprentissage adaptée pour le cas linéaire, on pourrait être tenté de faire l'approximation du cas non-linéaire par un cas linéaire par morceaux, ce qu'on appelle dans la littérature *Switching Linear Dynamic Systems* (SLDS).

Description

En reprenant les mêmes notations que dans la section précédente A.4, le système d'équations vérifié pour un SLDS est le suivant :

$$\begin{aligned} \mathbf{x}_t &= \mathbf{A}(s_t)\mathbf{x}_{t-1} + \mathbf{v}_t(s_t) \\ \mathbf{y}_t &= \mathbf{B}(s_t)\mathbf{x}_t + \mathbf{w}_t(s_t) \end{aligned} \tag{A.12}$$

Où s_t est une variable discrète relative à un système linéaire (index du système), $\mathbf{A}(s_t)$ est la matrice du système dynamique linéaire pour le système s_t à l'instant t, $\mathbf{B}(s_t)$ sa matrice d'émission et $\mathbf{v}_t(s_t), \mathbf{w}_t(s_t)$ leurs bruits additifs Gaussiens de moyenne nulle et de covariances respectives $\mathbf{A}(s_t)$ et $\mathbf{B}(s_t)$. Au LDS précédemment défini dans la section A.4, on ajoute un niveau supérieur d'états discrets conditionnant les états continus du système dynamique : chaque valeur prise par un état s_t correspond à un SDL défini par les matrices $\mathbf{A}, \mathbf{B}, \mathbf{Q}, \mathbf{R}$.

A.4.3 Systèmes non-linéaires : cas des filtres particulaires

Un système non-linéaire est régi par les équations suivantes :

$$\begin{aligned} \mathbf{x}_t &= f(\mathbf{x}_{t-1}, \mathbf{v}_t) \\ \mathbf{z}_t &= h(\mathbf{x}_t, \mathbf{w}_t) \end{aligned} \tag{A.13}$$

Où les fonctions f et h sont non-linéaires par rapport aux variables d'états et/ou au bruit. Les fonctions $f(\mathbf{x}_{t-1}, \mathbf{v}_t)$ et $h(\mathbf{x}_t, \mathbf{w}_t)$ sont dans le cas général non Gaussiennes. Contrairement au cas linéaire, il n'y a pas d'algorithme d'inférence exact et optimal. Les algorithmes considérés sont alors appelés *sub-optimaux*. Dans le panel des algorithmes d'inférence sub-optimaux, nous nous intéressons particulièrement au cas des filtres particulaires (Arulampalam et al., 2002).

Filtres particulaires

Les filtres particulaires sont des méthodes de Monte Carlo séquentielles et permettent d'estimer récursivement la distribution courante sur les états utilisant la méthode de d'échantillonnage séquentielle d'importance (ou préférentielle) (Geweke, 1989; Doucet et al., 2000). Une mesure aléatoire $\{\mathbf{x}_k^i, w_k^i\}_{i=1}^{N_s}$ est utilisée pour caractériser la distribution de probabilité a posteriori avec un ensemble de points sur l'ensemble d'états et des poids associés. Ainsi, la vraie distribution sur les états peut être estimée avec une série de fonctions Deltas de Dirac :

$$p(\mathbf{x}_k|\mathbf{x}_{1:k}, \mathbf{x}_0) \approx \sum_{i=1}^{N_s} w_k^i \delta(\mathbf{x}_k - \mathbf{x}_k^i) \tag{A.14}$$

Le terme $\mathbf{x}_0$ représente la distribution a priori (c'est à dire l'état initial), et la distribution a posteriori est actualisée à chaque pas de temps. Il s'ensuit une étape de rééchantillonnage (optionnel) qui permet de résoudre le problème de *dégénerescence*, commun aux approches particulaires, qui consiste au fait qu'en très peu d'itérations il ne reste que très peu de particules actives ce qui peut figer l'algorithme sur une mauvaise solution (ceci est discuté en détail dans (Arulampalam et al., 2002; Douc and Cappé, 2005)).

Annexe B

Modèles de suivi de geste

Les modèles présentés ici permettent le suivi temps réel d'un geste. Pour ce faire, un ensemble de gestes de référence est défini et à chaque nouvelle observation d'un geste capté, l'algorithme renvoie le geste de référence le plus probable de même que la position dans cette référence. Dans les deux modèles présentés ci-dessous, les gestes de référence sont définis à partir d'un seul exemple de chaque geste. Le premier modèle fait le suivi des trajectoires décrites par le geste en se basant sur un système dynamique linéaire et une inférence par filtrage particulaire B.1. Le deuxième modèle fait aussi le suivi en se basant sur les trajectoires de descripteurs gestuels, mais il se base sur un HMM et une inférence causale B.2.

B.1 Modèle basé sur l'algorithme CONDENSATION

Ici nous reportons le modèle présenté par Black et al. dans (Black and Jepson, 1998b). Le but est de reconnaître en temps réel le geste effectué sachant un ensemble de N gestes de référence. Ce modèle a inspiré celui présenté dans le chapitre 8. C'est un modèle à états défini par :

$$\mathbf{x}_t = \begin{pmatrix} \mu \\ \phi \\ \alpha \\ \rho \end{pmatrix}$$

Où, au temps t, μ est l'index d'une référence, ϕ la position (ou *phase*) dans la référence, α est un coefficient d'amplitude (facteur multiplicatif) et ρ la vitesse d'exécution. Le but est de trouver la valeur de l'état $\mathbf{x}_t$ qui a donné lieu aux observations $Z_t = (\mathbf{z}_t, \mathbf{z}_{t-1}, \dots)$. Si $Z_{t,i} = (z_{t,i}, z_{(t-1),i}, \dots)$ dénote la i-ème dimension de la séquence d'observations, la probabilité d'observation peut s'écrire :

$$p(Z_t|\mathbf{x}_t) = \prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-\sum_{j=0}^{w-1}(z_{(t-j),i} - \alpha m^{(\mu)}_{(\phi-\rho j),i})^2}{2\sigma^2(w-1)}\right]$$

Où w est la taille d'une fenêtre sur laquelle est effectuée la correspondance entre segment de geste d'entrée et segment de geste de référence. Aussi, $\alpha m^{(\mu)}_{(\phi-\rho j),i}$ est la i-ème dimension de la valeur de l'échantillon à l'index $(\phi - \rho j)$ pour le geste de référence μ.

Afin d'estimer les valeurs des variables d'états, l'auteur propose l'utilisation de l'algorithme CONDENSATION basé sur le filtrage particulaire. Tout d'abord, les variables d'états sont échantillonnées uniformément dans des intervalles pré-définis :

$$\begin{aligned} \mu &\in [0, \mu_{\max}] \\ \phi &\in [0, 1] \\ \alpha &\in [\alpha_{\min}, \alpha_{\max}] \\ \rho &\in [\rho_{\min}, \rho_{\max}] \end{aligned}$$

Notons N_s le nombre de particules distribuées initialement. On note la particule n par $\mathbf{x}_t^n = (\mu^n, \phi^n, \alpha^n, \rho^n)$. Chaque particule est pondérée par :

$$w_n = \frac{p(\mathbf{z}_t|\mathbf{x}_t^n)}{\sum_{i=1}^{N_s} p(\mathbf{z}_t|\mathbf{x}_t^i)}$$

L'algorithme suit les trois étapes suivantes

1. **Sélection**. L'estimation de la distribution $\mathbf{w} = (w_1, \ldots, w_n)$ sur les états à $t-1$ est utilisée pour sélectionner l'état à $t-1$.
2. **Prédiction**. À partir de la particule $\mathbf{x}_t^n$, on propage la particule de la manière suivante :

$$\begin{aligned} \mu_t &= \mu_{t-1} \\ \phi_t &= \phi_{t-1} + \rho_{t-1} + \mathcal{N}(\sigma_\phi) \\ \alpha_t &= \alpha_{t-1} + \mathcal{N}(\sigma_\alpha) \\ \rho_t &= \rho_{t-1} + \mathcal{N}(\sigma_\rho) \end{aligned}$$

3. **Mise à jour**. Étant donné la particule $\mathbf{x}_t^n$ on estime la probabilité $p(\mathbf{z}_t|\mathbf{x}_t^n)$ et a fortiori w_n à l'instant t.

B.2 Modèle hybride basé sur HMM

Nous présentons ici brièvement les principales caractéristiques du modèle hyrbide HMM-DTW présenté dans l'article (Bevilacqua et al., 2010). Le modèle possède une structure de HMM. L'apprentissage se base sur un seul exemple de référence gestuelle (ou *template*). Chaque échantillon de la référence définit un état du HMM comme indiqué sur la figure B.1.

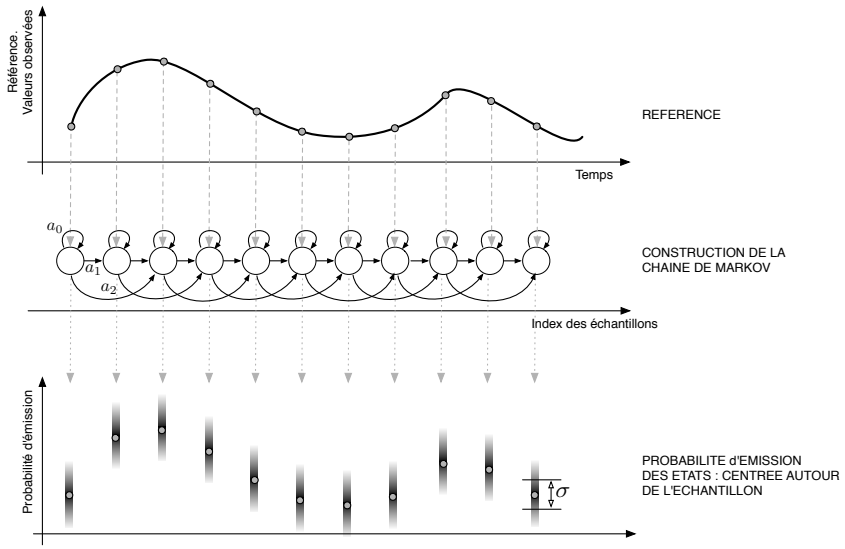

FIGURE B.1 – Apprentissage de modèle hybride HMM-DTW

Les états sont les index des échantillons dans la référence. La probabilité d'émission est une gaussienne centrée autour de l'échantillon correspondant à l'état émettant avec une variance constante σ^2 :

$$p(\mathbf{o}|q=i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\|\mathbf{o} - \boldsymbol{\mu}_i\|^2}{2\sigma^2}\right]$$

Où $\boldsymbol{\mu}_i$ est la valeur du i-ème échantillon dans la référence. Elle est multi-dimensionnelle, $\mathbf{o}$ est une observation (mutli-dimensionnelle), et σ est constante au cours du temps.

L'inférence est causale et utilise l'algorithme *forward* (cf. section A.2). Ainsi un geste d'entrée sera aligné sur la référence de manière causale (cf. figure B.2) et à chaque pas de temps une valeur de vraisemblance est calculée donnant la probabilité que le geste d'entrée soit reconnu comme étant le geste de référence.

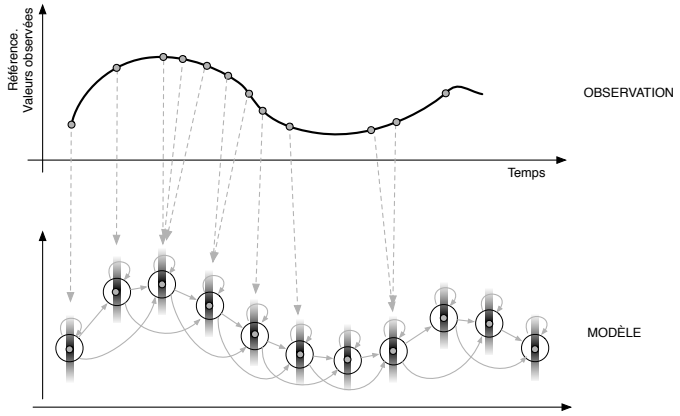

FIGURE B.2 – Décodage de modèle hybride HMM-DTW

En pratique plusieurs gestes de référence, différents, sont enregistrés, créant chacun une structure hybride HMM–DTW. Le décodage calcule la probabilité d'être dans chacun des gestes sachant le modèle (template). La normalisation sur tous les gestes de référence retourne la probabilité d'être dans un geste sachant tous ceux dans la base.

Annexe C

Towards Cross-Modal Gesture–Sound Analysis

B. Caramiaux[1], **F. Bevilacqua**[1], **N. Schnell**[1]

[1] UMR IRCAM-CNRS, Paris, France

Abstract : This article reports on the exploration of a method based on canonical correlation analysis (CCA) for the analysis of the relationship between gesture and sound in the context of music performance and listening. This method is a first step in the design of an analysis tool for gesture-sound relationships. In this exploration we used motion capture data recorded from subjects performing free hand movements while listening to short sound examples. We assume that even though the relationship between gesture and sound might be more complex, at least part of it can be revealed and quantified by linear multivariate regression applied to the motion capture data and audio descriptors extracted from the sound examples. After outlining the theoretical background, the article shows how the method allows for pertinent reasoning about the relationship between gesture and sound by analysing the data sets recorded from multiple and individual subjects.

Keywords : Gesture analysis, Gesture-Sound Relationship, Sound Perception, Canonical Correlation Analysis

C.1 Introduction

Recently, there has been an increasing interest in the multimodal analysis of the expression of emotion as well as expressivity in music. Several works reveal that motor expression components like body gestures are always accompanying other modalities (Scherer, Klaus R. and Ellgring, Heiner, 2007). For instance, human face-to-face communication often combines speech with non-verbal modalities like gestures. In this context, multimodal analysis reveals co-expressive elements that play an important role for the communication of emotions. In a similar way, we'd like to explore the relationship between gestures and sound in the context of music performance and listening.

We are particularly interested in the relationship between sound and the movements of an individual or a group in a listening situation as well as the movements of a music performer that are related primarily to the production of sound, in addition to the musical intention and the expression of emotion (Leman, 2007).

In our current project, we develop a set of methods for the analysis of the relationship between different aspects of gestures and sound. We would like to be able to apply these methods to a variety of contexts, covering the performance of traditional and electronic (virtual) instruments as well as different music listening scenarios. The goal of this work reaches the creation of tools for the study of gesture in musical expression and perception. In a greater context,

these tools contribute to the development of novel paradigms within the intersection between music performance and music listening technologies.

In this paper, we present a new approach to the quantitative analysis of the relationship between gesture and sound. The article is organized as follows. We first present a review of related works. Then we introduce in section C.3 the multivariate analysis method called canonical correlation analysis. In section C.4 we present the experimental context including our data capture methods and we show results on feature selection and correlation analysis of collected data. We discuss these results in C.5. Finally, we conclude and give the implications on further works in section C.6.

C.2 Related Work

The concept of embodied cognition has been adopted by a wide community of researchers. In this context, the relationship between gesture and sound has come into interest to interdisciplinary research on human communication and expression.

Some recent researches in neurosciences ((Kohler et al., 2002), (Metzinger and Gallese, 2003)) and others in perception ((Varela et al., 1991), (Berthoz, Alain, 1997), (Noë, 2005)) have shown that action plays a predominant role in perception insisting on the inherently multimodal nature of perception. In (Kita and Asli, 2003), (Kopp and Wachsmuth, 2004), (Bergmann and Kopp, 2007) the authors show that gesture and speech are to some extent complementary co-expressive elements in human communication.

Research in the domain of music and dance has studied the embodiment of emotion and expressivity in movement and gesture. Leman (Leman, 2007) has widely explored various aspects of music embodiment based on the correlation between physical measurements and corporeal articulations in respect to musical intention. Camurri et al. in (Camurri et al., 2003) show that emotion can be recognized in a dancing movement following dynamic features such as *quantity of motion* extracted from motion capture data. Dahl et al. in (Dahl and Friberg, 2003) show to what extent emotional intentions can be conveyed through musicians' body movements. Moreover, Nusseck and Wanderley in (Nusseck and Wanderley, 2009) show that music experience is multimodal and is less depend on the players' particular body movements than the player's overall motion characteristics.

Several recent works have studied gestures performed while listening to music revealing how an individual perceives and imagines sound and sound production as well as music and music performance. In (Godøy et al., 2005), (Jensenius, A. R., 2007) and (Haga, Egil, 2008) the authors explore the relationship between gesture and musical sound using qualitative analysis of the gestural imitation of musical instrument performance (*air-instruments*) as well as free dance and drawing movements associated with sounds (*sound-tracing*). For instance, (Godøy et al., 2005) shows that air-instrument performance can reflect how people perceive and imagine music highly depending on their musical skills.

On the other hand, only a few works have taken a quantitative approach and are mostly focussing on the synchronisation between gestures and music. In (Large, 2000), Large proposes a pattern-forming dynamical system modelling the perception of beat and meter that allows for studying the synchronisation and rhythmic correspondence of movement and music. Experiments in which subjects were asked to tap along with the musical tempo have revealed other pertinent characteristics of the temporal relationship between movement and music ((Repp, Bruno Hermann, 2006), (Luck and Toiviainen, 2006), (Styns, Frederik et al., 2007)) such as negative asynchrony, variability, and rate limits. In (Luck and Toiviainen, 2006), the authors give a quantitative analysis of the ensemble musicians' synchronization with the conductor's gestures. The authors have used cross-correlation analysis on motion capture data and beat patterns extracted from the audio signal to study the correspondence between the conduc-

tor's gestures and the musical performance of the ensemble. Lastly, Styns ((Styns, Frederik et al., 2007)) has studied how music influences the way humans walk analysing the correspondence between kinematic features of walking movements and beat patterns including the comparison of movement speed and walking tempo in addition to the analysis of rhythmic synchronicity. He shows that walking to music can be modelled as a resonance phenomenon (with resonance frequency at 2Hz).

In our work we attempt to introduce a method for the quantitative multimodal analysis of movement and sound that allows for the selection and analysis of continuous perceptively pertinent features and the exploration of their relationship. It focuses on free body movements performed while listening to recorded sounds. The mathematical approach is a general multivariate analysis method that has not been used yet in gesture-sound analysis, but that has given promising results in the analysis of multimedia data and information retrieval ((Kidron et al., 2005)).

C.3 Canonical Correlation Analysis : an Overview

Proposed by Hotelling in (Hotelling, 1936), Canonical Correlation Analysis (CCA) can be seen as the problem of measuring the linear relationship between two sets of variables. Indeed, it finds basis vectors for two sets of variables such that the correlations between the projections of the variables onto these basis vectors are mutually maximised. Thus, respective projected variables are a new representation of the variables in directions where variance and co-variance are the most explained.

Let us introduce some notations : bold type will be used for matrices (**X**, **Y**, etc...) and vectors (**u**, **v**, etc...). The matrix transpose of **X** will be written as $\mathbf{X}^T$. Finally, an observation of a random variable **v** will be written as v_i at time i.

Consider two matrices **X** and **Y** where the rows (resp. columns) are the observations (resp. variables). **X**, **Y** must have the same number of observations, denoted m, but can have different numbers of variables, denoted n_x resp. n_y. Then, CCA has to find two projection matrices, **A** and **B**, such as

$$\max_{\mathbf{A},\mathbf{B}} \left[\text{corr}\left(\mathbf{XA}, \mathbf{YB}\right)\right] \tag{C.1}$$

Here *corr* denotes the correlation operator between two matrices. Usually, the correlation matrix of a matrix **M** of dimension $m \times n$ is the correlation matrix of n random variables (the matrix columns $\mathbf{m}_1, ..., \mathbf{m}_n$) and is defined as a $n \times n$ matrix whose (i, j) entry is corr $(\mathbf{m}_i, \mathbf{m}_j)$. The correlation between two matrices is the correlation between the respective indexed columns. Therefore **XA** and **YB** must have the same number of variables. **A** and **B** are $n_x \times \min(n_x, n_y)$ and $n_y \times \min(n_x, n_y)$ matrices. Let h be one arbitrary variable index in **XA** (as in **YB**), equation (C.1) can be written as finding $\mathbf{a}_h$ and $\mathbf{b}_h$, $\forall h = 1... \min(n_x, n_y)$, that maximize :

$$\text{corr}\left(\mathbf{Xa}_h, \mathbf{Yb}_h\right) \tag{C.2}$$

We remind the reader that the correlation coefficient between two random variables is computed as the quotient between the covariance of these two random variables and the square root of the product of their variance. Let denote $\mathbf{C}(\mathbf{X}, \mathbf{Y})$ the covariance matrix. It is a positive semi-definite matrix and can be written as

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

Thus we can formulate the problem from equation (C.2) using the previous notations : find

$\mathbf{A}, \mathbf{B}$ such that the following quotient is maximized

$$\text{corr}\left(\mathbf{X}\mathbf{a}_h, \mathbf{Y}\mathbf{b}_h\right) = \frac{\text{cov}\left(\mathbf{X}\mathbf{a}_h, \mathbf{Y}\mathbf{b}_h\right)}{\sqrt{\text{var}\left(\mathbf{X}\mathbf{a}_h\right)\text{var}\left(\mathbf{Y}\mathbf{b}_h\right)}} = \frac{\mathbf{a}_h^T \mathbf{C}_{xy} \mathbf{b}_h}{\sqrt{\mathbf{a}_h^T \mathbf{C}_{xx} \mathbf{a}_h \mathbf{b}_h^T \mathbf{C}_{yy} \mathbf{b}_h}} \tag{C.3}$$

One can show that equation (C.3) leads to a generalized eigenproblem of the form (see (Hair, Joseph F. et al., 2009)) :

$$\mathbf{M}_1\mathbf{v} = \lambda\mathbf{M}_2\mathbf{v}$$

Efficient methods can be implemented to find interesting projection matrices. The key terms for an understanding of CCA are : *canonical weights* (coefficients in **A** and **B**) ; *canonical variates* (projected variables, **XA** and **YB**) ; *canonical function* (relationship between two canonical variates whose strength is given by the canonical correlation).

Interpreting canonical correlation analysis involves examining the canonical functions to determine the relative importance of each of the original variables in the canonical relationships. Precise statistics have not yet been developed to interpret canonical analysis, but several methods exist and we have to rely on these measures. The widely used interpretation methods are : canonical weights, canonical loadings and canonical cross-loadings. In this paper we use the second one because of its efficiency and simplicity. Canonical Loadings measure the simple correlation between variables in each set and its corresponding canonical variates, i.e. the variance that variables share with their canonical variates. Canonical Loadings are computed as :

$$\begin{aligned} \textit{Gesture loadings} &: \mathbf{L}_G = \text{corr}\left(\mathbf{X}, \mathbf{U}\right) \\ \textit{Sound loadings} &: \mathbf{L}_S = \text{corr}\left(\mathbf{Y}, \mathbf{V}\right) \end{aligned}$$

C.4 Cross-Modal Analysis

We applied the method based on CCA to some examples of data collected in an experiment with subjects performing free body movements while listening to sound recordings imagining themselves producing the sound. Given the setup of the experiment, gesture and sound can be assumed as highly correlated without knowing their exact relationship that may be related to the subjects' sound perception, their intention of musical control, and their musical and motor skills. In this sense, the collected data sets have been a perfect context to explore the developed method and its capability to support reasoning about the relationship between gesture and sound.

C.4.1 Collected Data

The data has been collected in May 2008 in the University of Music in Graz. For the experiment 20 subjects were invited to perform gestures while listening to a sequence of 18 different recorded sound extracts of a duration between 2.05 and 37.53 seconds with a mean of 9.45 seconds. Most of the sound extracts were of short duration. Since the experience was explorative, the sound corpus included a wide variety of sounds : environmental and musical of different styles (classical, rock, contemporary).

For each sound, a subject had to imagine a gesture that he or she performed three times after an arbitrary number of rehearsals. The gestures were performed with a small hand-held device that included markers for a camera-based motion capture system recording its position in Cartesian coordinates. A foot-pedal allowed to synchronise the beginning of the movement with the beginning of the playback of the sound extract in the rehearsal as well as for the recording of the final three performances.

C.4.2 Gesture Data

As input of the analysis method, a gesture is a multi-dimensional signal stream corresponding to a set of observations. The most basic kinematic features are the position coordinates x, y, z, velocity coordinates v_x, v_y, v_z and acceleration coordinates a_x, a_y, a_z derived from the motion capture data. These features give a basic and efficient representation of postures and body movements describing their geometry and dynamics. For instance, Rasamimanana in (Rasamimanana et al., 2006) shows that three types of bow strokes considered in the paper are efficiently characterized by the features $(a_{\min}, a_{\max})$. In order to abstract from absolute position and movement direction, we calculate vector norms for position, velocity, and acceleration. To also consider movement trajectories, we additionally represent the gestures in an adapted basis using Frenet-Serret formulas giving $curvature$ and $torsion$ in the coordinate system $(\mathbf{t}, \mathbf{n}, \mathbf{b})$. In the same coordinate system, we add $normal$ and $tangential\ accelerations$ denoted by acc_N and acc_T (that replace previous acceleration).

Finally, at the input of the method a gesture is represented by a finite sequence of observations of the following variables :

$$\{position, velocity, acc_N, acc_T, curvature, radius, torsion\}$$

The CCA here permits to select the most pertinent features used in further calculations eliminating non-significant parameters.

C.4.3 Sound Features

The perception of sound has been studied intensively since one century and it is now largely accepted that sounds can be described in terms of their pitch, loudness, subjective duration and "timbre". For our exploration, we extract a set of audio descriptors from the audio files used in the experiment that have been shown to be perceptively relevant (see (Peeters, 2004)). While we easily can rely on loudness and pitch the perceptual relevance of audio descriptors for timbre and its temporal evolution is less assured. Nevertheless, we have chosen to use *sharpness* corresponding to the perceptual equivalent to the spectral centroïd. Pitch has been discarded since in musical performance it generally requires high precision control associated to expert instrumental gestures (defined as *selection gestures* in (Cadoz and Wanderley, 2000)).

At the input of the method a sound is represented by a finite sequence of observations of the following variables :

$$\{loudness, sharpness\}$$

Their perceptual characteristic allows the easy interpretation of gesture-sound relationship analysis.

C.4.4 Results

For free body movements performed while listening to recorded sound extracts, we are interested in investigating how gesture can explain sound through sound features and how sound can highlight important gesture characteristics. Among the whole set of sounds we chose two : the sound of an ocean wave and a solo flute playing a single note with strong timbre modulation (extract from *Sequenza I* for flute (1958), by Luciano Berio). These two sounds appeared to be the most pertinent extracts given the selection of audio descriptors discussed in C.4.3. The set of two perceptual audio descriptors computed on each sound can be seen in figure C.1.

The wave sound is characterized by a spectral distribution similar to a white noise passing through a specific filter. It leads to a sharpness feature highly correlated with the loudness

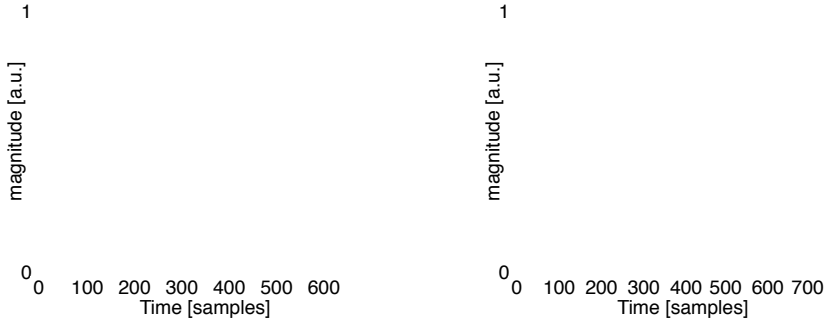

FIGURE C.1 – Loudness and Sharpness. On the left, feature values are plotted for the wave sound. The line corresponds to loudness, and the gray line sharpness. The same features for the flute timbre example are plotted on the right.

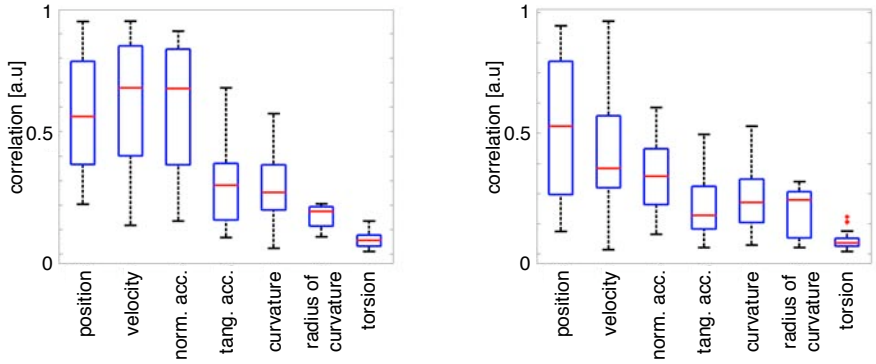

FIGURE C.2 – Relevant gesture parameters. Each parameter is analysed together with the audio descriptors using CCA for 42 gestures. Results for the wave sound are plotted on the left side while flute results can be seen on the right.

(correlation coefficient of 0.814). Since the flute example characteristic resides in a continuous transformation of its spectrum without significatively changing the fundamental frequency, its computed loudness and sharpness are less correlated (its correlation coefficient is -0.61).

First, gesture parameters considered as pertinent in the context cannot be chosen arbitrarily. Our analysis method can be applied to select a subset of pertinent gesture parameters using one gesture and many audio descriptors. In this way, the method operates as a multiple regression : the gesture parameter is predicted from audio descriptors. Each analysis returns one correlation coefficient corresponding to the canonical function strength between the current gesture parameter and the audio canonical component. 42 gestures are considered landing 42 canonical analysis iterations for each gesture parameter and each sound. Figure C.2 shows two box plots corresponding to this process as applied to the wave and flute sounds. Three principal features are emphasized : position (index 1), velocity (index 2), and normal acceleration (index 3). Since these features have the highest correlation means among those in the set of gesture parameters, they constitute a set of pertinent parameters related to the wave and flute sounds. Nevertheless, selection based on correlation means returns more significant results for the wave sound. For both cases, torsion has been discarded because the data derived from the motion capture recordings were very noisy.

Therefore, the selected subset of gesture parameters is $\{position, velocity, acc_N\}$. Canonical correlation analysis has been used as a selection tool ; now we apply this method in our search for the intrinsic relationships between the two sets of data. In the first step, we discard outliers

related to the first and the second canonical component. This leads to two subsets : 14 gestures among 42 for the wave example and 10 gestures for the flute example. Following the previous notations, CCA returns two projection matrices **A**, **B** whose dimensions are 3×2 and 2×2 for each gesture, respectively. Loadings are computed at each step ; figure C.3 and C.4 illustrate their statistics. The figures show the variance shared by each original variable with its canonical component for all gestures. Canonical gesture loadings are on the left side of the figures while audio descriptors respective canonical loadings are on the right. The first component is placed above the second one.

The wave case is illustrated by figure C.3 and can be interpreted as follows. Gesture parameter velocity and normal acceleration are the most represented in the first canonical component : around 90% of their variance is explained. In the audio space, one original variable is clearly highlighted : the loudness (at the top of figure C.3). In other words, the first canonical function correlates $\{velocity, acc_N\}$ to $\{loudness\}$.

Position contributes the most to the second canonical component in the gesture space while the sharpness descriptor is predominant in this case. So second canonical function correlates $\{position\}$ to $\{sharpness\}$ (at the bottom of figure C.3).

One can remark that analysis reveals that loudness and sharpness descriptors can be separated when considering sound with gesture while they were highly correlated (figure C.1).

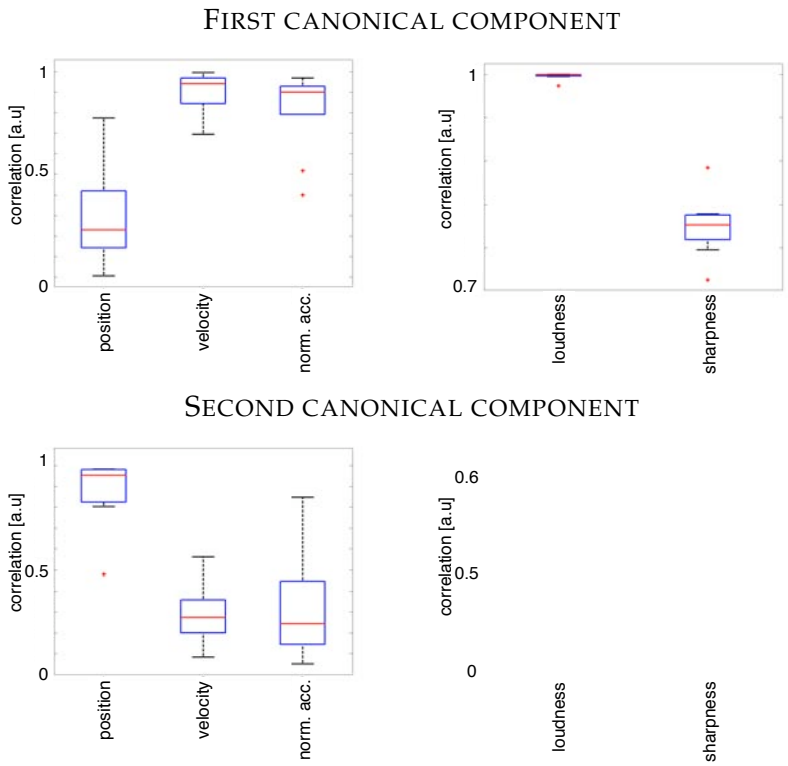

FIGURE C.3 – Canonical loadings for the wave sound. Each row is a canonical component. Gesture parameter loadings are plotted on the left while audio descriptors can be seen on the right. Top : $velocity$ and acc_N are correlated to $loudness$. Bottom : $position$ is correlated to $sharpness$.

A similar interpretation can be given for the flute timbre sound showed in figure C.4. In this case, we have :

$$\begin{array}{rrcl} \text{first function :} & \{position\} & \rightarrow & \{loudness\} \\ \text{second function :} & \{velocity, acc_N\} & \rightarrow & \{sharpness\} \end{array}$$

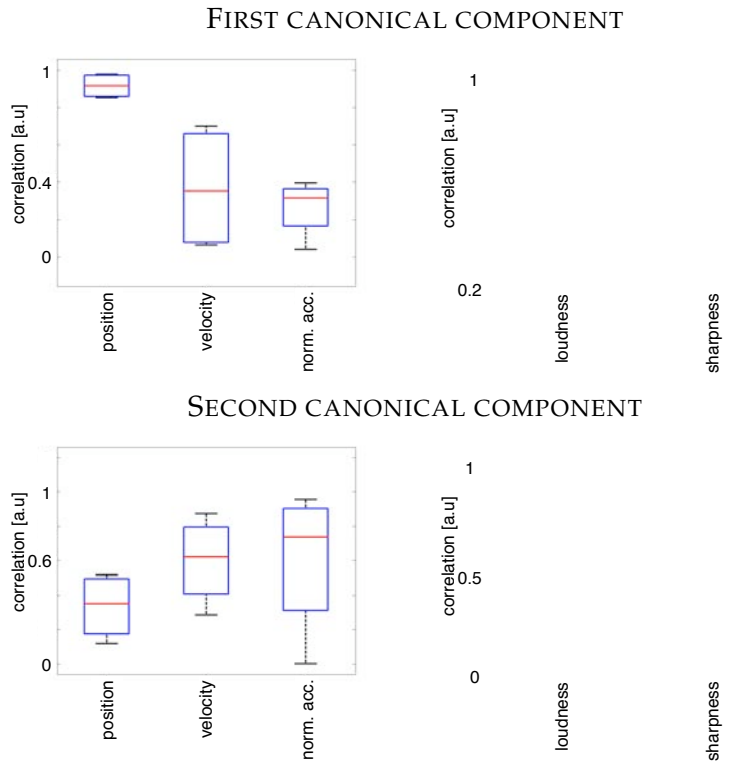

FIGURE C.4 – Canonical loadings for the flute sound. Each row is a canonical component. Gesture parameter loadings are plotted on the left while audio descriptors can be seen on the right. Top : $position$ is correlated to $loudness$. Bottom : $velocity$ and acc_N are correlated to $sharpness$.

C.5 Discussion

To analyse the cross-modal relationship between gesture and sound, a multivariate analysis method is used in two ways : first for the selection of pertinent gesture features, then for the analysis of the correlation between the selected features with the audio descriptors. In the first step, the selection yields a subset of movement features that best correlate with the audio descriptors. The low correlations obtained for some of the features have been discarded for further exploration. This seems to be coherent with kinematic studies of human gestures :

- Tangential acceleration is the acceleration component which is collinear to the velocity vector. If we consider the two-thirds power law by Viviani and Flash ($A = K.C^{2/3}$ where A is the angular velocity, C the curvature and K a constant), normal acceleration is related to curvature by $acc_N = K'.C^{1/3}$, where K' is a constant. In this case, tangential acceleration does not convey relevant information.
- The fact that curvature is no longer pertinent means there is no linear relation either between curvature and the audio descriptors or between the curvature and other gesture parameters. This result is in agreement with the previous kinematic law and can be also applied to the radius of curvature.

The next step of the analysis explores the correlation of selected movement features with the audio descriptors. The results of this analysis are correlations highlighting pertinent aspects of the gesture-sound relationship. Without surprise the subjects seem to favour gestures correlating with perceptual audio energy (*loudness*).

In the case of the wave sound, velocity or normal acceleration are highly correlated to loudness. Confronting this result with performance videos, one can see that the subjects are concerned about sonic dynamics and continuity. Increasing audio energy implies increasing velocity,

i.e. increasing kinetic energy. Here the analysis reveals that the subjects tend to embody sound energy through the energy of their movement.

On the other hand, for the gestures performed on the flute sound we observe a high correlation between the norm of the position and the loudness. Instead of embodying the sound dynamic the subjects rather tend to transcribe its temporal evolution tracing the modulation of the sound feature over time. As the variation of audio energy in the flute example is rather subtle compared to the wave sound, the subjects seem to adapt their strategy for the imagined sound control.

At last, we have started to inspect data of particular subjects that may reveal individual strategies and skills. For instance, considering the velocity feature, defined as $velocity^2 = v_x^2 + v_y^2 + v_z^2$, one can bring directional information to the analysis splitting $velocity^2$ into three specific variables : v_x^2, v_y^2, v_z^2. Canonical correlation analysis is no longer constrained to a uniform weight equal to 1 in the resulting linear combination but finds an optimal set of weights favouring directions. In other words, the analysis method takes into account the movement asymmetries. For the selection of movement parameters among a redundant set of extracted features, a trade-off has to be found between achieving a complete description of the movement and avoiding redundancies.

C.6 Conclusion and Future Works

Our goal was to study the relationship between gesture and sound. Gesture was considered as a set of kinematic parameters representing a free movement performed on a recorded sound. The sound was considered as a signal of feature observations. The method used in the paper arises from multivariate analysis research and offers a powerful tool to investigate the mutual shared variance between two sets of features. Objective results inferred from the application of CCA as a selection tool was presented. In addition, more subjective conclusions concerning mapping from the gesture parameter space to the audio descriptor space was highlighted. Thereby, we saw in this paper that gestural expression when relating to sounds can be retrieved considering gesture-sound as a pair instead of as individual entities.

However, the method suffers from some restrictive limitations. First of all, canonical functions correspond to linear relations so CCA cannot exhibit non-linear relations between variables. Besides, since we must restrict the variable sets to finite sets that encode only a part of the information contained in both gestures and sounds, the correlation (i.e. variance) as an objective function is not always relevant when real signals are analysed. The correlation involved in CCA could be replaced by the mutual information. By arising the statistical order of the multivariate relation, the main idea is to find canonical variates that are maximally dependent. It should lead to a more complete semantic interpretation of gesture-sound relationships in a musical context. To summarize, the method presented in this paper has given promising results and further works will consist in refining the method using information theory.

C.7 Acknowledgments

This work was supported by the ANR project 2PIM/MI3. Moreover, we would like to thank the COST IC0601 Action on Sonic Interaction Design (SID) for their support in the short-term scientific mission in Graz.

Annexe D

Analyzing Gesture and Sound Similarities with a HMM-Based Divergence Measure

B. Caramiaux[1], **F. Bevilacqua**[1], **N. Schnell**[1]

[1] UMR IRCAM-CNRS, Paris, France

Abstract : In this paper we propose a divergence measure which is applied to the analysis of the relationships between gesture and sound. Technically, the divergence measure is defined based on a Hidden Markov Model (HMM) that is used to model the time profile of sound descriptors. Particularly, we used this divergence to analyze the results of experiments where participants were asked to perform physical gestures while listening to specific sounds. We found that the proposed divergence is able to measure global and local differences in either time alignment or amplitude between gesture and sound descriptors.

Keywords : Gesture analysis, HMM, Divergence Measure, Gesture-Sound Relationship,

D.1 Introduction

Our research is concerned with the modelling of the relationships between gesture and sound in music performance. Several authors have recently shown the importance of these relations in the understanding of sound perception, cognitive musical representation and action-oriented meanings ((Leman, 2007), (Godøy, 2006), (Varela et al., 1991)), which constitutes a key issue for expressive virtual instrument design ((Van Nort, 2009), (Rasamimanana et al., 2009)).

A gesture is described here as a set of movement parameters measured by a motion capture system. In turn, a sound is described as a set of audio descriptors representing musical properties such as audio energy, timbre or pitch. Specifically, our goal is to propose a computational model enabling the measure of the similarities between the gesture parameters and sound descriptors.

Previous works on the quantitative analysis of the gesture-sound relationship often deal with variance-based statistical methods as principal correlation analysis (PCA) ((MacRitchie et al., 2009)) or canonical correlation analysis (CCA) (Caramiaux et al., 2010c). PCA allows for the determination of principal components that models the variation of the gesture parameters. Analyzing these components together with musical features (as tempo or metric) enabled to understand how listeners try to synchronize their movements on music beats ((MacRitchie et al., 2009), (Luck and Toiviainen, 2006)). In (Caramiaux et al., 2010c) the CCA method is used as a selection tool for mapping analysis. In this work, we showed that this method can return the gesture and sound predominant features. However, both variance-based methods suffer

from a lack of temporal modeling. Actually, these models assume as stationary both gesture parameters and audio descriptors, in the sense that statistical moments (mean, variance, etc.) do not depend on the ordering of the data. As a matter of fact, these models return a global static similarity measure without considering intrinsic dynamic changes.

To overcome these limitations, it is necessary to model the time profiles of the parameters. A large number of works dealing with time series modelling are based on hidden Markov models. HMM-based methods indeed allow for the temporal modeling of a sequence of incoming events, and have been used in audio speech recognition (Rabiner, 1989), gesture recognition ((Bobick and Wilson, 1997), (Bevilacqua et al., 2010)) and multimodal audio-visual speech recognition (Gurban, 2009). The common classification task generally considers a sequence as a unit to be classified and returns a decision once completed based on the computation of likelihood values. In (Bevilacqua et al., 2010) the authors present a HMM method designed for continuous modeling of gesture signals, that allows for the real-time assessment of the recognition process. Moreover, this method allows for the use of a single example for the learning procedure.

We propose to use in order to provide a measurement tool in a cross-modal fashion. HMM were already employed in cross-modal contexts : audio speech and video (Li and Shum, 2006), (Sargin et al., 2007). Here the novelty is to use HMM methods to model relationships between non-verbal sounds and hand gesture of passive listeners. More precisely, we propose here to use this method to further define a statistical distance between two time profiles, typically called a divergence measure (see for instance (Csiszár, 1967)) in information processing. Specifically, we report here that this HMM-based divergence measure has properties, induced by its underlying Markov process (Ephraim and Merhav, 2002), that makes it suitable to study the time evolution of the similarity between gesture parameters and sound descriptors.

This paper is structured as follows. First, we describe the general method and context of this work. Second, we present the theoretical framework of hidden Markov modeling (section D.3). In section D.4 we detail the divergence measure based on this framework and a specific learning process. Third, we report an experiment and the results that illustrate a possible use of our method (section D.5). Finally, we conclude and present future works in section D.6.

D.2 Context and Goal

Consider the following experiment : a participant listens to a specific sound several times, and then proposes a physical gesture that "mimics" the sound. The gesture is then performed (and captured) while the participant listens to the sound. Our general aim is to answer the following question : how can we analyse the gesture in relation to the sound ?

In this experiment, the gestures can be considered as a "response" to a "stimulus", which is actually the sound. In our framework, we will thus consider the sound as the "model" and the gestures as the "observations", as if they were generated by the model.

For each participant's gestures, as illustrated in figure D.1, our model should allow us to compute a divergence measure between each gesture and the corresponding sound (or in other words, to quantify similarity/dissimilarity between the gesture and sound). In the next section, we describe the mathematical framework enabling the computation of such a divergence measure. It is based on Hidden Markov Modeling permitting real time musical applications.

D.3 Hidden Markov Modeling

In this section we briefly report the theoretical HMM framework used to further define the divergence measure in section D.4.

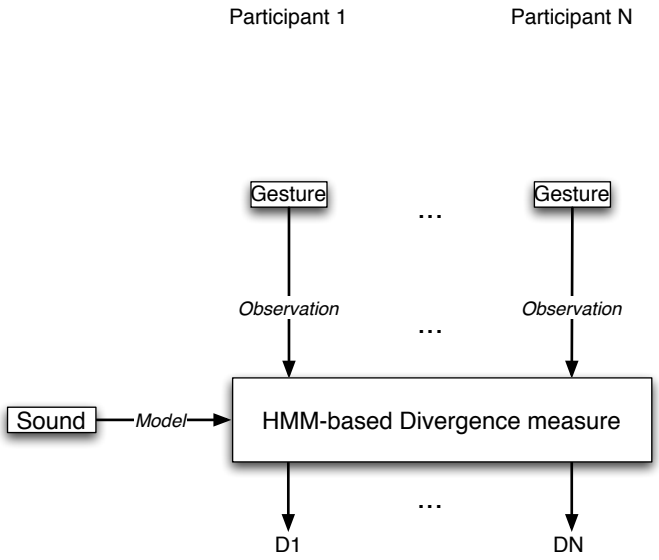

FIGURE D.1 – Methodology : Each participant's trials are taken as input and a selected sound is taken as model. We measure the divergence between each trial and the sound.

D.3.1 Definition

Hidden Markov modeling can be considered as two statistically dependent families of random sequences O, X ((Silva and Narayanan, 2006), (Ephraim and Merhav, 2002), (Rabiner, 1989)). The first family corresponds to the observations $\{O_t\}_{t\in\mathbb{N}}$ which represent measurements of a natural phenomenon. A single random variable O_t of this stochastic process takes value in a continuous finite dimensional space $\mathcal{O}$ (e.g $\mathbb{R}^p$). The second family of random process is the underlying state process $\{X_n\}_{n\in\mathbb{N}}$. A state process is a first-order time-homogenous Markov chain and takes values in a state space denoted by $\mathcal{X} = \{1, 2, \ldots, N\}$. If we note T the length of O, statistical dependency between the two processes can be written as

$$P(O_1 \ldots O_T | X_1 \ldots X_T) = \prod_{t=1}^{T} P(O_t | X_t) \tag{D.1}$$

We define a hidden Markov model as

$$\lambda = (A, B, \pi)$$

Where A is the time-invariant stochastic matrix, or transition matrix, $P(X_{t+1} = j_1 | X_t = j_2), (j_1, j_2) \in \mathcal{X}^2$; B is the time invariant observation distribution $b_j(o) = P(O_t = o | X_t = j), j \in \mathcal{X}$; and π is the initial state probability distribution $P(X_0 = j), j \in \mathcal{X}$. The HMM structure is reported in figure D.2.

In our case, $\{X_0 \ldots X_T\}$ corresponds to an index sequence of audio descriptor samples and $\{O_1 \ldots O_T\}$ a sequence of vector of samples from gesture parameter signals.

D.3.2 Topology

A and π must be fixed according to a modeling strategy. π describes where in the sequence model we start to decode. A is used to constrain the neighborhood of state j, taken at time t, in which a model state must be taken at the next time step $t + 1$. This data has a great influence on the resulting decoding computation. Let's consider two extreme situations for a forward Markov chain topology as illustrated in figure D.3.

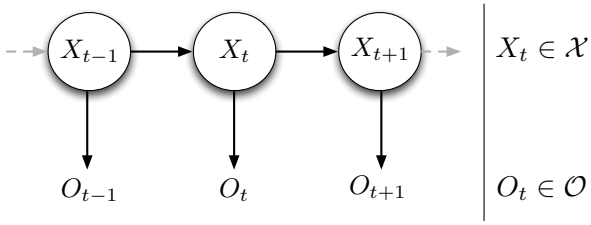

FIGURE D.2 – A general schema of HMM. $\{X_t\}_{t\in\mathbb{N}}$ is the model state random process where each state emits an observation O_t with a probability defined by B

In the first case, if current state is j we constrain to look forward until $j+1$ for the best state emitting O_{t+1} whereas in the second case we allow to look forward until the last state N to find this closest state. Usually, topology is learned from the data to have the most suitable model. Otherwise, we can tuned up the model according to a specific required behavior. For instance, as we work with continuous time series, a forward model will be chosen.

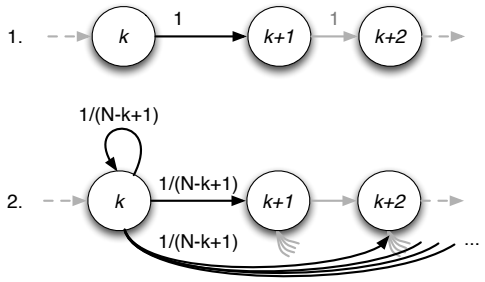

FIGURE D.3 – Two extreme cases of topology. First, one step forward is permitted in the state space. Second, each state from the current to the last one can be caught

D.3.3 Learning

Here we present how λ is learned using the approach presented in (Bevilacqua et al., 2010). The parameters A (transition probability matrix) and π (initial probability) are fixed according to user's choice of topology. B is such that time invariant observation distributions are gaussians, i.e

$$b_j(o) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-1}{2}\frac{(o-\mu_j)^2}{\sigma^2}\right] \tag{D.2}$$

Gaussian functions are centered on the model signal samples and the standard deviation σ can be adjusted by the user (see figure D.4). In our case, model signal samples are the audio feature samples computed from the chosen sound. A single example, namely the model, is needed for the learning procedure.

Thereby, rather learning based on training data, the observation probabilities are chosen such that the sound signal is the most likely observation sequence. In this way, we seek for the most likely gesture as the most similar to audio descriptor temporal evolution.

D.3.4 Decoding

Given an input sequence O and a HMM λ, one of the interesting problems is to compute the probability $P(O|\lambda)$. As mentioned in (Rabiner, 1989), in practice we usually compute the

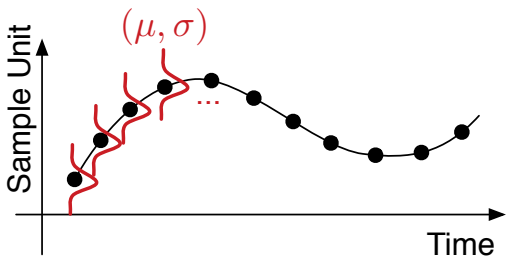

FIGURE D.4 – Learning phase. Gaussian functions are centered on the model signal samples and the standard deviation σ is *a priori* defined as a tolerance parameter.

logarithm of this probability as

$$\log\left[P(O|\lambda)\right] = \sum_{t=1}^{T} \log\left[\sum_{j=1}^{N} \alpha_t(j)\right] \tag{D.3}$$

Where $\alpha_t(i)$ is called the forward variable and is defined as $\alpha_t(i) = P(O_1 O_2 \dots O_t, X_t = i|\lambda)$, namely the probability of having the observation sequence $O_1 \dots O_t$ and the current state i. Also, this variable can be computed recursively providing an incremental method to find the desired probability (Rabiner, 1989), i.e $\forall j \in [1, N]$

$$\begin{aligned} t = 1 \quad & \alpha_1(j) = \pi_j b_j(O_1) \\ t > 1 \quad & \alpha_t(j) = \left(\sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij}\right) b_j(O_t) \end{aligned} \tag{D.4}$$

This forward inference allows for real time applications in which input signal is decoded inductively.

D.4 Divergence Measure

In this section we define the divergence measure based on the HMM framework and the learning method described in section D.3.3. Three main properties of this divergence are proved below : non-negativity ; global minimum ; non-symmetry.

D.4.1 Divergence Measure Definition

We consider two uniformly sampled signals : a model $M = \{M_1, \dots, M_N\}$ and an observation $O = \{O_1, \dots, O_T\}$. We define here the divergence measure between the observation O and a HMM learned from signal M as in section D.3.3, based on decoding presented in section D.3.4. We denote $\lambda_M = (A_M, B_M, \pi_M)$ the HMM learned from M. As mentioned in D.3.3, we fix A_M and π_M for the divergence independently to M. Observation distributions b_j^M are defined as equation (D.2) with $\mu_j = M_j$. Hence we have $\lambda_M = (A, B_M, \pi)$. We define the divergence measure as

$$D_{A,\pi}(O||M) = -\log\left[P(O|\lambda_M)\right] \tag{D.5}$$

In the following, for convenience $D_{A,\pi}$ will be noted D. Divergence measure corresponds to the logarithm of the likelihood of having the sequence of observations O given a model λ_M learned from a signal M. More precisely, $D(O\|M)$ measures the divergence between the input observation and a sequence of model states generating the observations. This sequence respects temporal structure of the model thanks to the underlying Markov chain. The result is a

temporal alignment of model states on observations with a probabilistic measure evaluating how the alignment fits the observation sequence in terms of time stretching and amplitude (cf figure D.5).

FIGURE D.5 – The HMM takes as input the sequence of observations $O_1 \dots O_t$. A sequence of model states (whose likelihood of emitting observations is maximum) approximates the observations. The quality of modeling is returned and defines $D(O\|M)$.

D.4.2 Divergence Properties

In this section, we present that divergence measure between observation O and model M defined by (D.5) satisfies important properties. We refer the reader to the appendix for more details.

Non-negativity Divergence $D(O\|M)$ is always positive. Theoretically, the divergence measure does not have to be finite. Actually, $D(O\|M)$ is finite because signals considered have a finite length ($T, N < +\infty$) and infinite values are theoretically impossible, due to numerical precision. The log of very small values can be either considered as zero or disregarded.

Lower bound. The most important corollary of non-negativity is the existence of a lower bound i.e a global minimum for our divergence measure which varies according to parameters A, π, σ. Moreover, the global minimum is explicit. Depending on A and π, the minimum $D(M\|M)$ is not necessarily zero. Minimum analysis returns how close the HMM learned from M can generate O. In section D.5.3 we will show that extremum analysis is pertinent in the analysis of the similarities between a sound and a gesture performed while listening to it.

For brevity, explicit global minimum is not reported here and its analytic formulation will not be explicitly used in the following.

Non-symmetry. The measure is not symmetric. Strategies to symmetrize divergence measures can be found in the literature (see for instance (Johnson and Sinanović, 2001) for the well known Kullback-Liebler divergence), but we are interested here in the analysis of the divergence from an observed gesture to a fixed sound model and there is *a priori* no reason why their relation should be symmetric.

D.4.3 Temporal evolution of the measure

The considered sample-based learning method trains an HMM that closely models the time evolution of the signal. Moreover, from forward decoding we can find at each time t which

model state emits the considered observation. Thus, at each time step the model can inform us on the close relation between both signals in terms of time evolution and amplitudes. This aims to an explicit temporal evolution of the divergence measure. Let any truncated observation signals be denoted by $O_{|t} = \{O_1 \dots O_t\}$ and the whole model λ_M. Hence D is defined as a function of time by,

$$D(O_{|t} \| M) = -\sum_{k=1}^{t} \log \left[\sum_{j=1}^{N} \alpha_k(j) \right] \tag{D.6}$$

D.5 Experiments

In this section we present an evaluation of the previously defined divergence measure to gesture and sound data. The measure returns an overall coefficient of the similarity between descriptors of both sound and performed gesture. Temporal evolution of this measure allows for the analysis of temporal coherence of both signals. We discuss the results at the end of this section.

D.5.1 Data Collection

The data was collected on May 2008 in the University of Music in Graz. For the experiment 20 subjects were invited to perform gestures while listening to a sequence of 18 different recorded sound extracts of a duration between 2.05 and 37.53 seconds with a mean of 9.45 seconds. Most of the sound extracts were of short duration. Since the experience was explorative, the sound corpus included a wide variety of sounds : environmental and musical of different styles (classical, rock, contemporary).

For each sound, a subject had to imagine a gesture that he or she performed three times after an arbitrary number of rehearsals. The gestures were performed with a small hand-held device that included markers for a camera-based motion capture system recording its position in Cartesian coordinates. The task was to imagine that the gesture performed with the hand-held device produces the listened sound. A foot-pedal allowed the beginning of the movement to be synchronized with the beginning of the playback of the sound extract in the rehearsal as well as for the recording of the final three performances.

D.5.2 Data Analysis

We refer the reader to the previously introduced method in figure D.1. We first select a sound as a model. This sound is *waves*. It is a sequence of five successive rising and falling ocean's waves at different amplitudes and durations. According to the sound model, we consider the three trials performed by each candidate while listening to it.

The divergence measure parameters are set as follows. The chosen transition matrix corresponding to the Markov chain topology is illustrated in figure D.6 (see (Bevilacqua et al., 2010) for further explanations). The initial probability distribution π is such that $\pi_1(O_1) = 1$ and $\forall i \neq 1, \pi_i(O_1) = 0$. The states of the Markov chain are the index of the audio descriptor samples (see section D.3.3).

The choice of audio description and gesture variables is based on our previous works (cf. (Caramiaux et al., 2010c)). We have shown that the predominant features when participants have performed gestures while listening to a wave sound is the audio loudness and gesture velocity. As we present some results based on the same data, we consider these two unidimensional signals for describing the data.

In the whole set of data captured, some trials had data missing ; for others gesture and sound were not synchronized and finally some trials were missing for some participants. A

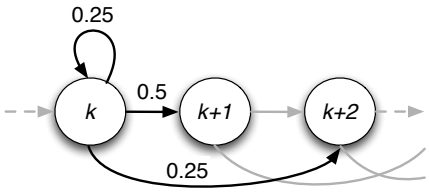

FIGURE D.6 – The chosen topology gives the predominant weight to a transition to the next state. An equal weight is given to the self-transition and to the transition above the next state.

selection is performed based on these criteria. Among the 20 participants, a set of 14 are kept. For all of the 14 participants, we measure the divergence between each trial and the selected sound. Gesture sequence for participant s and trial p is noted $O^{s,p}$, loudness signal is noted M. Figure D.7 reports the results.

In the following, we will focus result analysis on four key points.

1. *Divergence Extrema*. Participant performances for which the divergence measure is the lowest and the highest
$$\arg\min_{O^{s,p}} [D(O^{s,p} \| M)]$$
$$\arg\max_{O^{s,p}} [D(O^{s,p} \| M)]$$
2. *Gesture Variability*. Participant performances for which the standard deviation of resulting divergences is low or high.
3. *Temporal Alignment*. Alignment of the model (audio descriptor sample index) onto the incoming observations (gesture parameters) : the sequence of states returning the maximum likelihood.
4. *Temporal Evolution*. Evolution of divergence measure for the same selected participant performances as above.
$$D(O^{s,p}_{|t} \| M)$$

D.5.3 Results and Discussion

Divergence Extrema. Consider first the global minimum and maximum for divergence results obtained on the whole set of data (cf. figure D.7). It reveals that participant 4 holds the minimum 2.24 for the second trial. In the same way, participant 5 holds the maximum 19.42 for the second trial. In figure D.8, the participant 4's trial minimizing the divergence measure is plotted on the top-left together with the model. On the top-right of figure D.8, we report the participant 5's trial maximizing the divergence together with the model. It reveals that participant 4's gesture is more synchronized to the sound and the variations in velocity amplitude fit the best loudness proper variations than participant 5's performance. Actually, participant 4 tends to increase his arm's velocity synchronously with each wave falling. Otherwise, participant 5's gesture performance velocity does not globally correspond to the corresponding loudness variations.

Gesture Variability. Illustration of standard deviation between trial divergences in figure D.7 reflects the tendency of each participant to perform similar trials in terms of temporality and amplitude. Participant 7 performed very consistent trials compared to participant 4. Divergence medians suggest that a considered participant performed three different gesture performances (e.g. participant 2 or 11) or one really different compared to the remaining two (e.g. participant 4 : the first performance is very distinct from the other ones). Figure D.9 illustrates this analysis reporting the three trials performed by participants 4 and 7.

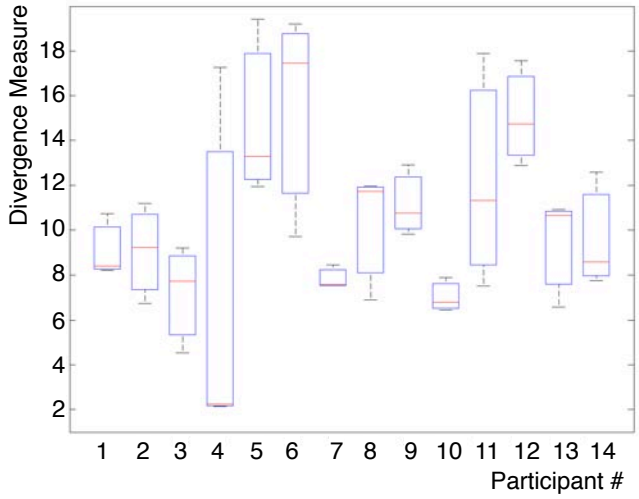

FIGURE D.7 – The figure reports statistics on divergence measures between each participant's trial and the sound *waves*. The figure reports each quartile.

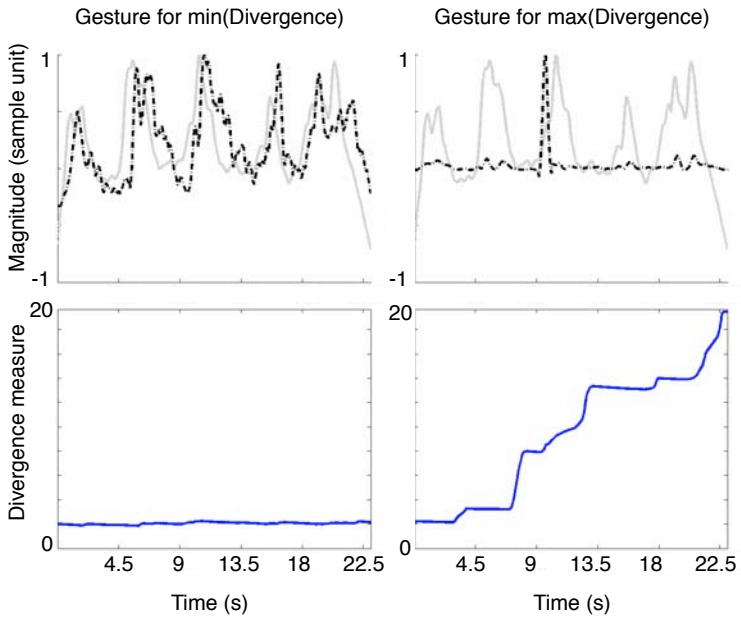

FIGURE D.8 – At the top, both gesture velocity signals are plotted in dashed line for both participant 4 (left) and participant 5 (right). The model (*waves* loudness) is also plotted in solid gray line. The bottom is divergence measure at each t between the respective signals above the plot.

In the following, temporal alignment and the resulting temporal evolution of divergence are analyzed on particular examples highlighting how we can interpret the use of such measure for cross-modal analysis.

Temporal Alignment. The divergence measure drastically decreases if both signal amplitude variations differ (see figure D.8). A standard correlation measure would behave similarly. The underlying stochastic structure overcomes this limitation by aligning both signals taking into account the ordering of the data. Figure D.10 illustrates participant 10's second performance :

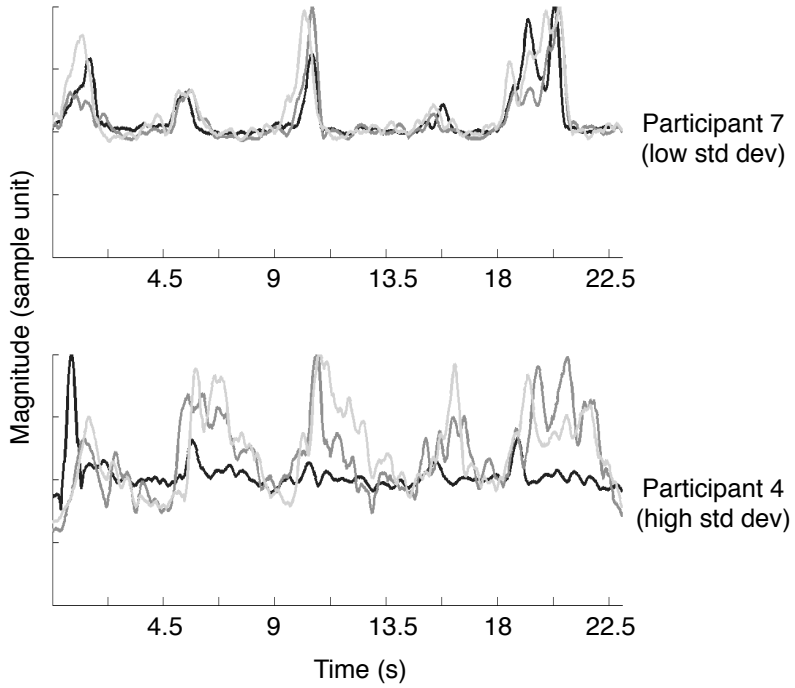

FIGURE D.9 – At the top are the trials for which variance in divergence measure is the lowest. Below we plot trials performed by participant 5 and 6 corresponding to the highest variance. Divergence median for participant 5 is roughly the mean (three different trials) of divergence values whereas divergence median for participant 6 is very low (one very different trial from the others)

at the top, original signals (*waves*' loudness and gesture's velocity) ; at the bottom, the aligned loudness onto the gesture's velocity signal. Even if both signals are not strictly synchronous, the divergence is quite low (6.79). Actually, both signal shapes are globally coherent. The alignment is roughly a time shift of the sound signal resulting from a delayed gesture during the performance. In this example, correlation coefficient before the alignment process would be 0.076 and 0.32 afterwards. Resulting aligned sound could be reconstructed and strategies of reconstruction should be investigated.

Temporal Evolution. As explained in section D.4.3, the quality of model state sequence according to observation signal can be measured at each time t. At the bottom of figure D.8 are the divergence measures evolving over time for the second trial of participant 4 (left) and the third of participant 5 (right). On the one hand, let's analyze bottom left plot corresponding to participant 4's performance (see figure D.11 for a better view of the divergence curve). The first samples of O and M are similar. Incoming observations have a tiny delay and the algorithm realigns both signals. The divergence decreases meaning that amplitudes are close (relatively to σ) and the signals are quite synchronous. Around 2 seconds, the divergence increases : gesture velocity is very low whereas sound loudness is still high. Performer's movement changed of direction involving a decreasing velocity. A peak of divergence informs us at which time a divergence occurs and its magnitude permits the degree of mismatching to be evaluated. In this example, a magnitude of 0.1 represents a small mismatch as illustrated in figure D.11 (top part). Thanks to the underlying stochastic structure, the state sequence corrects itself according to the new inputs. Indeed, the divergence measure is then decreasing slowly since the sum over time (from 1 to t, see equation D.3) of the log-probabilities induces a memory of the

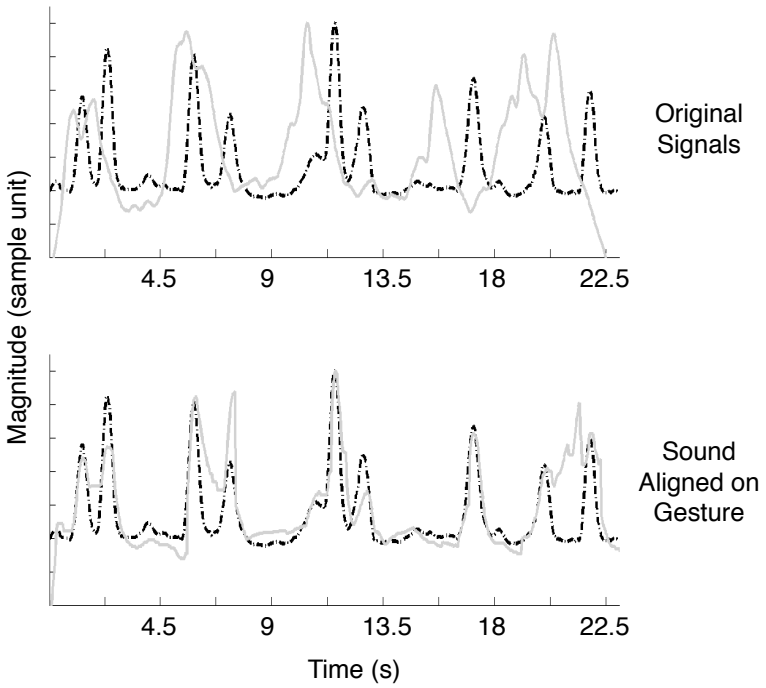

FIGURE D.10 – Temporal alignment of loudness onto gesture's velocity. At the top are plotted the original signals : gesture's velocity in dashed line and loudness in solid line. At the bottom, gesture's velocity is unchanged and loudness is aligned onto the velocity signal.

past signals' mismatching. Global shape presents sawtooth-type variations interpreted as local mismatching (peak which magnitude depends on the amplitude difference) and correction (release) (see figure D.11).

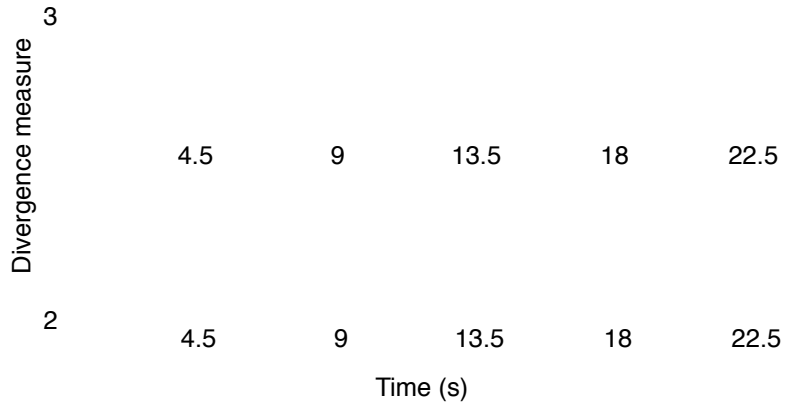

FIGURE D.11 – Zoom on divergence measure curve for participant 4. Zooming into this curve illustrates sawtooth-type behavior of the divergence.

Consider now gesture performed by participant 5, shown in the right part of figure D.8. The global evolution of the divergence measure is increasing indicating that they globally diverge, contrary to the previous behavior, and its magnitude is higher. The temporal shape shows constant parts (as around 4sec, 9sec, 13.5sec and 18sec). During these intervals, mismatching has less impact because amplitude of both signals is lower. The peaks occur for non-

synchronized peaks meaning highly divergent amplitude values. Contrary to the respective bottom-left plot, no decreasing can be seen due to the overall past divergence values that are not good enough to involve a decrease in the divergence : as seen before, the sum propagates past mismatching.

Thereby, two different dynamic behaviors for the divergence measure have been highlighted. Locally mismatching induced a saw shape for $D(O_{|t}\|M)$ whereas globally mismatching induced an ascending temporal curve which can roughly be approximated as piecewise constant. These behaviors give us useful hints to understand dynamic relationships between gesture and the sound which was listened to highlighting relevant parts of the signals where both signals are coherent or really distinct. Unfortunately, the current model does allow the speed of the decrease to be parametrized in the model. Otherwise, since the method considers a global model corresponding to the whole sound signal, it should be interesting to analyze gesture-sound relationship at an intermediate temporal scale between the sample and the global signal. Indeed, changes in gesture control could occur permitting a better fitting between loudness and velocity but the global divergence measure should not take such dynamic changes into account.

D.6 Conclusions

In this paper we have presented a divergence measure based on a HMM that is used to model the time profile of sound descriptors. Gestures are considered as observations for the HMM as if they were generated by the model. The divergence measure allows similarity/dissimilarity between the gesture and sound to be quantified. This divergence has the following properties : non-negativity ; global minimum ; non-symmetry. Experiments on real data have shown that the divergence measure is able to analyze either local or global relationships between physical gesture and the sound which was listened to in terms of time stretching and amplitude variations. Some constraints (changing parameters A, π or σ) could be added in order to reinforce or relax softness of the measure. The novelty is to use HMM methods to model relationships between non-verbal sounds and hand gesture of passive listeners. The use of HMM is motivated by possible real time implementation and interactive applications.

Actually, we are designing a gesture-driven sound selection system whose scenario is as follows. First, we build a sound corpus of distinct audio files with specific dynamic, timbre or melodic characteristics (environmental sounds, musical sounds, speech, etc.). Then we choose an interface allowing physical gesture capturing (e.g. WiiMote). Finally one can perform a gesture and the system will automatically choose the sound for which the divergence measure returns the minimal value. Such application could be useful for game-oriented systems, artistic installations or sound-design software.

D.7 Annex

D.7.1 Divergence Measure Properties

Non-negativity.

$$\forall t \in [1,T], \sum_{i=1}^{N} \alpha_t(i) = P(O_1 \dots O_t|\lambda_M) \in [0,1]$$

Hence,

$$D(O\|M) = -\sum_{t=1}^{T} \log \left[\sum_{j=1}^{N} \alpha_t(j) \right] \in [0,+\infty] \tag{D.7}$$

Lower bound.

Function $b_j^M(o)$ holds a global maximum in $\mathbb{R}^p$ for

$$\forall j \in [1, N],\ M_j = \arg\max_x b_j^M(x)$$

For brevity, the whole demonstration is not reported here, but it can be shown that this global maximum aims to a global maximum for $\alpha_t(j)$ leading to a global minimum for the divergence measure $D(O\|M)$ considering any inputs different from the model.

$$\forall O \neq M, D(O\|M) \geq D(M\|M) \tag{D.8}$$

Non-symmetry. From equation (D.4), let $\alpha_t(j)$ be rewritten as

$$\forall t \geq 1,\ \alpha_t(j) = C_{t,j} b_j(O_t)$$

Where $C_{1,j} = \pi_j$ and $C_{t,j} = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij}$. From respective expression of $D(O\|M)$ and $D(M\|O)$, we have $\forall t \geq 1$,

$$\sum_{j=1}^{N} \frac{C_{t,j}}{\sigma\sqrt{2\pi}} e^{\frac{-(O_t - M_j)^2}{2\sigma^2}} \neq \sum_{j=1}^{N} \frac{C_{t,j}}{\sigma\sqrt{2\pi}} e^{\frac{-(M_t - O_j)^2}{2\sigma^2}}$$

Meaning that the divergence is not symmetric.

$$D(O\|M) \neq D(M\|O) \tag{D.9}$$

Bibliographie

Achan, K., Roweis, S., and Frey, B. (2004). A segmental hmm for speech waveforms. Technical report, Univerity of Toronto Technical Report UTML-TR-2004-001. 75

Alexa, M. and Müller, W. (2000). Representing animations by principal components. In *Computer Graphics Forum*, pages 411–418. Wiley Online Library. 141

Appert, C. and Bau, O. (2010). Scale detection for a priori gesture recognition. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 879–882. ACM. 82, 100

Arikan, O., Forsyth, D., and O'Brien, J. (2003). Motion synthesis from annotations. In *ACM SIGGRAPH 2003 Papers*, pages 402–408. ACM. 103

Artières, T., Marukatat, S., and Gallinari, P. (2007). Online handwritten shape recognition using segmental hidden markov models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(2) :205–217. 76, 104, 143

Arulampalam, M., Maskell, S., Gordon, N., and Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on signal processing*, 50(2) :174–188. 74, 83, 86, 148

Aziz-Zadeh, L., Iacoboni, M., Zaidel, E., Wilson, S., and Mazziotta, J. (2004). Left hemisphere motor facilitation in response to manual action sounds. *European Journal of Neuroscience*, 19(9) :2609–2612. 36

Ballas, J. (1993). Common factors in the identification of an assortment of brief everyday sounds. *Journal of Experimental Psychology : Human Perception and Performance*, 19(2) :250–267. 43

Barbič, J., Alla, S., Jia-Yu, P., Faloutsos, C., Hodgins, J., and Pollard, N. (2004). Segmenting motion capture data into distinct behaviors. In *Proceedings of the 2004 Conference on Graphics Interface. London, Ontario, Canada*, pages 185–194. 103, 140

Bau, O. and Mackay, W. (2008). Octopocus : a dynamic guide for learning gesture-based command sets. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*, pages 37–46. ACM. 82, 100

Bengio, Y. and Frasconi, P. (1995). An input/output hmm architecture. In *Advances in Neural Information Processing Systems, NIPS 7*, pages 427–434. MIT Press. 74

Bergmann, K. and Kopp, S. (2007). Co-expressivity of speech and gesture : Lessons for models of aligned speech and gesture production. *Symposium at the AISB Annual Convention : Language, Speech and Gesture for Expressive Characters*, pages 153–158. 154

Berthoz, Alain (1997). *Le Sens du mouvement*. Odile Jacob, Paris, France. 8, 9, 36, 154

Bevilacqua, F., Baschet, F., and Lemouton, S. (2012). The augmented string quartet : experiments and gesture following. *Journal of New Music Research (Accepted)*. 82, 83

Bevilacqua, F., Guédy, F., Schnell, N., Fléty, E., and Leroy, N. (2007). Wireless sensor interface and gesture-follower for music pedagogy. In *Proceedings of the 7th international conference on New interfaces for musical expression*, pages 124–129. ACM. 82

Bevilacqua, F., Müller, R., and Schnell, N. (2005). Mnm : a max/msp mapping toolbox. In *Proceedings of the 2005 conference on NIME*, pages 85–88. National University of Singapore. 125

Bevilacqua, F., Ridenour, J., and Cuccia, D. (2003). 3d motion capture data : motion analysis and mapping to music. In *Proceedings of the workshop/symposium on sensing and input for media-centric systems*. Citeseer. 141

Bevilacqua, F., Schnell, N., and Fdili Alaoui, S. (2011a). Gesture capture : Paradigms in interactive music/dance systems. *Emerging Bodies : The Performance of Worldmaking in Dance and Choreography*, page 183. 82

Bevilacqua, F., Schnell, N., Rasamimanana, N., Zamborlin, B., and Guédy, F. (2011b). Online gesture analysis and control of audio processing. *Musical Robots and Interactive Multimodal Systems*, pages 127–142. 82, 83

Bevilacqua, F., Zamborlin, B., Sypniewski, A., Schnell, N., Guédy, F., and Rasamimanana, N. (2010). Continuous realtime gesture following and recognition. In *In Embodied Communication and Human-Computer Interaction, volume 5934 of Lecture Notes in Computer Science*, pages 73—84. Springer Berlin / Heidelberg. 82, 83, 88, 99, 127, 150, 164, 166, 169

Bilmes, J. (2002). What hmms can do. Technical report, University of Washington, Department of EE, Seattle WA, 98195-2500. 83, 142

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer New York. 71, 72

Black, M. and Jepson, A. (1998a). A probabilistic framework for matching temporal trajectories : Condensation-based recognition of gestures and expressions. *Computer Vision (ECCV 98)*, pages 909–924. 83, 88, 97, 98, 100

Black, M. and Jepson, A. (1998b). Recognizing temporal trajectories using the condensation algorithm. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 16–21. IEEE. 75, 86, 149

Bloit, J., Rasamimanana, N., and Bevilacqua, F. (2010). Modeling and segmentation of audio descriptor profiles with segmental models. *Pattern Recognition Letters*. 76, 104, 143

Bobick, A. F. and Wilson, A. D. (1997). A state-based approach to the representation and recognition of gesture. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(12) :1325–1337. 82, 83, 164

Bonebright, T. (2001). Perceptual structure of everyday sounds : A multidimensional scaling approach. In *Proceedings of the 2001 International Conference on Auditory Display*. 35

Bouenard, A. (2009). *Synthesis of Music Performances : Virtual Character Animation as a Controller of Sound Synthesis.* PhD thesis, European University of Brittany, France. 12

Bouenard, A., Wanderley, M., and Gibet, S. (2010). Gesture control of sound synthesis : Analysis and classification of percussion gestures. *Acta Acustica united with Acustica*, 96(4) :668–677. 13

Bowler, I., Purvis, A., Manning, P., and Bailey, N. (1990). On mapping n articulation onto m synthesiser-control parameters. In *Proceedings of the 1990 International Computer Music Conference*, pages 181–184. 17

Brand, M. and Hertzmann, A. (2000). Style machines. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 183–192. ACM Press/Addison-Wesley Publishing Co. 74, 83, 103

Brand, M., Oliver, N., and Pentland, A. (1997). Coupled hidden markov models for complex action recognition. In *cvpr*, page 994. Published by the IEEE Computer Society. 74

Bregler, C. (1997). Learning and recognizing human dynamics in video sequences. In *cvpr*, page 568. Published by the IEEE Computer Society. 78

Bretzner, L., Laptev, I., and Lindeberg, T. (2002). Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 423–428. IEEE. 83

Brown, A. and Hinton, G. (2001). Products of hidden markov models. In *Proceedings of Artificial Intelligence and Statistics*, pages 3–11. Citeseer. 74

Bryson, A. and Ho, Y. (1979). *Applied optimal control*. American Institute of Aeronautics and Astronautics. 9

Buchsbaum, D., Griffiths, T. L., Gopnik, A., and Baldwin, D. (2009). Learning from actions and their consequences : Inferring causal variables from continuous sequences of human action. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. 103, 134

Bui, H. (2003). A general model for online probabilistic plan recognition. In *International Joint Conference on Artificial Intelligence*, volume 18, pages 1309–1318. Citeseer. 77

Bui, H., Phung, D., and Venkatesh, S. (2004). Hierarchical hidden markov models with general state hierarchy. In *Proceedings of the National Conference on Artificial Intelligence*, pages 324–329. Menlo Park, CA ; Cambridge, MA ; London ; AAAI Press ; MIT Press ; 1999. 77

Cabe, P. and Pittenger, J. (2000). Human sensitivity to acoustic information from vessel filling. *Journal of experimental psychology : human perception and performance*, 26(1) :313. 34

Caclin, A., McAdams, S., Smith, B., and Winsberg, S. (2005). Acoustic correlates of timbre space dimensions : A confirmatory study using synthetic tones. *The Journal of the Acoustical Society of America*, 118 :471. 36

Cadoz, C. (1988). Instrumental gesture and musical composition. In *Proceedings of the 1988 International Computer Music Conference*, pages 1–12. 12

Cadoz, C. and Wandereley, M. (2000). Gesture-music. *Trends in Gestural Control of Music*. 12, 101

Cadoz, C. and Wanderley, M. M. (2000). Gesture-music. In *Trends in Gestural Control of Music*, pages 1–55. Ircam, Paris, France. 157

Camurri, A., Lagerlöf, I., and Volpe, G. (2003). Recognizing emotion from dance movement : comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies*, 59(1-2) :213–225. 58, 154

Caramiaux, B., Bevilacqua, F., and Schnell, N. (2010a). Analysing gesture and sound similarities with a hmm-based divergence measure. In *Proceedings of the 6th Sound and Music Conference*, Barcelona, Spain. 22, 127, 129

Caramiaux, B., Bevilacqua, F., and Schnell, N. (2010b). Mimicking sound with gesture as interaction paradigm. Technical report, IRCAM - Centre Pompidou. 51

Caramiaux, B., Bevilacqua, F., and Schnell, N. (2010c). Towards a gesture-sound cross-modal analysis. In *In Embodied Communication and Human-Computer Interaction, volume 5934 of Lecture Notes in Computer Science*, pages 158–170. Springer Verlag. 21, 22, 25, 42, 50, 51, 54, 58, 60, 102, 124, 163, 169

Caramiaux, B., Bevilacqua, F., and Schnell, N. (2011). Sound selection by gestures. In *New Interfaces for Musical Expression (NIME2011)*. 66, 123, 126

Caramiaux, B., Montecchio, N., and Bevilacqua, F. (2012a). Realtime adaptive continuous gesture recognition. *(in review)*. 78

Caramiaux, B., Susini, P., Houix, O., and Bevilacqua, F. (expected 2012b). Study of the impact of sound causality on gesture responses. *(in review)*. 38

Caramiaux, B., Wanderley, M. M., and Bevilacqua, F. (2012c). Segmenting and parsing instrumentalists' gestures. *Journal of New Music Research*, 41(1) :13–29. 79

Carello, C., Anderson, K., and Kunkler-Peck, A. (1998). Perception of object length by sound. *Psychological Science*, 9(3) :211. 34

Cermak, G. and Cornillon, P. (1976). Multidimensional analyses of judgments about traffic noise. *Journal of the Acoustical Society of America*, 59(6) :1412–1420. 35

Chadabe, J. (2002). The limitations of mapping as a structural descriptive in electronic instruments. In *Proceedings of the 2002 conference on New interfaces for musical expression*, pages 1–5. National University of Singapore. 16, 17, 18

Chen, J., Penhune, V., and Zatorre, R. (2008). Moving on time : Brain network for auditory-motor synchronization is modulated by rhythm complexity and musical training. *Journal of Cognitive Neuroscience*, 20(2) :226–239. 57

Chen, M., Kundu, A., and Zhou, J. (1994). Off-line handwritten word recognition using a hidden markov model type stochastic network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 481–496. 73

Chion, M. (1983). *Guide des objets sonores : Pierre Schaffer et la recherche musicale*. Buchet/Chastel. 33, 37, 38

Comaniciu, D., Ramesh, V., and Meer, P. (2003). Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 564–577. 74

Cont, A., Dubnov, S., and Assayag, G. (2007). Anticipatory model of musical style imitation using collaborative and competitive reinforcement learning. *Anticipatory Behavior in Adaptive Learning Systems*, pages 285–306. 134

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2 :229–318. 164

Daffertshofer, A., Lamoth, C., Meijer, O., and Beek, P. (2004). Pca in studying coordination and variability : a tutorial. *Clinical Biomechanics*, 19(4) :415–428. 141

Dahl, S. (2000). The playing of an accent–preliminary observations from temporal and kinematic analysis of percussionists. *Journal of New Music Research*, 29(3) :225–233. 101, 102

Dahl, S. (2006). *On the beat : Human movement and timing in the production and perception of music"*. PhD thesis, KTH, Royal Institute of Technology, Stockholm, Sweden. 12

Dahl, S., Bevilacqua, F., Bresin, R., Clayton, M., Leante, L., Poggi, I., and Rasamimanana, N. (2009). *Gestures in performance*, chapter 3. Routeledge. Rolf Inge Godøy and Marc Leman editors. 12

Dahl, S. and Friberg, A. (2003). Expressiveness of musician's body movements in performances on marimba. *Gesture-Based Communication in Human-Computer Interaction, 5th International Gesture Workshop, GW 2003*, LNAI 2915 :479–486. 12, 13, 154

Dannenberg, R., Birmingham, W., Pardo, B., Hu, N., Meek, C., and Tzanetakis, G. (2007). A comparative evaluation of search techniques for query-by-humming using the musart testbed. *Journal of the American Society for Information Science and Technology*, 58(5) :687–701. 124

Davidson, J. (1993). Visual perception of performance manner in the movements of solo musicians. *Psychology of Music*, 21(2) :103. 13, 102

De La Torre, F. and Black, M. (2003). A framework for robust subspace learning. *International Journal of Computer Vision*, 54(1) :117–142. 140

Delalande, F. (1988). La gestique de gould : éléments pour une sémiologie du geste musical. *G.Guertin, ed. Glenn Gould, Pluriel*, pages 83–111. 12, 13

Dellaert, F., Fox, D., Burgard, W., and Thrun, S. (1999). Monte carlo localization for mobile robots. In *Robotics and Automation, 1999. Proceedings. 1999 IEEE International Conference on*, volume 2, pages 1322–1328. Ieee. 74

Demoucron, M. (2008). *On the control of virtual violins - Physical modeling and control of bowed string instruments*. PhD thesis, KTH, Royal Institute of Technology and University Paris VI. 12

Digalakis, V. (1992). *Segment-based stochastic models of spectral dynamics for continuous speech recognition*. PhD thesis, Elect. Comput. Syst. Eng. Dept., University of Boston. 76

Ding, Y. and Fan, G. (2007). Segmental hidden markov models for view-based sport video analysis. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE. 76

Dobrian, C. (2004). Strategies for continuous pitch and amplitude tracking in real-time interactive improvisation software. In *Proceedings of the 2004 Sound and Music Computing conference (SMC04)*. 15

Douc, R. and Cappé, O. (2005). Comparison of resampling schemes for particle filtering. In *Image and Signal Processing and Analysis, 2005. ISPA 2005. Proceedings of the 4th International Symposium on*, pages 64–69. IEEE. 86, 148

Doucet, A., De Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo methods in practice*. Springer Verlag. 83

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing*, 10(3) :197–208. 74, 148

Dubnov, S., McAdams, S., and Reynolds, R. (2006). Structural and affective aspects of music from statistical audio signal analysis. *Journal of the American Society for Information Science and Technology*, 57(11) :1526–1536. 15

Durbin, R. (1998). *Biological sequence analysis : Probabilistic models of proteins and nucleic acids*. Cambridge university press. 141

Eitan, Z. and Granot, R. (2006). How music moves : Musical parameters and listeners' images of motion. *Music Perception*, 23(3) :221–248. 41, 58, 64

Ekman, P. and Friesen, W. (1969). The repertoire of nonverbal behavior : Categories, origins, usage, and coding. *Semiotica*, 1(1) :49–98. 11

Engel, K., Flanders, M., and Soechting, J. (1997). Anticipatory and sequential motor control in piano playing. *Experimental brain research*, 113(2) :189–199. 101

Ephraim, Y. and Merhav, N. (2002). Hidden markov processes. *IEEE Trans. on Info. Theory*, 48(6) :1518–1569. 164, 165

Fadiga, L., Craighero, L., Buccino, G., and Rizzolatti, G. (2002). Short communication : Speech listening specifically modulates the excitability of tongue muscles : A tms study. *European Journal of Neuroscience*, 15 :399–402. 34

Fiebrink, R. (2011). *Real-time Human Interaction with Supervised Learning Algorithms for Music Composition and Performance*. PhD thesis, Princeton University. 12

Fiebrink, R., Cook, P., and Trueman, D. (2011). Human model evaluation in interactive supervised learning. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 147–156. ACM. 134

Fine, S., Singer, Y., and Tishby, N. (1998). The hierarchical hidden markov model : Analysis and applications. *Machine learning*, 32(1) :41–62. 77, 144, 145

Flash, T. and Hogan, N. (1985). The coordination of arm movements : an experimentally confirmed mathematical model. *The journal of Neuroscience*, 5(7) :1688. 9

Fletcher, P., Lu, C., Pizer, S., and Joshi, S. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *Medical Imaging, IEEE Transactions on*, 23(8) :995–1005. 141

Fod, A., Matarić, M., and Jenkins, O. (2002). Automated derivation of primitives for movement classification. *Autonomous robots*, 12(1) :39–54. 103

Fodor, I. (2002). A survey of dimension reduction techniques. *Center for Applied Scientific Computing, Lawrence Livermore National Laboratory*. 139

Forbes, K. and Fiume, E. (2005). An efficient search algorithm for motion data using weighted pca. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 67–76. ACM. 82

Françoise, J., Caramiaux, B., and Bevilacqua, F. (2011). Realtime segmentation and recognition of gestures using hierarchical markov models. Technical report, Ircam- Centre Pompidou, University Paris 6, Telecom ParisTech. 134

Françoise, J., Caramiaux, B., and Bevilacqua, F. (2012). A hierarchical approach for the design of gesture-to-sound mappings. In *Sound and Music Computing (SMC'2012)*. 134

Fremiot, M., Mandelbrojt, J., Formosa, M., Delalande, G., Pedler, E., P.Malbosc, and Gobin, P. (1996). Les unités sémiotiques temporelles : éléments nouveaux d'analyse musicale. *Diffusion ESKA. MIM Laboratoire Musique et Informatique de Marseille (1996), documents musurgia*, 13. 15

Gales, M. and Young, S. (1993). Segmental hmms for speech recognition. In *Proc. Eurospeech*, volume 3, pages 1579–1582. 76

Gallese, V., Fadiga, L., Fogassi, L., and Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119(2) :593. 9

Gaver, W. (1993a). How do we hear in the world ? explorations in ecological acoustics. *Ecological psychology*, 5(4) :285–313. 35, 37, 42, 53, 54

Gaver, W. (1993b). What in the world do we hear ? : An ecological approach to auditory event perception. *Ecological psychology*, 5(1) :1–29. 35, 37, 42, 53, 54

Gavrila, D. and Davis, L. (1995). Towards 3-d model-based tracking and recognition of human movement : a multi-view approach. In *International workshop on automatic face-and gesture-recognition*, pages 272–277. Citeseer. 82

Ge, X. and Smyth, P. (2000). Segmental semi-markov models for change-point detection with applications to semiconductor manufacturing. Technical report, Citeseer. 76

Gérard, Y. (2004). *Mémoire sémantique et sons de l'environnement*. PhD thesis, Université de Bourgogne. 35, 42

Geweke, J. (1989). Bayesian inference in econometric models using monte carlo integration. *Econometrica : Journal of the Econometric Society*, pages 1317–1339. 148

Ghahramani, Z. and Jordan, M. (1997). Factorial hidden markov models. *Machine learning*, 29(2) :245–273. 74

Giordano, B. and Mcadams, S. (2006). Material identification of real impact sounds : Effects of size variation in steel, glass, wood, and plexiglass plates. *The Journal of the Acoustical Society of America*, 119 :1171. 34

Glardon, P., Boulic, R., and Thalmann, D. (2004). Pca-based walking engine using motion capture data. In *In : Proceedings. Computer Graphics International*. IEEE Computer Society. 102, 141

Glass, J. (2003). A probabilistic framework for segment-based speech recognition. *Computer Speech & Language*, 17(2-3) :137–152. 75, 76

Godøy, R. (2009). *Musical gestures : Sound, movement, and meaning*. Routledge. 12

Godøy, R. and Jensenius, A. (2009). Body movement in music information retrieval. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*. 58

Godøy, R. I. (2001). Imagined action, excitation, and resonance. *Musical imagery*, pages 237–250. 37

Godøy, R. I. (2006). Gestural-sonorous objects : embodied extensions of schaeffer's conceptual apparatus. *Organised Sound*, 11(2) :149–157. 18, 21, 24, 33, 38, 54, 57, 163

Godøy, R. I., Haga, E., and Jensenius, A. R. (2005). Playing "air instruments" : Mimicry of sound-producing gestures by novices and experts. *Gesture in Human-Computer Interaction and Simulation, 6th International Gesture Workshop, GW 2005*, 3881/2006 :256–267. 14, 41, 154

Godøy, R. I., Haga, E., and Jensenius, A. R. (2006a). Exploring music-related gestures by sound-tracing. a preliminary study. In *2nd International Symposium on Gesture Interfaces for Multimedia Systems (GIMS2006)*. 14, 26

Godøy, R. I., Haga, E., and Jensenius, A. R. (2006b). Exploring music-related gestures by sound-tracing : A preliminary study. In *Proceedings of the COST287-ConGAS 2nd International Symposium on Gesture Interfaces for Multimedia Systems (GIMS2006)*. 42, 54, 58

Godøy, R. I., Haga, E., and Jensenius, A. R. (2006c). Playing "air instruments" : Mimicry of sound-producing gestures by novices and experts. In *Lecture Notes in Computer Science*. Springer-Verlag. 42

Godøy, R. I., Jensenius, A., and Nymoen, K. (2010). Chunking in music by coarticulation. *Acta Acustica united with Acustica*, 96(4) :690–700. 19, 103

Goldstone, R. and Kersten, A. (2003). Concepts and categorization. *Handbook of psychology*. 35

Goudeseune, C. (2002). Interpolated mappings for musical instruments. *Organised Sound*, 7(2) :85–96. 17

Grey, J. and Moorer, J. (1977). Perceptual evaluations of synthesized musical instrument tones. *The Journal of the Acoustical Society of America*, 62 :454. 35

Guastavino, C. (2007). Categorization of environmental sounds. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 61(1) :54. 35

Guerra-Filho, G. and Aloimonos, Y. (2007). A language for human action. *Computer*, 40(5) :42–51. 102, 103

Gurban, M. (2009). *Multimodal Feature Extraction and Fusion for Audio-Visual Speech Recognition*. PhD thesis, École Polytechnique Fédérale de Lausanne. 23, 164

Guyot, F. (1996). *Étude de la perception sonore en termes de reconnaissance et d'appréciation qualitative : une approche par la catégorisation*. PhD thesis, Université du Maine. 35

Gygi, B., Kidd, G., and Watson, C. (2004). Spectral-temporal factors in the identification of environmental sounds. *The Journal of the Acoustical Society of America*, 115 :1252. 44

Haga, Egil (2008). *Correspondences between music and body movement*. PhD thesis, University of Oslo, Department of Musicology. 154

Hair, J., Black, W., Babin, B., Anderson, R., and Tatham, R. (1998). *Multivariate data analysis*, volume 5. Prentice hall New Jersey. 60

Hair, Joseph F., Black, William C., Babin, Barry J., and Anderson, Rolph E. (February, 2009). *Multivariate Data Analysis (7th Edition)*. Prentice Hall, New Jersey, USA. 156

Handschin, J. and Mayne, D. (1969). Monte carlo techniques to estimate the conditional expectation in multi-stage non-linear filtering. *International journal of control*, 9(5) :547–559. 74

Harris, C. and Wolpert, D. (1998). Signal-dependent noise determines motor planning. *Nature*, 394(6695) :780–784. 9

Heloir, A., Courty, N., Gibet, S., and Multon, F. (2006). Temporal alignment of communicative gesture sequences. *Computer animation and virtual worlds*, 17(3-4) :347–357. 82

Hermann, T., Hansen, M., and Ritter, H. (2001). Sonification of markov chain monte carlo simulations. In *Proc. of 7th Int. Conf. on Auditory Display*, pages 208–216. Citeseer. 75

Hollands, K., Wing, A., and Daffertshofer, A. (2004). Principal components analysis of contemporary dance kinematics. In *Proceedings of the 3rd IEEE EMBSS UK &RI PostGraduate Conference in Biomedical Engineering & Medical Physics. Southampton, England*. 141

Holmes, W. and Russell, M. (1995). Experimental evaluation of segmental hmms. In *icassp*, pages 536–539. IEEE. 76

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4) :321–377. 23, 155

Howard Jr, J. (1977). Psychophysical structure of eight complex underwater sounds. *The Journal of the Acoustical Society of America*, 62 :149. 35

Hu, J., Brown, M., and Turin, W. (1996). Hmm based online handwriting recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(10) :1039–1045. 73

Hunt, A. and Wanderley, M. (2002). Mapping performer parameters to synthesis engines. *Organised Sound*, 7(2) :97–108. 17

Iazzetta, Fernando (2000). Meaning in musical gesture. *Trends in Gestural Control of Music*, pages 259–268. 16

Isard, M. and Blake, A. (1998a). Condensation conditional density propagation for visual tracking. *International journal of computer vision*, 29(1) :5–28. 75, 83

Isard, M. and Blake, A. (1998b). A mixed-state condensation tracker with automatic model-switching. In *Computer Vision, 1998. Sixth International Conference on*, pages 107–112. IEEE. 75

Jain, A. K., Duin, R. P. W., and Mao, J. (2000). Statistical pattern recognition : A review. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(1) :217–220. 22

Jang, J., Lee, H., and Yeh, C. (2001). Query by tapping : A new paradigm for content-based music retrieval from acoustic input. *Advances in Multimedia Information Processing*, pages 590–597. 124

Jeannerod, M. (1997). *The cognitive neuroscience of action.* Blackwell Publishing. 8, 9

Jenkins, O. and Matarić, M. (2004). A spatio-temporal extension to isomap nonlinear dimension reduction. In *Proceedings of the twenty-first international conference on Machine learning*, page 56. ACM. 141

Jensen, F. (1996). *An introduction to Bayesian networks*, volume 210. UCL press London. 72

Jensenius, A. R., Wanderley, M., Godøy, R. I., and Leman, M. (2009). Musical gestures : concepts and methods in research. In *Musical gestures : Sound, Movement, and Meaning*. Rolf Inge Godoy and Marc Leman eds. 12, 101, 102

Jensenius, A. R. (2007). *Action-Sound, Developing Methods and Tools to Study Music-Related Body Movement*. PhD thesis, University of Oslo, Department of Musicology. 11, 12, 13, 154

Johnson, D. H. and Sinanović, S. (2001). Symmetrizing the kullback-leibler distance. *IEEE Trans. on Info. Theory*. 168

Jordà, S. (2005). *Digital Lutherie Crafting musical computers for new musics' performance and improvisation*. PhD thesis, University Pompeu Fabra. 16

Jordà, S. (2008). On stage : the reactable and other musical tangibles go real. *International Journal of Arts and Technology*, 1(3) :268–287. 16, 17, 19, 81, 130

Jordan, M. and Wolpert, D. (1999). Computational motor control. *The cognitive neurosciences*, 601. 9

Kannan, A. and Ostendorf, M. (1993). A comparison of trajectory and mixture modeling in segment-based word recognition. In *icassp*, pages 327–330. IEEE. 76

Kawato, M., Furukawa, K., and Suzuki, R. (1987). A hierarchical neural-network model for control and learning of voluntary movement. *Biological cybernetics*, 57(3) :169–185. 9

Kela, J., Korpipää, P., Mäntyjärvi, J., Kallio, S., Savino, G., Jozzo, L., and Marca, S. (2006). Accelerometer-based gesture control for a design environment. *Personal and Ubiquitous Computing*, 10(5) :285–299. 95

Kendon, A. (1988). How gestures can become like words. *Cross-cultural perspectives in nonverbal communication*, pages 131–141. 11

Kendon, A. (2004). *Gesture : Visible action as utterance*. Cambridge University Press. 10, 11, 13, 34, 50

Kidron, E., Schechner, Y. Y., and Elad, M. (2005). Pixels that sound. *IEEE Computer Vision & Pattern Recognition (CVPR 2005)*, 1 :88–95. 155

Kidron, E., Schechner, Y. Y., and Elad, M. (2007). Cross-modal localization via sparsity. *IEEE Trans. on Signal Processing*, 55(4) :1390–1404. 23

Kim, S. and Smyth, P. (2006). Segmental hidden markov models with random effects for waveform modeling. *The Journal of Machine Learning Research*, 7 :969. 76, 106, 113, 143

Kirk, D. (2004). *Optimal control theory : an introduction*. Dover Pubns. 9

Kita, S. and Asli, O. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal ? : Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48 :16–32. 154

Kohler, E., Keysers, C., Umilta, M., Fogassi, L., Gallese, V., and Rizzolatti, G. (2002). Hearing sounds, understanding actions : action representation in mirror neurons. *Science*, 297(5582) :846. 9, 154

Kopp, S. and Wachsmuth, I. (2004). Synthesizing multimodal utterances for conversational agents. *Computer Animation and Virtual Worlds*, 15(1) :39–52. 154

Krumhansl, C. (1989). *Why is musical timbre so hard to understand*, volume 9. Amsterdam : Elsevier. 35, 36

Kurby, C. and Zacks, J. (2008). Segmentation in the perception and memory of events. *Trends in cognitive sciences*, 12(2) :72–79. 102

Lahav, A., Saltzman, E., and Schlaug, G. (2007). Action representation of sound : audiomotor recognition network while listening to newly acquired actions. *Journal of Neuroscience*, 27(2) :308. 37

Large, E. (2000). On synchronizing movements to music. *Human Movement Science*, 19(4) :527–566. 14, 42, 154

Large, E. and Palmer, C. (2002). Perceiving temporal regularity in music. *Cognitive Science*, 26(1) :1–37. 42

Le Groux, S. and Verschure, P. (2010). Towards adaptive music generation by reinforcment learning of musical tension. In *Proceedings of the 6th Sound and Music Conference*, Barcelona, Spain. 134

Lee, M. A. and Wessel, D. (1992). Connectionist models for real-time control of synthesis and compositional algorithms. In Association, I. C. M., editor, *International Computer Music Conference*, pages 277–280. 18

Lemaitre, G., Houix, O., Misdariis, N., and Susini, P. (2010). Listener expertise and sound identification influence the categorization of environmental sounds. *Journal of Experimental Psychology : Applied*, 16(1) :16–32. 35, 42, 43, 44, 49, 50, 53

Leman, M. (2007). *Embodied Music Cognition and Mediation Technology*. Massachusetts Institute of Technology Press, Cambridge, USA. 12, 14, 18, 26, 33, 36, 37, 41, 50, 123, 153, 154, 163

Leman, M., Desmet, F., Styns, F., Van Noorden, L., and Moelants, D. (2009). Sharing musical expression through embodied listening : A case study based on chinese guqin music. *Music Perception*, 26(3) :263–278. 14, 18, 21, 24, 33, 41

Li, X., Logan, R., and Pastore, R. (1991). Perception of acoustic source characteristics : Walking sounds. *Journal of the Acoustical Society of America*. 42

Li, Y. and Shum, H. (2006). Learning dynamic audio-visual mapping with input-output hidden markov models. *Multimedia, IEEE Transactions on*, 8(3) :542–549. 164

Li, Z., Hofemann, N., Fritsch, J., and Sagerer, G. (2005). Hierarchical modeling and recognition of manipulative gesture. In *Proc. of the Workshop on Modeling People and Human Interaction at the IEEE Int. Conf. on Computer Vision*. 77

Liberman, A. and Mattingly, I. (1985). The motor theory of speech perception revised* 1. *Cognition*, 21(1) :1–36. 8, 34

Liberman, A. and Mattingly, I. (1989). A specialization for speech perception. *Science*, 243(4890) :489. 8, 34

Liu, J., Zhong, L., Wickramasuriya, J., and Vasudevan, V. (2009). uwave : Accelerometer-based personalized gesture recognition and its applications. *Pervasive and Mobile Computing*, 5(6) :657–675. 82, 92, 95, 96, 97, 98

Loehr, J. and Palmer, C. (2007). Cognitive and biomechanical influences in pianists finger tapping. *Experimental brain research*, 178(4) :518–528. 13, 50, 101

Luck, G. and Toiviainen, P. (2006). Ensemble musicians' synchronization with conductors' gestures : An automated feature-extraction analysis. *Music Perception*, 24(2) :189–200. 14, 58, 154, 163

Lutfi, R., Oh, E., Storm, E., and Alexander, J. (2005). Classification and identification of recorded and synthesized impact sounds by practiced listeners, musicians, and nonmusicians. *The Journal of the Acoustical Society of America*, 118 :393. 35

MacRitchie, J., Buck, B., and Bailey, N. (2009). Visualising musical structure through performance gesture. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*. 58, 141, 163

Maestre, E. (2009). *Modeling instrumental gestures : an analysis/synthesis framework for violin bowing*. PhD thesis, Ph. D. dissertation, Univ. Pompeu Fabra, Barcelona, Spain, 2009. 12, 101, 102

Maestre, E., Blaauw, M., Bonada, J., Guaus, E., and Pérez, A. (2010). Statistical modeling of bowing control applied to violin sound synthesis. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(4) :855–871. 13

Marcell, M., Borella, D., Greene, M., Kerr, E., and Rogers, S. (2000). Confrontation naming of environmental sounds. *Journal of Clinical and Experimental Neuropsychology*, 22(6) :830–864. 35, 42, 53

Mazzola, G. (2011). *Musical performance. A comprehensive approach : Theory, analytical tools, and case studies*. Springer. 12

McAdams, S. (1993). Recognition of sound sources and events. *Thinking in Sound : The Cognitive Psychology of Human Audition, Oxford University Press, Oxford 1993*. 35, 42

McAdams, S., Winsberg, S., Donnadieu, S., Soete, G., and Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres : Common dimensions, specificities, and latent subject classes. *Psychological research*, 58(3) :177–192. 36

McMillen, K. (1994). Zipi : Origins and motivations. *Computer Music Journal*, 18(4) :47–51. 17

McNeill, D. (1996). *Hand and mind : What gestures reveal about thought*. University of Chicago Press. 11, 54, 55

McNeill, D. (2000). *Language and gesture*, volume 2. Cambridge Univ Pr. 11

Merer, A. (2011). *Caractérisation acoustique et perceptive du mouvement évoqué par les sons pour le contrôle de la synthèse*. PhD thesis, Université de Provence – Aix-Marseille 1. 12, 15

Merer, A., Ystad, S., Kronland-Martinet, R., and Aramaki, M. (2008). Semiotics of sounds evoking motions : Categorization and acoustic features. *Computer Music Modeling and Retrieval. Sense of Sounds*, pages 139–158. 58

Merleau-Ponty, M. (1945). La phénomenologie de la perception. *Gallimard Paris*. 8

Merleau-Ponty, M. (1968). *Résumé de cours au Collège de France*. Paris, Gallimard. 8

Merrill, D. and Paradiso, J. (2005). Personalization, expressivity, and learnability of an implicit mapping strategy for physical interfaces. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, Extended Abstracts*, pages 2152–2161. 82

Metois, E. (1996). *Musical Gestures and Embedding Synthesis*. PhD thesis, Massachusetts Institute of Technology. 15

Metzinger, T. and Gallese, V. (2003). The emergence of a shared action ontology : Building blocks for a theory. *Consciousness and Cognition*, 12(4) :549–571. 154

Miall, R. and Wolpert, D. (1996). Forward models for physiological motor control. *Neural networks*, 9(8) :1265–1279. 9

Middleton, R. (1993). Popular music analysis and musicology : bridging the gap. *Popular Music*, 12(2) :177–190. 15

Misdariis, N., Minard, A., Susini, P., Lemaitre, G., McAdams, S., and Parizet, E. (2010). Environmental sound perception : meta-description and modeling based on independent primary studies. *EURASIP Journal on Audio, Speech, and Music Processing*. 59

Mitra, S. and Acharya, T. (2007). Gesture recognition : A survey. *Systems, Man, and Cybernetics, Part C : Applications and Reviews, IEEE Transactions on*, 37(3) :311–324. 81, 83

Modler, P. (2000). Neural networks for mapping hand gestures to sound synthesis parameters. *Trends in Gestural Control of Music*. 18

Moeslund, T., Hilton, A., and Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 104(2-3) :90–126. 102

Murphy, K. (1998). Switching kalman filters. *Dept. of Computer Science, University of California, Berkeley, Tech. Rep*. 78

Murphy, K. (2002). *Dynamic Bayesian Networks : Representation, Inference and Learning*. PhD thesis, UC Berkeley. 72, 109, 142

Murphy, K. and Paskin, M. (2001). Linear time inference in hierarchical hmms. In *Advances in neural information processing systems 14 : proceedings of the 2001 conference*, page 833. 77, 146

Nag, R., Wong, K., and Fallside, F. (1986). Script recognition using hidden markov models. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86.*, volume 11, pages 2071–2074. IEEE. 73

Naveda, L. (2010). *Gesture in Samba : A cross-modal analysis of dance and music from the Afro-Brazilian culture*. PhD thesis, IPEM, University of Ghent, Belgium. 12

Nelson, W. (1983). Physical principles for economies of skilled movements. *Biological Cybernetics*, 46(2) :135–147. 9

Nettl, B. (2000). An ethnomusicologist contemplates universals in musical sound and musical culture. *The origins of music*, pages 463–472. 57

Nguyen, N., Phung, D., Venkatesh, S., and Bui, H. (2005). Learning and detecting activities from movement trajectories using the hierarchical hidden markov models. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. IEEE Computer Society. 77

Noë, A. (2005). *Action in Perception*. Massachusetts Institute of Technology Press, Cambridge, USA. 7, 8, 154

Nusseck, M. and Wanderley, M. M. (2009). Music and motion - how music-related ancillary body movements contribute to the experience of music. *Music Perception*, 26 :335–353. 154

Nymoen, K., Caramiaux, B., Kozak, M., and Tørresen, J. (2011). Analyzing sound tracings - a multimodal approach to music information retrieval. In *ACM Multimedia – MIRUM 2011 (accepted)*. 39, 42, 51, 54

Nymoen, K., Glette, K., Skogstad, S., Torresen, J., and Jensenius, A. (2010). Searching for cross-individual relationships between sound and movement features using an svm classifier. In *Proceedings, New Interfaces for Musical Expression, NIME 2010 Conference*. 14, 42, 64

Oh, S., Rehg, J., Balch, T., and Dellaert, F. (2008). Learning and inferring motion patterns using parametric segmental switching linear dynamic systems. *International Journal of Computer Vision*, 77(1) :103–124. 78

O'Regan, J. (1992). Solving the" real" mysteries of visual perception : The world as an outside memory. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 46(3) :461. 8

Ostendorf, M., Digalakis, V., and Kimball, O. A. (1996). From hmms to segment models : a unified view of stochastic modeling for speech recognition. *IEEE Trans. on Speech and Audio Processing*, 4 :360–378. 75, 76, 104, 109, 143

Özyürek, A. (2010). The role of iconic gestures in production and comprehension of language : evidence from brain and behavior. In *In Embodied Communication and Human-Computer Interaction, volume 5934 of Lecture Notes in Computer Science*, pages 1–10. Springer Verlag. 34

Palmer, C., Koopmans, E., Carter, C., Loehr, J., and Wanderley, M. (2009). Synchronization of motion and timing in clarinet performance. In *International Symposium on Performance Science*, pages 1–6. 13, 14, 105

Patterson, R., Uppenkamp, S., Johnsrude, I., and Griffiths, T. (2002). The processing of temporal pitch and melody information in auditory cortex. *Neuron*, 36(4) :767–776. 36

Pavlovic, V., Rehg, J., and MacCormick, J. (2001). Learning switching linear models of human motion. *Advances in Neural Information Processing Systems*, pages 981–987. 78

Peeters, G. (2004). A large set of audio features for sound description. *CUIDADO Project*. 22, 157

Peeters, G. and Deruty, E. (2009). Sound indexing using morphological description. *IEEE Transactions on Audio, Speech and Language Processing*. 38

Pizzamiglio, L., Aprile, T., Spitoni, G., Pitzalis, S., Bates, E., D'Amico, S., and Di Russo, F. (2005). Separate neural systems for processing action-or non-action-related sounds. *Neuroimage*, 24(3) :852–861. 36, 37

Potamianos, G., Neti, C., Luettin, J., and Matthews, I. (2004). Audio-visual automatic speech recognition : an overview. *Issues in Visual and Audio-Visual Speech Processing*. 22, 23

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, pages 257–286. 83, 141, 142, 164, 165, 166, 167

Rajko, S., Qian, G., Ingalls, T., and James, J. (2007). Real-time gesture recognition with minimal training requirements and on-line learning. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE. 77

Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis*. 2nd edition, Springer Science. 49, 50

Rasamimanana, N. (2008). *Geste instrumental du violoniste en situation de jeu : analyse et modélisation*. PhD thesis, Université Pierre et Marie Curie. 12

Rasamimanana, N. (2012). Towards a conceptual framework for exploring and modeling expressive musical gestures. *Journal of New Music Research (Accepted)*. 17

Rasamimanana, N. and Bevilacqua, F. (2008). Effort-based analysis of bowing movements : evidence of anticipation effects. *Journal of New Music Research*, 37(4) :339–351. 9, 13, 101, 102

Rasamimanana, N. and Bevilacqua, F. (2012). Urban musical game. In *Proceedings of the 2012 annual conference on Human factors in computing systems (CHI2012)*. 82

Rasamimanana, N., Bevilacqua, F., Schnell, N., Guedy, F., Flety, E., Maestracci, C., Zamborlin, B., Frechin, J., and Petrevski, U. (2011). Modular musical objects towards embodied control of digital music. In *Proceedings of the fifth international conference on Tangible, embedded, and embodied interaction*, pages 9–12. ACM. 81, 82

Rasamimanana, N., Fléty, E., and Bevilacqua, F. (2006). Gesture analysis of violin bow strokes. *Gesture in Human-Computer Interaction and Simulation*, pages 145–155. 13, 157

Rasamimanana, N. H., Kaiser, F., and Bevilacqua, F. (2009). Perspectives on gesture-sound relationships informed from acoustic instrument studies. *Organised Sound*, 14(2) :208 – 216. 163

Repp, B. (1987). The sound of two hands clapping : An exploratory study. *Journal of the Acoustical Society of America*, 81(4) :1100–1109. 42

Repp, Bruno Hermann (2006). *Musical Synchronization*, pages 55–76. Music, motor control and the brain, Oxford University Press, e. altenmüller, m. wiesendanger, j. kesselring (eds.) edition. 14, 154

Rizzolatti, G. and Craighero, L. (2004). The mirror-neuron system. *Annu. Rev. Neurosci.*, 27 :169–192. 34

Röbel, A. (2003). A new approach to transient processing in the phase vocoder. In *Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx03)*, pages 344–349. Citeseer. 127

Rocchesso, D. (2011). *Explorations in Sonic Interaction Design*. COST–SID. 21

Rovan, J., Wanderley, M., Dubnov, S., and Depalle, P. (1997). Instrumental gestural mapping strategies as expressivity determinants in computer music performance. In *Proceedings of Kansei-The Technology of Emotion Workshop*, pages 3–4. Citeseer. 17

Roweis, S. (1998). Em algorithms for pca and spca. *Advances in neural information processing systems*, pages 626–632. 140

Roweis, S. and Ghahramani, Z. (1999). A unifying review of linear gaussian models. *Neural computation*, 11(2) :305–345. 146

Roweis, S. and Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290 :2323–2326. 141

Rubine, D. (1991). Specifying gestures by example. In *Proceedings of the 18th annual conference on Computer graphics and interactive techniques*, pages 329–337. ACM. 82

Russell, M. (1993). A segmental hmm for speech pattern modelling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 1993. ICASSP-93*, volume 2, pages 499–502. 75, 76

Sargin, M., Yemez, Y., Erzin, E., and Tekalp, A. (2007). Audiovisual synchronization and fusion using canonical correlation analysis. *Multimedia, IEEE Transactions on*, 9(7) :1396–1403. 23, 164

Schaal, S., Ijspeert, A., and Billard, A. (2003). Computational approaches to motor learning by imitation. *Philosophical Transactions of the Royal Society of London. Series B : Biological Sciences*, 358(1431) :537. 10, 75

Schaeffer, P. (1966). *Traité des Objets Musicaux*. Éditions du Seuil. 28, 37, 54

Scherer, Klaus R. and Ellgring, Heiner (2007). Multimodal expression of emotion : Affect programs or componential appraisal patterns ? *Emotion*, 7(1) :158–171. 153

Schnell, N., Röbel, A., Schwarz, D., Peeters, G., Borghesi, R., et al. (2009). Mubu & friends-assembling tools for content based real-time interactive audio processing in max/msp. In *Proceedings of the ICMC, Montreal*. 126

Schoner, G., Dose, M., and Engels, C. (1995). Dynamics of behavior : Theory and applications for autonomous robot architectures. *Robotics and Autonomous Systems*, 16(2-4) :213–245. 10

Schoonderwaldt, E. (2009). *Mechanics and acoustics of violin bowing : Freedom, constraints and control in performance*. PhD thesis, KTH, Royal Institute of Technology, Sweden. 12

Schoonderwaldt, E. and Demoucron, M. (2009). Extraction of bowing parameters from violin performance combining motion capture and sensors. *The Journal of the Acoustical Society of America*, 126 :2695. 13

Shafiro, V. (2008). Identification of environmental sounds with varying spectral resolution. *Ear and hearing*, 29(3) :401. 44

Silva, J. and Narayanan, S. (2006). Upper bound kullback-leibler divergence for hidden markov models with application as discrimination measure for speech recognition. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*. 165

Smalley, D. (1997). Spectromorphology : explaining sound-shapes. *Organised Sound*, 2(2) :107–126. 38

Smith, E. (1995). Concepts and categorization. *An invitation to cognitive science : Thinking*, 3. 35

Styns, Frederik, van Noorden, Leon, Moelants, Dirk, and Leman, Marc (2007). Walking on music. *Human Movement Science*, 26(5) :769–785. 14, 154, 155

Susini, P., McAdams, S., Winsberg, S., Perry, I., Vieillard, S., and Rodet, X. (2004). Characterizing the sound quality of air-conditioning noise* 1. *Applied Acoustics*, 65(8) :763–790. 26, 35, 59

Sutton, R. (1984). *Temporal credit assignment in reinforcement learning*. PhD thesis, University of Massachusetts. 134

Tardieu, J., Susini, P., Poisson, F., Kawakami, H., and McAdams, S. (2009). The design and evaluation of an auditory way-finding system in a train station. *Applied Acoustics*, 70(9) :1183–1193. 49

Taylor, G. (2009). *Composable, distributed-state models for high-dimensional time series*. PhD thesis, University of Toronto. 71

Taylor, G. and Hinton, G. (2009). Products of hidden markov models : It takes n>1 to tango. In *Proc. of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*. 74

Teixeira, E., Loureiro, M., and Yehia, H. (2010). Methodological aspects of the research in musical expressiveness based on corporal movement information. *Unpublished report. Available at* http://hal.archives-ouvertes.fr/docs/00/61/16/60/PDF/Teixeira_Loureiro_Yehia.pdf. 102

Tenenbaum, J. B., deSliva, V., and Langford, J. C. (2000). A global framework for nonlinear dimensionality reduction. *Science*, 290 :2319–2323. 141

Thierry, G., Giraud, A., and Price, C. (2003). Hemispheric dissociation in access to the human semantic system. *Neuron*, 38(3) :499–506. 36

Thompson, M. R. and Luck, G. (2008). Effect of pianists' expressive intention on amount and type of body movement. In *ICMPC 10 - Proceedings of the 10th International Conference on Music Perception and Cognition*, pages 540–544. 13

Thoresen, L. and Hedman, A. (2007). Spectromorphological analysis of sound objects : an adaptation of pierre schaeffer's typomorphology. *Organised Sound*, 12(2) :129–141. 38

Tipping, M. and Bishop, C. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 611–622. 72, 140

Todd, N. (1993). Multi-scale analysis of expressive signals : Recovery of structure and motion. In *Proceedings of the Stockholm Music Acoustics Conference*, volume 79, pages 14–69. 15

Todorov, E. (2009). Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences*, 106(28) :11478. 9

Toiviainen, P., Luck, G., and Thompson, M. (2010). Embodied meter : Hierarchical eigenmodes in music-induced movement. *Music Perception*, 28(1) :59–70. 141

Tournier, M., Wu, X., Courty, N., Arnaud, E., and Reveret, L. (2009). Motion compression using principal geodesics analysis. In *Computer Graphics Forum*, pages 355–364. Wiley Online Library. 141

Turaga, P., Chellappa, R., Subrahmanian, V., and Udrea, O. (2008). Machine recognition of human activities : A survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(11) :1473–1488. 77, 103

Turvey, M. (1977). Contrasting orientations to the theory of visual information processing. *Psychological Review*, 84(1) :67. 8

Urtasun, R., Glardon, P., Boulic, R., Thalmann, D., and Fua, P. (2004). Style-based motion synthesis. In *Computer Graphics Forum*, pages 799–812. Wiley Online Library. 141

Van Nort, D. (2009). Instrumental listening : sonic gesture as design principle. *Oganised Sound*, 14(02) :177–187. 18, 57, 163

Van Nort, D., Wanderley, M. M., and Depalle, P. (2004). On the choice of mappings based on geometric properties. In *Proceedings of the 2004 Conference on New Interfaces for Musical Expression*, Hamamatsu, Japan. NIME04. 17

VanDerveer, N. (1980). *Ecological acoustics : Human perception of environmental sounds.* PhD thesis, ProQuest Information & Learning. 34, 35, 42

Varela, F., Thompson, E., and Rosch, E. (1991). *The Embodied Mind : Cognitive Science and Human Experience*. Massachusetts Institute of Technology Press, Cambridge, USA. 8, 21, 37, 154, 163

Vermersch, P. (1990). Questionner l'action : l'entretien d'explicitation. *Psychologie française*, 35(3) :227–235. 49

Vines, B., Wanderley, M., Krumhansl, C., Nuzzo, R., and Levitin, D. (2004). Performance gestures of musicians : What structural and emotional information do they convey ? *Gesture-based communication in human-computer interaction*, pages 468–478. 12, 13, 102

Visell, Y. and Cooperstock, J. (2007a). Enabling gestural interaction by means of tracking dynamical systems models and assistive feedback. In *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*, pages 3373–3378. IEEE. 75

Visell, Y. and Cooperstock, J. (2007b). Modeling and continuous sonification of affordances for gesture-based interfaces. In *Proc. of the International Conference on Auditory Display (submitted), Montréal, Canada*. 75

Viviani, P. (1990). Motor-perceptual interactions : the evolution of an idea. *Cognitive Sciences in Europe : Issues and trends*, pages 11–39. 8

Viviani, P. and Flash, T. (1995). Minimum-jerk, two-thirds power law, and isochrony : converging approaches to movement planning. *Journal of Experimental Psychology : Human Perception and Performance*, 21(1) :32. 9, 26

Volpe, G. (2003). Computational models of expressive gesture in multimedia systems. *InfoMus Lab, DIST – University of Genova*. 12

Wandereley, M. (2001). *Performer-Instrument Interaction : Applications to Gestural Control of Sound Synthesis*. PhD thesis, Université Paris VI. 12

Wanderley, M., Schnell, N., and Rovan, J. (1998). Escher-modeling and performing composed instruments in real-time. In *Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on*, volume 2, pages 1080–1084. IEEE. 17

Wanderley, M. M. (2002). Quantitative analysis of non-obvious performer gestures. *Gesture and sign language in human-computer interaction. Springer-Verlag*, pages 241–253. 12, 13, 102, 105, 113

Wanderley, M. M. and Depalle, P. (2004). Gestural control of sound synthesis. *Proceedings of the IEEE*, 92(4) :632–644. 12

Wanderley, M. M. and Depalle, P. (2005). Gestural control of sound synthesis. *Proceedings of the IEEE*, 92(4) :632–644. 101

Wanderley, M. M., Vines, B. W., Middleton, N., McKay, C., and Hatch, W. (2005). The musical significance of clarinetists' ancillary gestures : An exploration of the field. *Journal of New Music Research*, 34(1) :97–113. 13, 14, 102, 105

Wang, J., Fleet, D., and Hertzmann, A. (2008). Gaussian process dynamical models for human motion. *IEEE transactions on pattern analysis and machine intelligence*, pages 283–298. 130

Warren, W. (2006). The dynamics of perception and action. *Psychological review*, 113(2) :358. 9

Warren, W. and Verbrugge, R. (1984). Auditory perception of breaking and bouncing events : A case study in ecological acoustics. *Journal of Experimental Psychology : Human Perception and Performance*, 10(5) :704. 34

Watkins, K., Strafella, A., and Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, 41(8) :989–994. 34

Wessel, D. and Wright, M. (2002). Problems and prospects for intimate musical control of computers. *Computer Music Journal*, 26(3) :11–22. 16, 17

Widmer, G., Dixon, S., Goebl, W., Pampalk, E., and Tobudic, A. (2003). In search of the horowitz factor. *AI Magazine*, 24(3) :111. 101, 103

Widmer, G. and Goebl, W. (2004). Computational models of expressive music performance : The state of the art. *Journal of New Music Research*, 33(3) :203–216. 101

Williamson, J. and Murray-Smith, R. (2005). Sonification of probabilistic feedback through granular synthesis. *Multimedia, IEEE*, 12(2) :45–52. 75

Wilson, A. and Bobick, A. (1999). Parametric hidden markov models for gesture recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(9) :884–900. 74, 83

Wilson, A. and Bobick, A. (2000). Realtime online adaptive gesture recognition. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 1, pages 270–275. IEEE. 83

Wilson, A. D. (2000). *Adaptive Models for Gesture Recognition*. PhD thesis, Massachusetts Institute of Technology. 83

Wobbrock, J., Wilson, A., and Li, Y. (2007). Gestures without libraries, toolkits or training : a $1 recognizer for user interface prototypes. In *Proceedings of the 20th annual ACM symposium on User interface software and technology*, pages 159–168. ACM. 82, 92, 93, 94, 99

Wolpert, D. and Ghahramani, Z. (2000). Computational principles of movement neuroscience. *nature neuroscience*, 3 :1212–1217. 9

Wundt, W. (1973). The language of gestures. *Mouton, The Hague*. 10

Xu, J., Gannon, P., Emmorey, K., Smith, J., and Braun, A. (2009). Symbolic gestures and spoken language are processed by a common neural system. *Proceedings of the National Academy of Sciences*, 106(49) :20664. 34

Yamato, J., Ohya, J., and Ishii, K. (1992). Recognizing human action in time-sequential images using hidden markov model. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on*, pages 379–385. IEEE. 73, 141

Yankov, D., Keogh, E., Medina, J., Chiu, B., and Zordan, V. (2007). Detecting time series motifs under uniform scaling. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 844–853. ACM. 118

Yu, S. (2010). Hidden semi-markov models. *Artificial Intelligence*, 174(2) :215–243. 74, 143

Zatorre, R., Chen, J., and Penhune, V. (2007). When the brain plays music : auditory–motor interactions in music perception and production. *Nature Reviews Neuroscience*, 8(7) :547–558. 18, 36, 41, 42

Zhou, S., Chellappa, R., and Moghaddam, B. (2004). Visual tracking and recognition using appearance-adaptive models in particle filters. *Image Processing, IEEE Transactions on*, 13(11) :1491–1506. 83

Zue, V., Glass, J., Philips, M., and Seneff, S. (1989). Acoustic segmentation and phonetic classification in the summit system. In *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, pages 389–392. IEEE. 76