Perspectives pour l'apprentissage interactif du couplage geste-son

Jules Françoise, Ianis Lallemand, Thierry Artières, Frédéric Bevilacqua, Norbert Schnell, Diemo Schwarz

PERSPECTIVES POUR L'APPRENTISSAGE INTERACTIF DU COUPLAGE GESTE–SON

Jules Françoise
Ircam — UMR STMS
IRCAM–CNRS–UPMC
jules.francoise@ircam.fr

Ianis Lallemand
Ircam — UMR STMS
IRCAM–CNRS–UPMC
ianis.lallemand@ircam.fr

Thierry Artières
LIP6
CNRS–UPMC
thierry.artieres@lip6.fr

Frédéric Bevilacqua
Ircam — UMR STMS
IRCAM–CNRS–UPMC
frederic.bevilacqua@ircam.fr

Norbert Schnell
Ircam — UMR STMS
IRCAM–CNRS–UPMC
norbert.schnell@ircam.fr

Diemo Schwarz
Ircam — UMR STMS
IRCAM–CNRS–UPMC
diemo.schwarz@ircam.fr

RÉSUMÉ

L'apprentissage de mappings du geste vers le son constitue aujourd'hui un enjeu de recherche majeur. Dans un travail précédent, nous avons proposé un modèle hiérarchique permettant de modéliser des structures temporelles à différentes échelles [8]. Nous nous intéressons ici à l'apprentissage de structures temporelles de plus haut niveau. Plus spécifiquement, nous nous proposons de formuler la problématique de l'*articulation* entre différents mappings geste–son dans un contexte d'apprentissage interactif. Ce champ de recherche émergent, aux croisements de l'apprentissage automatique et de l'interaction homme-machine, permet à notre sens de poser correctement la question de l'apprentissage « par démonstration ». Nous présentons d'abord successivement les cadres de l'apprentissage interactif et de la modélisation du couplage geste–son, puis les perspectives ouvertes par la réunion de ces problématiques, ainsi qu'une première extension de nos travaux précédents dans ce cadre.

1. INTRODUCTION : SYSTÈMES INTERACTIFS MUSICAUX BASÉS SUR LE GESTE

1.1. Contexte

Nous considérons dans cet article des dispositifs interactifs musicaux impliquant une interaction gestuelle expressive avec des processus sonores. Ces systèmes sont fréquemment utilisés dans les domaines artistique et scientifique : installations sonores interactives, contrôle temps-réel pour la performance scénique ou nouvelles interfaces pour l'expression musicale [1] . De manière générale, ils s'articulent autour de trois composantes principales, représentées sur la figure 1 : les gestes de l'utilisateur [2] sont d'abord captés et analysés (1), puis couplés (2) à un dispositif de synthèse sonore (3). Cette opération de couplage, souvent appelée *mapping*, est un élément crucial qui conditionne les possibilités d'interaction avec le dispositif sonore. Des choix différents lors de la création du mapping influenceront fortement l'expressivité du système, sa facilité d'utilisation ou ses possibilités d'exploration [13].

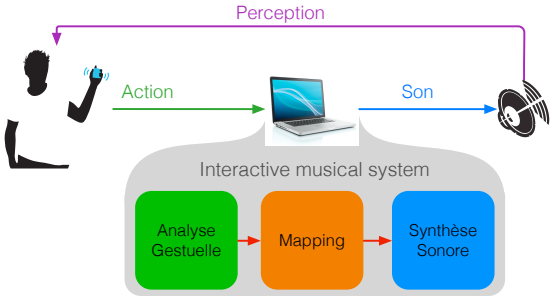


Figure 1: Schéma général d'un système interactif musical.

Souvent, ce mapping est construit par une formulation analytique, réalisée par exemple par la création de liens explicites entre paramètres gestuels et sonores.

1.1.1. « Mapping par démonstration »

Nous considérons ici un cadre alternatif qui s'inscrit dans la continuité des recherches menées au sein de l'équipe Interactions Musicales Temps-Réel (IMTR) de l'Ircam. Appelée « mapping par démonstration », notre démarche consiste à définir le couplage geste–son par l'interaction elle-même en fournissant des exemples de réalisations concrètes du mapping [3] . Ceci induit une boucle d'interaction composée de deux phases :

1 . Voir en particulier la conférence internationale NIME (*New Interfaces for Musical Expression*) : `http://www.nime.org/`.

2 . Dans cet article, nous nommons « utilisateur » toute personne interagissant avec un système interactif, sans considération pour les problèmes esthétiques que pourrait poser une telle notion d'« utilisation », comprise dans un sens littéral.

3 . Cette approche se rapproche des notions de « mapping par l'écoute » développés dans la thèse de Baptiste Caramiaux [3], et du « *play-along mapping* » introduit par Fiebrink [6].

1. **Apprentissage**. Dans un premier temps, l'utilisateur « joue » un ensemble d'exemples sonores : simultanément à l'écoute de ceux-ci, il réalise des gestes qu'il souhaite associer aux sons entendus, afin de fournir au système des exemples de réalisations du mapping à apprendre. Le système doit ensuite abstraire le couplage depuis les données d'exemple, par le biais de méthodes d'apprentissage automatique (cf. fig. 2a).
2. **Performance**. Dans un second temps, l'utilisateur peut réinterpréter les exemples sonores de départ en réalisant de nouveau les gestes fournis au système au cours de la première phase. Le couplage geste–son appris par le système permet de traduire les modulations des gestes réinterprétés en modulations sonores (cf. fig. 2b).

1.1.2. Travaux récents

Dans ce contexte de mapping par démonstration, nous nous sommes plus particulièrement intéressés à la réinterprétation des sons enregistrés. Les techniques musicales permettant d'aborder la question du *jeu* fondé sur un matériau sonore concret sont nombreuses : *cut-up*, compression ou dilatation temporelle, réalisation de boucles, ...

Nous avons récemment réalisé plusieurs systèmes utilisant le système de suivi de geste *Gesture Follower* [4] afin de combiner de manière expressive ces éléments de vocabulaire musical [1]. En particulier, cette technologie permet de capturer avec précision les dynamiques temporelles de gestes associés à des sons. Pendant la phase d'apprentissage, l'utilisateur effectue un geste aligné à un exemple sonore. Il peut ensuite exécuter de nouveau le geste qui, aligné en temps réel à la référence, permet la réinterprétation du son. Les variations temporelles gestuelles son alors traduite en variations sonores par des opération de time–stretching utilisant un vocodeur de phase.

Cette approche a été récemment étendue par une modélisation hiérarchique des gestes, détaillée en 3.2. Ceci permet le développement de représentations gestuelles plus complexes, autorisant par exemple l'ajout de contraintes sur certaines phases du son (typiquement, les phases d'attaque peuvent ainsi être conservées) [5].

1.1.3. Réalisations musicales

Ces approches développées dans l'équipe IMTR ont donné lieu à des réalisations à différentes échelles, du dispositif instrumental à l'installation interactive. Citons par exemple :

- Les *Modular Musical Objects* (MO) [6], des interface gestuelles portables permettant de construire de nombreux scénarios d'interaction musicale par leur intégration dans des objets du quotidien [18].
- *Urban Musical Game*, un jeu sonore interactif utilisant un ballon augmenté [7].

1.2. Enjeux scientifiques

De nombreux enjeux scientifiques résultent des problématiques posées par les modes d'expérience spécifiques aux systèmes interactifs musicaux. Ces enjeux sont liés à des questions de recherche difficiles dans le champ de l'apprentissage automatique.

1.2.1. Faible nombre d'exemples

Dans le cas d'un système interactif musical basé sur le couplage entre geste et son, les « règles » conditionnant la création des contenus sonores sont inférées uniquement à partir de données gestuelles et sonores fournies par l'utilisateur. Cette caractéristique limite nécessairement le nombre d'exemples d'apprentissage disponibles : en pratique, de tels systèmes doivent pouvoir être appris à partir d'un petit nombre d'exemples, voire d'un seul exemple (enjeu 1, section 1.2.4).

1.2.2. Apprentissage et interaction simultanés

Dans le cadre des systèmes interactifs musicaux de « mapping par démonstration » que nous considérons ici, les données d'apprentissage (initiales ou « d'adaptation », cf. 1.1.1) sont fournies par les gestes mêmes de l'utilisateur, c'est-à-dire par son interaction avec le dispositif. En d'autres termes, l'apprentissage ne peut pas être distingué du fonctionnement « réel » du système. Il se déroule simultanément à l'interaction des utilisateurs avec le dispositif (enjeu 2, section 1.2.4).

1.2.3. Adaptation de l'apprentissage à l'utilisateur

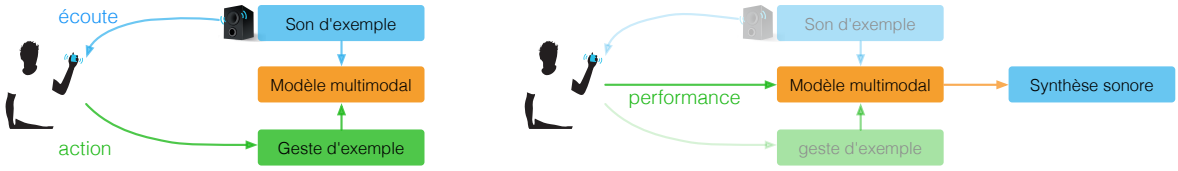
La prise en compte des attentes des utilisateurs lors de l'apprentissage constitue un critère fondamental pour des usages inscrits dans un domaine artistique. Il est en effet souvent très difficile de formuler explicitement des critères d'évaluation de la qualité « artistique » ou « musicale » des résultats produits par un système interactif, dans la mesure où les attentes à l'égard de ceux-ci sont fortement dépendantes du contexte et, par extension, du public. L'interaction avec le dispositif, en impliquant activement l'utilisateur, propose de reformuler la question de l'apprentissage d'un certain critère de qualité (et donc, incidemment, de la qualité de l'apprentissage) en termes d'apprentissage par l'interaction avec l'utilisateur (enjeu 3, section 1.2.4).

1.2.4. Résumé des enjeux scientifiques

Nous avons ainsi pu mettre en évidence trois enjeux scientifiques majeurs :

Enjeu 1 : Apprendre à partir de très peu d'exemples

4. Voir des exemples d'utilisation de *Gesture Follower* : `http://imtr.ircam.fr/imtr/Gesture_Follower`
5. Voir une demonstration du système : `http://vimeo.com/julesfrancoise/smc2012hierarchicalmapping`
6. `http://www.youtube.com/watch?v=Uhps_U2E9OM`
7. `http://www.youtube.com/watch?v=jXGlvmrGBgY`



(a) Phase d'apprentissage.

(b) Phase de performance.

Figure 2: Apprentissage d'un modèle geste–son par « mapping par démonstration ».

Enjeu 2 : Apprendre à partir de données acquises pendant l'interaction

Enjeu 3 : Adapter interactivement l'apprentissage à l'utilisateur

Ces problématiques, prises indépendamment, constituent des champs disciplinaires vastes. Elles sont ici abordées dans la perspective transversale de l'apprentissage automatique interactif (*interactive machine learning*).

1.3. Problématique et aperçu

Situé dans la continuité de nos travaux actuels, cet article vise à proposer un ensemble de perspectives sur le problème de l'apprentissage du couplage geste–son à la lumière des développements récents du domaine de l'apprentissage automatique interactif. Ce champ de recherche émergent est introduit dans la section 2 ; nous y présentons l'état de l'art de la recherche actuelle, et établissons un lien avec les problématiques intrinsèques au contexte musical. Synthétisant des recherches récentes menées au sein de l'équipe IMTR, la section 3 présente un cadre de modélisation du couplage geste–son centré sur les aspects temporels. Enfin, la section 4 présente un ensemble de perspectives d'extension des modèles présentés dans la section 3 au cadre de l'apprentissage interactif.

2. APPRENTISSAGE INTERACTIF

2.1. Définition

L'apprentissage interactif est l'objet d'un intérêt récent dans le domaine de l'apprentissage automatique [8]. Il se positionne généralement à l'interface des disciplines de l'apprentissage automatique et de l'interaction Homme–machine, et se distingue du contexte de l'apprentissage automatique « traditionnel » en ce sens qu'il entend intégrer explicitement la nature *humaine* de l'utilisateur dans le processus d'apprentissage.

8. Voir par exemple les ateliers :
- https://sites.google.com/site/iui2013imlw/
- http://www.cs.utexas.edu/~bradknox/AAAIFSS-RLIHT12/RLIHT__2012_AAAI_Fall_Symposium.html
- http://research.microsoft.com/en-us/um/people/sumitb/adaiml09/

2.2. Approche transversale des problématiques

Certaines exigences imposées par les contextes artistique et musical inscrivent de fait les enjeux scientifiques présentés en 1.2 dans une perspective transversale. Ces problématiques nous semblent pouvoir être étudiées de manière pertinente dans le cadre de l'apprentissage interactif, lui-même situé à l'interface entre plusieurs domaines de recherche.

2.2.1. Exigence de généralisation

La question de la généralisation d'un apprentissage réalisé à partir de peu d'exemples se pose ici en termes spécifiques : il ne s'agit pas de construire à partir d'un seul exemple une représentation *fixe* des données d'entrée, permettant la classification ultérieure de données similaires. L'enjeu réside plutôt dans la capacité d'apprendre à générer des variations alternatives du matériau de départ, en interaction continue avec l'utilisateur. Le cadre d'apprentissage interactif mêle donc les problèmes de généralisation et d'apprentissage interactif, dans une perspective de *généralisation spécifique à un utilisateur* basée sur l'interaction.

2.2.2. Exigence de simultanéité des phases d'apprentissage et d'interaction

Apprendre par et pendant l'interaction avec un utilisateur établit une identité entre la durée d'interaction avec le dispositif et le nombre d'exemples à partir desquels ce dernier peut être appris. Dans le cadre étudié ici, la question de l'apprentissage à partir de peu d'exemples croise donc celle de l'apprentissage à partir de données d'interaction, dans la perspective commune d'un apprentissage réalisé en une durée suffisamment courte (c'est-à-dire adaptée aux durées typiques d'interaction avec des dispositifs artistiques).

2.3. Façonnage interactif

Nous nous intéressons plus particulièrement à une classe spécifique de méthodes, désignées par le terme de « façonnage interactif » (*interactive shaping*).

Le façonnage interactif, récemment formalisé par W. Bradley Knox dans sa thèse de doctorat [14], constitue un cadre permettant de traiter des situations d'apprentissage

dans lesquelles un instructeur humain évalue le comportement d'un système informatique de manière interactive. Dans ce contexte, l'évaluation humaine se compose de « récompenses » scalaires, dont le caractère numérique positif ou négatif correspond au caractère positif ou négatif de l'appréciation.

2.3.1. Relations avec l'apprentissage par renforcement

Cette notion de récompense s'inspire directement de celle utilisée dans le domaine de l'apprentissage par renforcement (*reinforcement learning*) [19]. Les méthodes d'apprentissage par renforcement fournissent un cadre permettant l'apprentissage d'un système à partir de récompenses attribuées par son *environnement*. Dans la mesure où le terme d'environnement se réfère à tout élément extérieur avec lesquel le système est susceptible d'interagir, l'utilisateur d'un dispositif interactif geste–son peut être inclus dans cette notion. Cependant, la formulation « classique » de l'apprentissage par renforcement [17] ne prévoit pas que l'environnement récompense directement le système. Au sens propre, les récompenses ne sont en effet pas reçues par le système, mais calculées par le système à partir d'une spécification interne des issues possibles de son interaction avec l'environnement. Considéré comme l'« environnement » d'un système geste–son, l'utilisateur n'attribuerait donc pas de récompense de manière interactive : une spécification « objective » des attentes de celui-ci devrait être fournie préalablement au système. Comme évoqué en 1.2, cette contrainte est fortement problématique dans le domaine artistique, où la notion de qualité peut varier énormément selon le public et le contexte.

2.3.2. Intégration d'un instructeur humain : spécificités du façonnage interactif

Les approches de façonnage interactif [14] entendent contourner cette difficulté. À la différence des méthodes « classiques » d'apprentissage par renforcement, elles modélisent explicitement la nature humaine de l'instructeur et considèrent des récompenses attribuées de manière interactive. Elles sont ainsi particulièrement adaptées aux cas où une spécification explicite et *a priori* des attentes de l'utilisateur par rapport au système serait difficile, voire impossible (si l'on s'attend par exemple à ce que des utilisateurs différents formulent des attentes différentes).

La prise en compte de la dimension humaine de l'instructeur est à l'origine d'une autre spécificité des méthodes de façonnage interactif par rapport aux approches d'apprentissage par renforcement. Si la majeure partie de ces dernières calculent à tout moment une prédiction des récompenses attendues sur un horizon de temps futur, la plupart des méthodes de façonnage interactif sont « myopes » [14] : elles tiennent peu ou aucunement compte d'une prédiction de l'impact futur des récompenses attribuées. Lorsque l'instructeur humain attribue une récompense, le modèle compare cette récompense à celle qu'il attendait pour l'instant actuel uniquement. Les

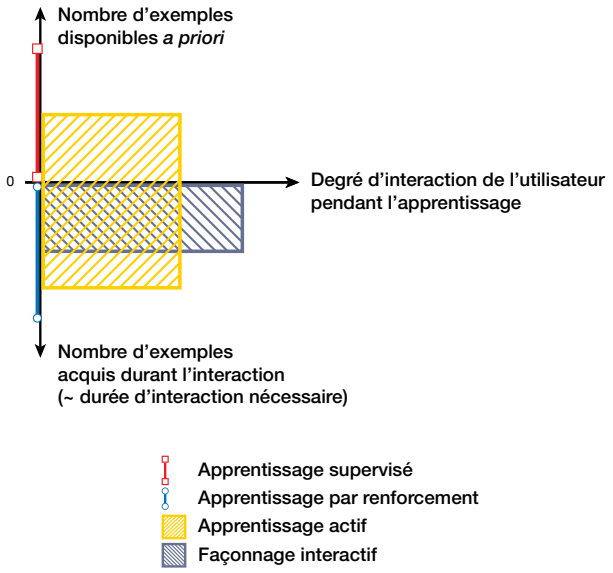


Figure 3: Relations du façonnage interactif à d'autres champs de recherche.

méthodes de façonnage interactif entendent ainsi trouver un compromis entre leurs capacités de prévision et leur réactivité.

Ces deux spécificités rendent les approches de façonnage interactif particulièrement intéressantes au regard des exigences des contextes artistiques.

2.3.3. Résumé : intérêt du façonnage interactif pour l'apprentissage du couplage geste–son

La figure 3 propose une comparaison avec d'autres approches d'apprentissage automatique, comme l'apprentissage supervisé [16] et l'apprentissage par renforcement. Les approches de façonnage interactif y sont représentées comme impliquant le plus l'utilisateur, et requérant le moins d'exemples (tant *a priori* que collectés pendant l'interaction) pour leur apprentissage. Ces spécificités du façonnage interactif nous semblent particulièrement intéressantes dans le contexte de l'apprentissage du couplage geste–son, détaillé en 3. Nous verrons en particulier en 4.1 comment l'intégration explicite des réactions d'appréciation positive ou négative de l'utilisateur permettent de formuler de manière pertinente le problème de l'apprentissage de structures de haut niveau, difficile à poser dans ce cadre.

3. MODÉLISATION DU COUPLAGE GESTE–SON

Cette partie vise à synthétiser des développements récents menés dans l'équipe IMTR dans le domaine de la modélisation statistique de mappings liant geste et son. Nous nous plaçons donc dans un cadre de « mapping par démonstration ». Cette situation implique la réalisation simultanée à l'écoute d'une performance gestuelle, qui rend prépondérante la dynamique temporelle du son. Nous nous intéressons donc ici à chaque exemple — sonore ou

gestuel — en tant que *morphologie temporelle* de courte durée (typiquement inférieure à quelques secondes). Dans la suite, nous désignons par « segment geste–son » la réunion de deux profils temporels gestuel et sonore.

Nous décrivons dans un premier temps une méthode d'apprentissage du couplage temporel entre geste et son, puis présentons une extension hiérarchique récente.

3.1. Modèles statistiques de segments geste–son

Dans *Gesture Follower* (voir 1.1.2), le système modélise un geste d'exemple par un modèle de Markov caché (*hidden Markov model* ou HMM) de topologie gauche-droite, qui permet d'encoder sa temporalité de façon continue et flexible. Un mapping temporel est alors formulé en alignant les morphologies gestuelle et sonore [2]. Le geste peut ensuite être rejoué en faisant varier expressivement son déroulement temporel. À chaque instant, un calcul d'alignement permet d'estimer l'avancement temporel à l'intérieur du geste, ce qui permet une resynthèse adaptative du son associé par des opérations d'étirement ou de compression temporelle basées sur des techniques de vocodeur de phase (cf. fig. 4).

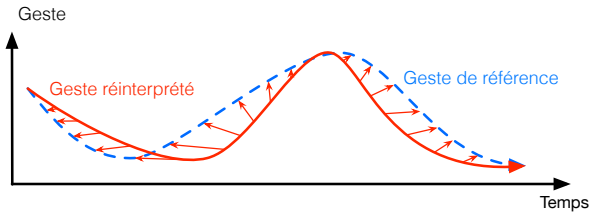


Figure 4: Fonctionnement de *Gesture Follower*. Le geste de référence (bleu pointillé) est modélisé par un modèle de markov caché. En temps réel, une nouvelle interprétation du geste (rouge plein) peut être alignée sur la référence afin de rejouer dynamiquement le son associé.

3.2. Du segment à la séquence : extension hiérarchique

3.2.1. Motivation

Le modèle présenté dans la section précédente fournit une modélisation fine de la structure temporelle de segments geste–son de courte durée. Cependant, afin d'élargir les possibilités d'interaction, ces segments nécessitent d'être combinés pour former des séquences plus complexes. Les séquences musicales peuvent souvent être décomposées de manière hiérarchique d'un point de vue temporel, par exemple en phases d'attaque/sustain, notes, phrase, etc. Il semble donc cohérent d'appliquer une décomposition similaire au couplage geste–son.

3.2.2. Modélisation hiérarchique

Nous avons récemment proposé une approche hiérarchique pour la création de mappings temporels [8], représentée sur la figure 5. L'approche consiste à appliquer au geste une segmentation multi-échelles, afin de structurer hiérarchiquement sa représentation temporelle **(1)**. Cette structure est alors mise en relation avec une représentation analogue des processus sonores **(2)**, permettant le développement de stratégies d'interaction complexes **(3)**.

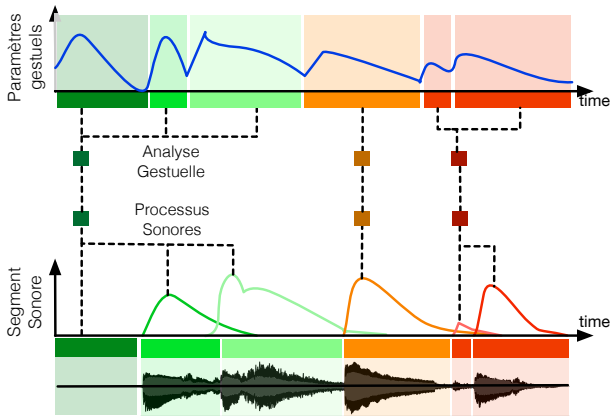


Figure 5: Représentation hiérarchique des relations geste–son. Chaque modalité est structurée hiérarchiquement par une segmentation multi-échelles, qui permet la création de mappings selon différentes échelles temporelles.

Formellement, le modèle introduit dans la section précédente peut être étendu à une structure multi-échelles, sous la forme d'un modèle de Markov caché hiérarchique [7]. Les segments geste–son appris sont alors liés les uns aux autres par une structure de haut niveau, dont les probabilités de transition peuvent être fixées par l'utilisateur.

3.2.3. PASR : Préparation — Attaque — Sustain — Relâchement

Afin d'illustrer l'intérêt de telles structures multi-échelles, nous avons présenté dans des travaux précédents une décomposition gestuelle spécifique appliquée au contrôle de la synthèse sonore [8]. Nous reprenons ici cette représentation appelée PASR (par analogie avec la représentation sonore ADSR) qui décompose chaque geste en quatre phases. Une phase de préparation (**P**) est associée au geste d'anticipation précédant le début du son. Suivent des phases d'attaque (**A**) et sustain (**S**). Enfin, une phase de relâchement (**R**) décrit le geste de rétraction accompagnant la fin du son. La figure 6 détaille la segmentation d'un geste selon cette représentation et présente la topologie du modèle associé. Les flèches indiquent les transitions possibles : un utilisateur peut donc entrer dans un geste par les phases de préparation ou d'attaque. Le parcours se poursuit en suivant l'axe temporel, et le geste peut être achevé par une phase de sustain ou de relâchement.

Cette représentation offre plusieurs avantages. D'une part, elle rend possibles différents modes de jeu : les gestes peuvent être entièrement rejoués ou rapidement enchainés par des transitions des phases de sustain aux phases d'attaque. D'autre part, l'apport d'une structure de haut ni-

veau permet une précision accrue de la segmentation du geste en termes de reconnaissance et de précision temporelle. Enfin, elle permet l'imposition de contraintes sur certaines phases du son lors de la synthèse. dans une situation typique où le son est réinterprété temporellement en utilisant un vocodeur de phase (alignement du son à la performance gestuelle), il est possible d'imposer une conservation des transitoires sur les phases d'attaque, ce qui garantit une meilleure cohérence du rendu sonore.

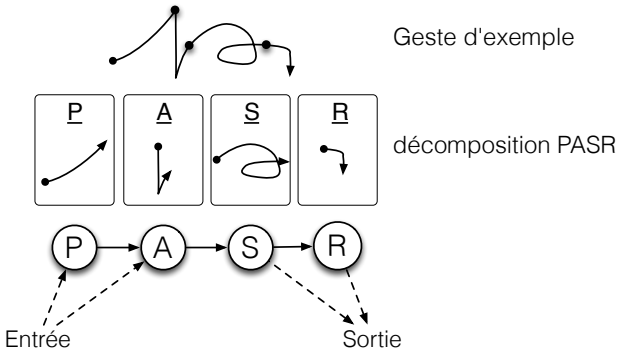


Figure 6: Décomposition PASR d'un geste. Les flèches entre les états **P**, **A**, **S** et **R** indiquent les possibilités de transition au sein du modèle.

3.2.4. Une question ouverte : la co-articulation

Dans la représentation hiérarchique actuelle, un geste est modélisé par une séquences de segments, une représentation qui semble pertinente d'un point de vue cognitif [11]. Cependant, le modèle est limité par deux aspects. Sa construction est soumise à une segmentation manuelle de l'utilisateur, qui ne peut définir des segments ou modèles de séquences que manuellement. D'autre part, le modèle ne laisse que peu d'espace à d'éventuelles variations à l'interface entre deux segments successifs. Ainsi, les phénomènes de *co-articulation* pouvant survenir entre des segments « significatifs » pendant la performance d'une nouvelle séquence seront difficilement caractérisés par le modèle actuel.

Considérons en effet le cas d'un système interactif composé de deux exemples de mapping geste–son, modélisés par deux représentations PASR. Dans la version actuelle du modèle, la seule « articulation » que l'on puisse concevoir entre ces deux représentations est détaillée dans la figure 7. Les états de fin (**S1**, **R1**) de la première représentation PASR1 sont connectés aux états initiaux (**P2**, **A2**) de la seconde représentation PASR2. Il est donc uniquement possible de réaliser ces deux gestes « à la suite », c'est-à-dire de naviguer de la fin d'un geste (au sortir de la phase de sustain **S1** ou de relâchement **R1**) vers le début de l'autre (entrée dans la phase de préparation **P2** ou d'attaque **A2**). Dans l'optique de dépasser de ces contraintes d'articulation, nous détaillons dans la section suivante un programme prospectif d'apprentissage de phases de *co-articulation* entre différents modèles PASR. Il s'agit de développer un système capable d'interpréter les gestes

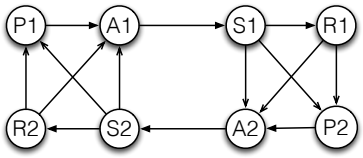


Figure 7: Topologie d'une structure comprenant deux modèles de couplage PASR.

d'un utilisateur souhaitant augmenter le système avec des mouvements de co-articulation, en se basant uniquement sur ses gestes effectués durant l'interaction.

4. PROSPECTIVE : APPRENTISSAGE DU COUPLAGE GESTE–SON PAR FAÇONNAGE INTERACTIF

4.1. Problématique

Nous nous proposons d'exploiter le paradigme du façonnage interactif exposé dans la section 2.3 afin d'apprendre incrémentalement des modalités de co-articulation entre plusieurs représentations hiérarchiques de type PASR, telles qu'introduites en section 3.2. Cette approche consiste à interconnecter des modèles PASR appris pour différents exemples de couplage geste–son. Dans la suite de notre explication, nous considérons une situation d'exemple composée de deux modèles PASR. La topologie de la structure de haut niveau réunissant ces deux modèles est initialement la même que celle présentée dans la figure 7.

Comme expliqué en 3.2.4, ces possibilités de transition correspondent simplement à l'enchaînement de plusieurs actions distinctes. Nous nous proposons ici de laisser la possibilité à l'utilisateur d'étendre cette structure de base de manière incrémentale et interactive, en ajoutant une dimension de *co-articulation* entre paires de modèles PASR. Formellement, l'enjeu est de construire un « pont » supplémentaires entre deux modèles par l'ajout d'un état **C** connectant, par exemple, l'état **A1** du premier modèle à l'état **S2** du second (figure 8). L'état **C** correspond à une phase de haut niveau, au même titre que les phases de type **P, A, S** et **R**. Il génère un segment modélisant la co-articulation entre les phases **A1** et **S2**.

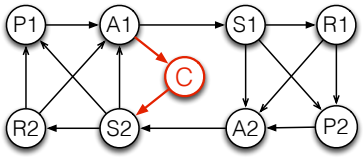


Figure 8: Ajout d'un état de *co-articulation* dans la topologie.

Cette inférence d'une nouvelle configuration d'une topologie de haut niveau relève d'une problématique d'*apprentissage de structure de modèles statistiques*. Il s'agit d'un domaine actif de l'apprentissage automatique [16] :

usuellement traité dans un cadre classique d'apprentissage supervisé où de nombreux exemples du matériau à modéliser sont disponibles, il constitue alors un problème extrêment difficile. Nous envisageons ici de tirer parti de la dimension interactive intrinsèque à nos contextes applicatifs pour proposer une formulation de ce problème en termes d'*apprentissage interactif* : cette approche nous semble poser le problème de manière plus cohérente avec le cadre d'étude des systèmes interactifs musicaux.

4.2. Proposition d'approche pour l'apprentissage incrémental et interactif de l'état de co-articulation C

Nous souhaitons permettre à l'utilisateur de « forcer » l'apprentissage d'une telle co-articulation par le seul moyen de l'interaction gestuelle. La difficulté de cette approche est double :

1. Il s'agit d'abord de détecter les gestes réalisés par l'utilisateur qui imposent effectivement au modèle une transition entre **A1** et **S2**. Ces gestes doivent être distingués efficacement de possibles variations interprétatives, s'éloignant du geste modélisé par la structure PASR de départ sans pour autant être investies d'une intention de modification de cette structure (i.e., l'utilisateur souhaite pouvoir rejouer différemment le mapping de départ, sans souhaiter toutefois articuler deux mappings distincts).
2. Une fois la phase de co-articulation détectée, celle-ci doit être apprise de manière incrémentale, à mesure que l'utilisateur réalise de nouveaux gestes dans lesquels elle apparaît (fournissant ainsi de nouveaux exemples d'apprentissages). Le système réalise en ce sens une adaptation de sa représentation interne du couplage (telle qu'évoquée en 1.1.1).

4.2.1. Formalisation

Notre formulation entend traiter ces deux étapes simultanément, dans un cadre d'apprentissage par façonnement. Pour ce faire, une fois les différents modèles PASR appris séparément, nous initialisons un tableau de probabilités de transitions potentielles, entre les états du modèle PASR1 et l'état **C**, et entre l'état **C** et le modèle PASR2. Ces probabilités sont au départ égales à 0 ; de même, l'état **C** est initialement purement hypothétique, puisqu'aucun exemple de co-articulation n'a encore été fourni par l'utilisateur.

Durant la phase d'interaction, le système détecte les « erreurs », c'est à dire les transitions théoriquement « interdites » qui peuvent être empruntées sous la contrainte des mouvements de l'utilisateur. La détection de telles erreurs est un bon indicateur de l'introduction d'un mouvement de co-articulation entre deux segments.

Supposons qu'une co-articulation entre **A1** et **S2** soit ainsi détectée. On considère alors que l'utilisateur *récompense* la transition de **A1** vers **C**. La probabilité de cette transition, initialement nulle, est mise à jour pour intégrer cette récompense. Parallèlement, la phase identifiée comme relevant de la co-articulation dans le geste de l'utilisateur permet d'apprendre un premier modèle du segment **C**. L'état **C** est pour l'instant construit « à l'écart » : il n'est pas encore considéré comme une co-articulation effective entre les modèles PASR, mais comme une co-articulation *potentielle*.

À mesure que l'utilisateur continue de récompenser cette co-articulation par son interaction avec le système, l'algorithme de façonnage interactif donne de plus en plus de poids au modèle « potentiel » intégrant l'état de co-articulation **C**. Lorsque ce poids atteint une valeur suffisante, l'état **C** est effectivement intégré comme lien de co-articulation entre les modèles PASR1 et PASR2. La modalité sonore correspondante est alors évaluée à partir de **A1** et **S2**, à l'aide d'un algorithme d'interpolation audio [9] .

4.2.2. Implémentation

Nous disposons actuellement d'une implémentation des modèles de Markov hiérarchiques, sous la forme d'un object Max/MSP. Celui-ci permet l'enregistrement de profils gestuels liés à des segments sonores, et intègre les algorithmes temps-réel de suivi. Le modèle peut donc être utilisé pour le contrôle sonore, par exemple en utilisant des techniques de vocodeur de phase. Nous développons actuellement en parallèle les deux composantes de la dimension d'apprentissage interactif : la méthode de détection de transitions relevant d'un aspect de co-articulation, ainsi que l'algorithme de façonnage.

4.3. Perspectives

Les modèles actuels possèdent des limitations quand à l'estimation de la dimension sonore. En particulier, la modélisation de type suivi de geste fournit une représentation statistique du geste mais non du son. Ceci peut limiter les possibilités d'apprentissage des variations du mapping apparues sur différentes performances.

Pour dépasser cette limitation, nous étudions actuellement une extension multimodale des modèles de Markov cachés pour l'apprentissage du couplage geste–son. Cette approche s'inspire de travaux réalisés en animation d'avatars [10] et en inversion acoustique–articulatoire [12]. Cette fois, les processus gestuels et sonores sont modélisés conjointement. Un algorithme d'apprentissage multi-exemples permet d'entraîner le modèle sur plusieurs variations d'un même segment geste–son. Ce modèle multimodal peut ensuite être inversé pour le contrôle sonore : une nouvelle performance gestuelle permet la synthèse des paramètres sonores en temps réel. De par la modélisation statistique conjointe, l'estimation des paramètres de contrôle sonore est réalisée de manière probabiliste, ce qui induit une meilleure fluidité de la synthèse. Ce modèle

9 . Rappelons que dans le cadre présenté dans cette section, seule la dimension gestuelle est modélisée de manière probabiliste. L'estimation de la modalité sonore constitue une question ouverte, qui fera l'objet d'un travail ultérieur (esquissé en 4.3).

fournit une perspective intéressante pour l'estimation de l'aspect sonore de segments de co-articulation gestuelle. D'une part, la capacité d'apprentissage multi-exemples permet une meilleure approximation des mouvements de co-articulation gestuelle. D'autre part, le modèle pourrait utiliser le contexte de la co-articulation gestuelle comme premier estimateur de la co-articulation sonore correspondante. Par apprentissage incrémental, ce modèle pourrait ensuite être affiné par les exemples successifs fournis par l'utilisateur.

5. CONCLUSION

Cet article se place dans la continuité de nos recherches sur l'interaction gestuelle avec des processus sonores. Nous considérons spécifiquement le problème de l'apprentissage de mappings entre geste et son présentant une structure temporelle hiérarchique. Si les modèles actuels permettent de modéliser le mapping par un ensemble de segments associant geste et son, leurs relations dans une structure de haut niveau doivent pour l'instant être fixées *a priori*. Au regard des récents développements dans le champ de l'apprentissage interactif, nous présentons une perspective pour l'apprentissage de telles structures de haut niveau. Nous détaillons en particulier une application visant à apprendre et à intégrer incrémentalement des gestes de co-articulation au sein d'une structure de transition prédéfinie.

6. REMERCIEMENTS

Ces recherches sont soutenues par le projet ANR LEGOS (11 BS02 012). Nous remercions tous les membres de l'équipe IMTR.

7. REFERENCES

[1] Bevilacqua, F., Zamborlin, B., Sypniewski, A., Schnell, N., Guédy, F., et Rasamimanana, N. "Continuous realtime gesture following and recognition", *Gesture in Embodied Communication and Human-Computer Interaction*, 73–84, 2010.

[2] Bevilacqua, F., Schnell, N., Rasamimanana, N., Zamborlin, B., et Guédy, F. "Online Gesture Analysis and Control of Audio Processing", *Musical Robots and Interactive Multimodal Systems*, 127–142. Springer, 2011.

[3] Caramiaux, B. *Etudes sur la relation geste–son en performance musicale*, Thèse de doctorat, Ircam – Université Pierre et Marie Curie, 2012.

[4] Cohn, D., Atlas, L., et Ladner, R. "Improving generalization with active learning", *Machine Learning, 15(2)*, 201–221, 1994.

[5] Cont, A., Dubnov, S., et Assayag, G. "Anticipatory model of musical style imitation using collaborative and competitive reinforcement learning", *Anticipatory Behavior in Adaptive Learning Systems*, Springer, New York, 2007.

[6] Fiebrink, R., Cook, P. R., et Trueman, D. "Play-along mapping of musical controllers", *Proceedings of the International Computer Music Conference (ICMC)*, 2009.

[7] Françoise, J. *Realtime Segmentation and Recognition of Gestures using Hierarchical Markov Models*, Mémoire de Master, Université Pierre et Marie Curie – Ircam, 2011.

[8] Françoise, J., Caramiaux, B., et Bevilacqua, F. "A Hierarchical Approach for the Design of Gesture–to–Sound Mappings", *Proceedings of the 9th Sound and Music Conference (SMC)*, Copenhague, Danemark, 2012.

[9] Franklin, J. A., et Manfredi, V. U. "Nonlinear Credit Assignment for Musical Sequences", *Second International Workshop on Intelligent Systems Design and Applications (ISDA)*, Atlanta, USA, 2002.

[10] Fu, S., Gutierrez-Osuna, R., Esposito, A., Kakumanu, P. K., et Garcia, O. N. "Audio/visual mapping with cross-modal hidden Markov models", *Multimedia, IEEE Transactions on*, 7(2), 243–252, 2005.

[11] Godøy, R. I., Jensenius, A. R., et Nymoen, K. "Chunking in music by co-articulation." *Acta Acustica united with Acustica,* 96(4), 690–700, 2010.

[12] Hofer, G. *Speech-driven animation using multimodal hidden Markov models.* Thèse de doctorat, University of Edinburgh, 2009.

[13] Hunt, A. et Kirk, R. "Mapping Strategies for Musical Performance." *Trends in Gestural Control of Music*, 231–258, 2000.

[14] Knox, W. B. *Learning from Human-Generated Reward*, Thèse de doctorat, Austin, TX, 2012.

[15] Le Groux, S., et Verschure, P. F. "Towards Adaptive Music Generation by Reinforcement", *Proceedings of the 7th Sound and Music Conference (SMC)*, Barcelona, Spain, 2010.

[16] Murphy, K. P. *Machine Learning : a Probabilistic Perspective*, MIT Press, Cambridge, MA, 2012.

[17] Puterman, Martin L. *Markov Decision Processes : Discrete Stochastic Dynamic Programming*, John Wiley & Sons, Inc., Hoboken, NJ, 1994.

[18] Rasamimanana, N., Bevilacqua, F., Schnell, N., Fléty, E. et Zamborlin, B. "Modular Musical Objects Towards Embodied Control Of Digital Music Real Time Musical Interactions", *Proceedings of the fifth international conference on Tangible, embedded, and embodied interaction*, 2011.

[19] Sutton, R. et Barto, A. *Reinforcement Learning : An Introduction*, MIT Press, Cambridge, MA, 1998.

[20] Wanderley, Marcelo M. et Battier, Marc. *Trends in Gestural Control of Music*, Ircam - Centre Pompidou, 2000.