Mémoire Master ATIAM
Ircam - UPMC

ISMM - Interaction Son Musique Mouvement

Description et synthèse sonore dans le cadre de l'apprentissage mouvement-son par démonstration

Auteur :
Pablo Arias

Sous la direction de :
Jules Françoise
Frédéric Bevilacqua
Norbert Schnell

Mars 2014 - août 2014

Remerciements

Je tiens à remercier Frédéric Bevilacqua pour m'avoir accueilli dans son équipe, Jules Françoise pour son aide et bonne humeur tout au long du stage ainsi que Norbert Schnell pour ses conseils et enthousiasme. Par ailleurs, un grand merci à toute l'équipe "Interaction Son Musique Mouvement" et particulièrement à Karim Barkati, Victor Saiz, Diemo Schwarz et Sébastien Robaszkiewicz. Enfin, à tout le personnel et chercheurs de l'IRCAM pour l'excellente ambiance de travail.

Table des matières

Résumé

Ce mémoire présente premièrement le développement d'une synthèse sonore paramétrique hybride composée d'une partie additive pour la synthèse de contenu harmonique et une partie granulaire avec conservation d'attaques pour les parties bruités et percussives. Deuxièmement, nous présentons l'intégration de cette synthèse dans le système d'apprentissage par démonstration, modèle d'apprentissage automatique utilisant des modèles de Markov cachés multimodales pour le suivi gestuel et l'inférence temps réel de paramètres sonores.

Par la suite, les résultats pour le contrôle de la synthèse paramétrique par l'inférence temps réel d'amplitudes et fréquences de partiels sont présentés et évalués. Enfin, nous présentons des possibilités de morphing sonore par interpolation des paramètres de partiels dans le but de créer un son intermédiaire contrôlé par un geste intermédiaire en utilisant le moteur de synthèse développé et le système d'apprentissage par démonstration.

Mots clés

— Mapping par démonstration, geste, synthèse audio, mapping, description sonore, synthèse granulaire, conservation d'attaques, synthèse additive, HMM

Abstract

This dissertation presents in the first place the developement of a hybrid sound synthesis engine composed by an additive part for the resynthesis of harmonic content and a granular part with onset preservation for noise and transients content. Furthermore, we present the integration of this synthesis engine in the mapping by demonstration system, a machine learning model using multimodal hidden Markov models in order to perform gesture following and real time inference of sound parameters. We present and evaluate the obtained results of the regression for the control of the parametric synthesis engine by infering the frequencies and amplitudes of sound partials. Finally, we present the possibility to perform sound parameter interpolation in order to create an intermediate sound controlled by an interemediate gesture recognised from examples using the mapping by demonstration system.

Key Words

— Mapping by demonstration, gesture, sound synthesis, mapping, sound description, granular synthesis, onset preservation, additive synthesis, HMM

Terminologie

— ISMM : Interaction Son Musique Mouvement
— IRCAM : Institut de Recherche et Coordination Acoustique Musique
— HMM : Hidden Markov Models
— HHMM : Hierachical Hidden Markov Model
— STS : Short Term Spectrum
— LFO : Low Frequency Oscilator
— ADSR : Attack, Decay, Sustain, Release
— SDIF : Sound Description Interchange Format
— MO : Modular Objects
— MFCC : Mell Frequency Cepstral Coefficient
— knn : k-nearest neighbors

1 Introduction

1.1 Présentation générale

L'interaction mouvement-son a fait l'objet de nombreuses recherches scientifiques et artistiques du fait non seulement de sa complexité mais aussi des enjeux et des possibilités qu'elle accapare. Le but de la recherche dans l'interaction mouvement-son est de comprendre et formaliser ces interactions afin de pouvoir les recréer de manière interactive pour proposer de nouveaux types d'expressions musicales et sonores.

Nous définissons une interaction comme une communication bidirectionnelle entre deux entités. L'utilisateur est au centre d'une boucle Action-Perception. Une interaction se compose de trois blocs, chacun jouant un rôle fondamental : Le contenu, la transmission du contenu et la perception ou interprétation.

L'interaction mouvement-son se doit de créer un lien intime entre la perception, la technique et l'art. Il faut donc être capable de modéliser un lien entre le geste et les paramètres de synthèse engageants.

Le but de ce stage est de créer un outil de synthèse temps réel qui est contrôlé par le système d'analyse de geste et de mapping par démonstration, système qui apprend le lien entre le geste et le son par l'exemple. Il s'agit d'étudier d'une part les différentes méthodes de suivi gestuel et d'autre part de concevoir une synthèse qui s'adapte au mapping par démonstration afin d'ouvrir des possibilités de contrôle gestuel.

Dans ce rapport nous étudions la description et synthèse sonore dans le cadre de l'apprentissage mouvement-son par démonstration. Ce rapport aborde premièrement une présentation succincte de l'état de l'art puis présente une description précise de la problématique du stage pour enfin exposer le travail effectué, les résultats obtenus et les perspectives futures.

1.2 Cadre du stage

Ce stage est effectué dans le cadre du master *Acoustique, Traitement du signal et Informatique Appliqués à la Musique* de l'Université Pierre et Marie Curie, en partenariat avec l'IRCAM. Il se déroule dans l'équipe ISMM (Interaction Son Musique Mouvement) à l'IRCAM sous la direction de Jules Françoise, Norbert Schnell et Frédéric Bevilacqua. L'équipe ISMM mène des recherches et développements sur des systèmes interactifs dédiés à la musique et au spectacle vivant.

1.3 Contexte

Le moteur de synthèse à étudier et à réaliser pendant ce stage s'ancre dans un système d'apprentissage par démonstration du couplage mouvement-son. Celui-ci vise à déterminer le "mapping" entre mouvement et son à partir d'exemples fournis par l'utilisateur. Le but de l'étude est de créer un système de synthèse qui sera contrôlé par ce système de reconnaissance

de geste. L'apprentissage par démonstration des relations entre mouvement et son s'appuie sur des modèles d'apprentissage automatique multimodaux. Ces modèles permettent d'apprendre le mapping entre mouvement et son par des exemples fournis par l'utilisateur. Ces exemples sont généralement des mouvements effectués parallèlement à l'écoute d'un son ou à la production du son (e.g. des vocalisations) qui illustrent explicitement la relation entre son et geste. Une fois entrainé, le système est utilisé pour le contrôle d'un moteur de synthèse sonore. Ce système nécessite donc des modèles cohérents d'analyse-synthèse de sons, permettant un contrôle temps-réel fluide. Il s'agit de proposer de nouvelles méthodes de synthèse adaptées à ce système.

1.4 Objectifs

Ce stage vise à formaliser l'articulation entre description sonore et synthèse pour le cas particulier de l'apprentissage par démonstration. Il s'agit de proposer une approche hybride de synthèse sonore (liant synthèses additive, granulaire, concaténative) de manière à améliorer la qualité de la resynthèse de sons enregistrés. Dans le système original cette synthèse se fait uniquement par synthèse granulaire.

Une problématique principale dans ce cadre est la décomposition et description des contenus sonores et leur resynthèse pilotée par le mouvement à travers des modèles de "mapping par démonstration". Cette description et décomposition concernent à la fois des caractéristiques instantanées du son (i.e. hauteurs, énergie, timbre) que des aspects temporels (i.e. segmentation, évolution temporelle). La finalité de ce travail est donc l'amélioration de la synthèse sonore dans un but d'expressivité : il s'agit de fournir une représentation et un contrôle modulaire sur différents aspects du son.

Les objectifs à réaliser au cours de ce stage et abordés par ce rapport sont premièrement une bibliographie autour des méthodes de synthèse, description sonore, et interaction mouvement-son. Deuxièmement, une formalisation du problème de description/synthèse dans le cadre de l'approche "mapping par démonstration" pour pouvoir proposer et intégrer un système de synthèse sonore hybride (additive, granulaire, concaténative, soustractive) qui s'adapte à ce cadre. Enfin, une dernière partie consiste à étudier des alternatives de contrôle du synthétiseur, comme le contrôle par HMM, et d'évaluer les performances (qualité sonore, cohérence, expérience utilisateur).

2 État de l'art

Cette partie présente différentes avancées de la recherche scientifique dans plusieurs domaines. Ces avancées ont inspiré le travail réalisé au cours du stage pour répondre à des problèmes variés. Ce stage constitue un lien entre ces différents domaines scientifiques.

2.1 Mapping

Le musicologue Rolf Inge Godøy discute des liens et relations entre le son et le geste. En général, les mouvements corporels ont toujours une conséquence sonore. Godøy explique que l'écoute et le mouvement sont intimement liés. Nous les associons naturellement ensembles [11].

Cependant, dans le monde virtuel le passage du geste au son doit être modélisé. C'est le rôle du mapping, étape permettant de passer des paramètres du mouvement (notamment les paramètres issus de la reconnaissance de geste) à la génération sonore, afin de créer un lien entre le mouvement et le son.

Arfib et al. proposent de distinguer différents types de mapping : le mapping explicite/implicite, simple/complex et dynamique/statique [1].

Le mapping explicite consiste à avoir un contrôle direct d'un paramètre sonore par un paramètre gestuel. Par exemple, contrôler la hauteur d'un instrument par le mouvement vertical de la main. Le mapping implicite suppose l'utilisation d'un modèle intermédiaire créant le lien entre mouvement et son, par exemple un modèle dynamique, ou un modèle d'apprentissage automatique statistique.

D'autre part, le mapping dynamique en opposition au mapping statique se concentre sur l'aspect temporel des relations entre paramètres sonores aux paramètres gestuels. Le mapping dynamique est adaptatif : il évolue dans le temps. Celui-ci prend en compte l'évolution du geste ainsi que son passé. Ainsi, un même mouvement peut potentiellement contrôler plusieurs sons ou plusieurs paramètres de synthèse. Un mapping dynamique peut donc changer et évoluer. Pour le mapping statique, au contraire, les paramètres contrôlés sont fixes et choisis à l'avance.

Enfin, une dernière catégorie est le mapping complexe/simple présenté par Hunt [12]. Le mapping complexe est défini comme un mapping de "plusieurs paramètres vers plusieurs". Par exemple, dans notre cas, plusieurs paramètres du geste seraient associés à plusieurs paramètres de la synthèse (utilisation de réseaux de neurones, bases de données ...). Un mapping simple est une association "un vers un" entre les paramètres : Pour chaque paramètre contrôlé il y a un contrôleur et inversement. En général, pour le mapping statique il existe des relations analytiques entre les paramètres gestuels et les paramètres sonores. Le mapping complexe ajoute une étape entre les paramètres.

Un exemple de mapping complexe est présenté par Françoise [14] où une approche par des HHMM (Hierachical Hidden Markov Model) pour le mapping est utilisé. Dans l'application implémentée dans l'article le geste est analysé et contrôle un vocodeur de phase en temps

réel afin de piloter l'algorithme de "time-stretch" en relation avec la vitesse et la temporalité du geste. Le geste ne contrôle pas un paramètre directement mais c'est l'analyse du geste qui permet de contrôler l'algorithme. Ce type de mapping est une extension du mapping temporel introduit par Bevilacqua [3] où l'évolution temporelle est utilisée pour le contrôle d'effets audionumériques.

Par ailleurs, Arfib et al. [1] soulignent l'importance du choix des paramètres pour contrôler de façon intuitive le moteur de synthèse. En effet, l'évolution sonore peut être contrôlée par des paramètres liés à la perception afin de donner au son les caractéristiques du geste.

Enfin, Caramiaux et al. [5] présentent une formalisation du mapping selon des modes d'écoute : dans cette approche chaque dispositif interactif est informé par un mode d'écoute particulier. Ces travaux soulignent l'importance de prendre en compte l'ensemble de la boucle Action-Perception pour la conception de systèmes sonores interactifs. L'approche de "Mapping par démonstration" vise à intégrer cette boucle du point de vue utilisateur, en lui permettant de définir les relations entre mouvement et son par démonstration en effectuant des gestes durant l'écoute.

2.2 Analyse du geste

L'analyse du geste et son interprétation automatique sont très étudiés depuis quelques années. Plusieurs nouvelles technologies et applications grand public ont été commercialisées. Au niveau scientifique, nombreuses techniques de traitement du geste ont été développées au sein de l'équipe ISMM notamment par l'utilisation de méthodes d'apprentissage automatique [3, 14, 9].

Par exemple, Bevilacqua et al. [4] présentent une méthode d'apprentissage et reconnaissance du geste utilisant des HMM qui permet de se situer temporellement dans un geste. Les gestes captés sont comparés aux gestes enregistrés par une étude de maximum de vraisemblance ce qui permet de connaître la similarité entre eux.

Au niveau de la reconnaissance de geste, le système proposé par Françoise et al. [14, 9] pour le suivi de l'évolution temporelle d'un geste par l'utilisation de Modèles de Markov cachés nous intéresse particulièrement.

2.2.1 Modèles de Markov cachés pour le suivi gestuel

Dans l'équipe ISMM, les modèles de Markov cachés ou HMM (Hidden Markov model) [19] ont été utilisés dans plusieurs systèmes de reconnaissance de geste pour la création sonore [14, 9].

Dans ce stage nous nous intéressons particulièrement aux modèles de Markov cachés qui permettent de modéliser et de suivre l'évolution temporelle d'un geste et d'estimer des paramètres sonores. Le système de mapping par démonstration utilisé au long de ce stage se sert de ces modèles.

Chaînes de Markov

Les chaînes de Markov permettent de modéliser l'évolution d'un processus stochastique ou aléatoire qui respecte la propriété de Markov. Celle-ci stipule que l'état futur dépend uniquement de l'état présent et pas de la séquence d'événements passés. Ainsi, une chaîne de Markov peut être définie comme un processus de Markov possédant un nombre fini d'états discrets.

Modèles de Markov cachés

Les modèles de Markov cachés permettent de modéliser des séquences d'observations, sous l'hypothèse que celles-ci sont générées par un processus de Markov. Contrairement aux chaînes de Markov, ce processus n'est pas directement observable. L'information est encodée dans des états cachés, qui génèrent des observations au travers d'une fonction d'observation probabiliste.

Les HMM (Hidden Markov Models) sont donc une modélisation statistique de l'évolution temporelle d'un processus de Markov. Dans les HMM, à différence des chaînes de Markov, les états sont cachés. En clair, pour les HMM, nous considérons que chaque état correspond à un phénomène observable. Les états ne peuvent pas être eux-mêmes observés. Par contre, il est possible de construire une séquence d'observation à partir d'un ensemble de processus stochastiques.

D'après Rabiner [19], un HMM peut être représenté par un ensemble de cinq variables dans le cas continu :

- Nombre d'états N du modèle, l'ensemble d'états est noté $S = S_1, S_2, ..., S_N$ et q_t est l'état présent.
- La distribution initiale de probabilités $\pi = [\pi_i]$ où $\pi_i = P[q_1 = S_i], 1 \leq i \leq N$.
- La matrice des probabilité de transition $A = [a_{ij}]$ où $a_{ij} = P[q_{t+1} = S_j | q_t = S_i]$ avec $1 \leq i, j \leq N$
- La probabilité d'observer l'observation o_t à l'instant t depuis l'état j, $b_j(o_t) = P[o_t | q_t = S_j]$ avec $1 \leq j \leq N$. Dans le cas d'un modèle Gaussien, $b_j(o_t) = \mathcal{N}(o_t; \mu_j, \Sigma_j)$ où μ_j et Σ_j représentent respectivement la moyenne et la covariance de la distribution Gaussienne associée à l'état j.

Un exemple de modèle de Markov caché pour le suivi gestuel est présenté en figure 1 .

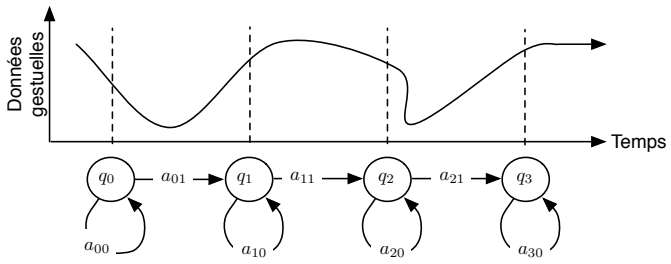


FIGURE 1 – Exemple de modèle de Markov caché pour le suivi gestuel

Pour une description plus précise des modèles de Markov et des modèles de markov cachés se référer à [19, 17, 18].

2.2.2 Le système d'apprentissage mouvement-son par démonstration

La méthode utilisée dans le cadre de ce stage pour le contrôle du moteur de synthèse a été développée par Françoise [9]. Cette méthode est un système de reconnaissance gestuelle par apprentissage. Sa structure se divise en trois étapes : Apprentissage, description, synthèse comme présenté dans la figure 2.

Premièrement, l'utilisateur apprends au système un son par démonstration : il joue le geste et réalise le son en simultané par vocalisation. Le système enregistre en temps réel

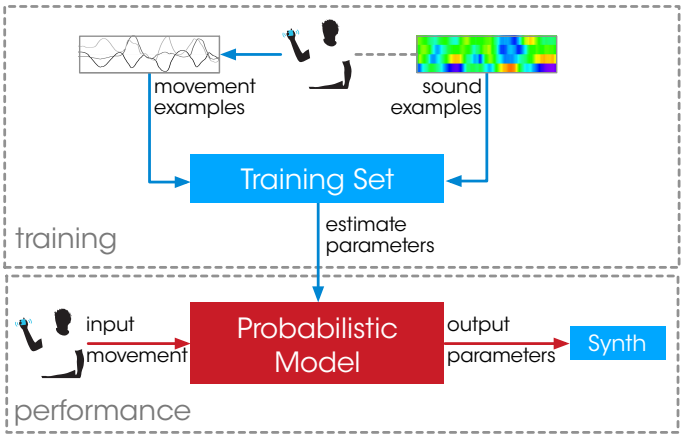


FIGURE 2 – Mapping geste son en utilisant des HMM multimodaux. (Tiré de [10])

le déroulement du geste et du son et leur concordance temporelle. Une analyse du son est faite, plusieurs descripteurs audio sont utilisés. Ainsi, l'évolution des descripteurs sonores au cours du temps peut être comparée à l'évolution gestuelle. C'est la dimension multimodale du système.

Une fois l'enregistrement fini le système apprend les relations entre séquences de descripteurs gestuels et sonores par un modèle de Markov caché multimodal. Dans la phase de "performance", le modèle est utilisé pour générer en temps réel les paramètres sonores associés à la performance d'un nouveau mouvement. Les auteurs proposent d'utiliser une synthèse granulaire par sélection d'unités, pour laquelle les grains contenus dans le corpus d'exemples sont sélectionnés et joués selon leur description sonore à l'aide d'un algorithme de "K-Nearest Neighbors" (KNN) [26].

Les HMM multimodaux pour l'inférence de données sonores

Les HMM multimodaux sont une extension des HMM pour prendre en compte l'évolution de données multimodales. Les données multimodales sont créées par la combinaison de plusieurs moyens de communication simultanés entre l'utilisateur et la machine. Le système d'apprentissage par démonstration est dit multimodal car la voix et le geste ont en commun leur évolution temporelle. Nous modélisons leurs séquences d'observations respectives par un même modèle de Markov caché qui encode leur évolution temporelle et les dépendances entre les deux modalités.

Pour le modèle multimodal, tant le son que le geste sont générés par le même processus de Markov, de ce fait nous modélisons conjointement leurs séquences d'observations. Durant la phase d'entraînement, les données gestuelles et sonores sont utilisées pour former des séquences multimodales. Les vecteurs de données de son et de geste sont concaténés pour créer une seule séquence. Ainsi, pour l'état j d'un HMM de paramètre λ la loi de probabilité est définie comme une distribution gaussienne multimodale jointe :

$$P(\sigma_t^g | q_t = j, \lambda) = \mathcal{N}([\sigma_t^g, \sigma_t^s]; \mu_j, \Sigma_j) \tag{1}$$

Où q_t est l'état caché à l'instant t, σ_t^g le vecteur d'observation du geste, σ_t^s le vecteur d'observation du son et μ_j est la concaténation des vecteurs moyens pour le son et le geste.

$$\mu_j = [\mu_j^g, \mu_j^s] \tag{2}$$

Enfin, Σ_j est la matrice de covariance définie par l'équation 3 :

$$\Sigma_j = \begin{pmatrix} \Sigma_j^{gg} & \Sigma_j^{gs} \\ \Sigma_j^{sg} & \Sigma_j^{ss} \end{pmatrix} \tag{3}$$

La matrice de covariance 3 est composée de quatre sous matrices de covariance du fait de la dimension multimodale du système.

Apprentissage et entraînement
Le HMM est entraîné grâce à un algorithme d'espérance-maximisation (Expectation–maximization ou EM). Cet algorithme permet de trouver le maximum de vraisemblance de paramètres de modèles probabilistes.

La régression
En suivant ce modèle, grâce à l'apprentissage des données gestuelles et sonores en simultanné et à leur concatenation pour l'apprentissage nous posons :

$$p(\sigma_t^s | \sigma_t^g, q_t = j, \lambda) = \mathcal{N}(\sigma_t^s; \hat{\mu}_j^s(\sigma_t^g), \hat{\Sigma}_j^{ss}) \tag{4}$$

où la moyenne $\hat{\mu}_j^s$ et la covariance $\hat{\Sigma}_j^{ss}$ du son sont ré-estimés par la combinaison de la moyenne apprise du son et une régression linéaire sur les données gestuels à partir de l'équation 3 :

$$\hat{\mu}_j^s = \mu_j^s + \Sigma^{sg}(\Sigma_j^{gg})^{-1}(\sigma_t^g - \mu_j^g) \tag{5}$$

$$\hat{\Sigma}_j^{ss} = \Sigma_j^{ss}(\Sigma_j^{gg})^{-1}(\sigma_j^{gs}) \tag{6}$$

Nous pouvons déduire les données sonores correspondantes à un geste en entrée pour un apprentissage donné. Ce modèle probabiliste permet donc d'apprendre le lien entre les variations de données de mouvement et les variations de données sonores. Cependant, la resynthèse du son enregistré est de fait limitée au contenu du corpus d'origine. Ce stage vise à améliorer le moteur de synthèse sonore par la combinaison de synthèse additive, qui permet un contrôle continu sur les parties harmoniques, avec la synthèse granulaire.

2.3 Synthèse Sonore

Ce stage vise au développement d'un système d'analyse-synthèse sonore permettant un contrôle modulaire et indépendant des différentes composantes du son. Comme vu précédemment, le système d'apprentissage par démonstration permet d'inférer des paramètres sonores en temps réels par rapport à l'exemple, il s'agit de créer un synthétiseur paramétrique intégrant synthèses granulaire, concaténative, et additive dans le but d'être contrôlé par ce système. Cette section détaille les recherches sur la synthèse sonore. L'objectif est d'adapter la synthèse au contrôle par estimation de paramètres sonores.

2.3.1 Synthèse granulaire

L'idée de la synthèse granulaire est de créer des sons à partir de fragments sonores appelés grains d'une durée de quelques dizaines de millisecondes. Cette synthèse s'appuie sur une superposition de ces grains sans cohérence de phase. Le point fort au niveau design sonore est la quantité de paramètres de contrôle qu'elle offre ainsi que les effets sonores produits par ces derniers. En effet, les grains et leurs superpositions peuvent être ajustées. Leurs durée, période, enveloppe entre autres peuvent être adaptés, chaque paramètre changeant les caractéristiques sonores. Cette synthèse est très utilisée pour la création de textures sonores.

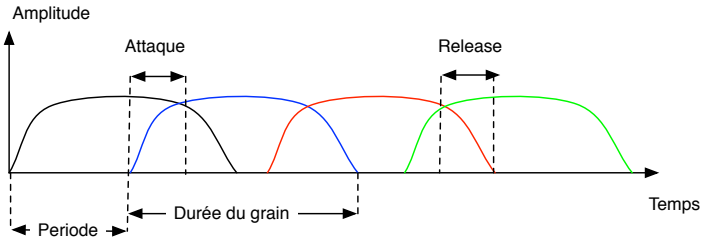


FIGURE 3 – Paramètres et superposition de grains dans la synthèse granulaire.

La synthèse granulaire est intéressante dans notre cas car elle permet de réaliser de la resynthèse en modifiant la temporalité du son. D'ailleurs, la première application du système d'apprentissage par démonstration est le contrôle d'un indice temporel par le geste. La resynthèse étant faite par une synthèse granulaire classique.

Le synthétiseur `mubu.granular~` développé par l'équipe ISMM a été largement utilisé au cours de ce stage. Dans celui-ci chaque grain est modulé par une enveloppe ASR (Attack - Sustain - Release) et autres paramètres comme montré dans la figure 3.

2.3.2 Synthèse concaténative

La synthèse concaténative, dont l'historique est décrit par Schwarz [24], est une technique de création sonore basée sur la concaténation d'échantillons présents dans un corpus. Ces échantillons sont décris par des descripteurs sonores et sont généralement de quelques millisecondes. La recherche des échantillons à jouer se fait en temps réel en utilisant des méthodes de recherche comme l'algorithme knn.

Par exemple, le système catart de Schwartz [25] utilise un ensemble de descripteurs sonores pour définir un plan à deux dimensions où les échantillons du corpus sont placés dans l'espace par rapport à leurs description. Ainsi, il est possible de concaténer des unités sonores qui sont proches entre elles au niveau de la description choisie [25, 27]. Par ailleurs, du fait de la superposition de grains, la synthèse concatenative est intéressante pour la synthèse de textures sonores et de contenu bruité.

Dans le cadre de ce stage la synthèse concaténative peut permettre de distinguer le son par rapport à ses caractéristiques gestuelles, notamment à la distinction de sons avec attaques afin de conserver la structure sonore tout en contrôlant l'évolution temporelle par le geste.

2.3.3 Synthèse additive

Un signal périodique peut être décomposé en somme de composantes sinusoïdales de fréquences multiples de sa fréquence fondamentale. Or, les signaux musicaux peuvent souvent être considérés comme pseudopériodiques. De ce fait, ils peuvent être modélisés par superposition de signaux sinusoïdaux variant en fréquence, en amplitude et en phase au cours du temps. Ceci afin de recréer leur timbre.

Ainsi, un modèle de synthèse additive peut s'écrire :

$$y(t) = \sum_{k=1}^{K} A_k(t) cos(2\pi k f_0 t + \phi_k) \tag{7}$$

Où A_k, f_0, ϕ_k sont respectivement l'amplitude, la fréquence et la phase de l'harmonique de rang k.

Par ailleurs, certains auteurs se sont intéressés à l'utilisation de la synthèse additive pour la resynthèse sonore et notamment sur l'extension de cette technique pour resynthétiser les phénomènes percussifs et bruités du son [13, 22, 28].

Dans sa thèse Serra présente une méthode qui part de l'hypothèse que le son est composé d'un contenu déterministe et d'un contenu stochastique. La méthode propose une synthèse en deux parties, une partie sinusoïdale synthétisant le contenu harmonique et une partie résiduelle synthétisant le contenu bruité [28] .

Ainsi, au contenu harmonique de l'équation 7 s'ajoute le contenu stochastique représenté par une série d'enveloppes spectrales fonctionnant comme des filtres variant dans le temps et excités par un bruit blanc :

$$y(t) = \sum_{k=1}^{K} A_k(t) cos[\theta_k(t)] + e(t) \tag{8}$$

Où K est le nombre de sinusoïdes du modèle et A_k et $\theta_k(t)$ sont respectivement l'amplitude et la phase instantanée. Les calculs pour obtenir l'expression de la phase instantanée $\theta_k(t)$ sont présentés dans [28].

Il y a deux différences entre le modèle de Serra [28] et le modèle sinusoïdal présenté en équation 7. D'une part, une partie résiduelle est calculée afin de modéliser la partie bruité. D'autre part, les sinusoïdes sont restreintes à être stables (suivre les composantes quasi sinusoïdales du signal original) et donc de suivre les partiels du son.

Rodet décrit une autre méthode de synthèse additive par FFT^{-1}. Pour cette resynthèse Rodet considère de même que plusieurs signaux musicaux peuvent être décrits comme une combinaison d'ondes pseudo-périodiques et de bruit coloré. Cette méthode repose sur la construction d'un STS (Short Term Spectrum) et d'un ajout de bruit à certains endroits du signal [21].

Enfin, d'autres modèles proposent des extensions de la synthèse additive en ajoutant une étape de modélisation des transitoires [2, 13].

Les avantages de la synthèse additive dans le cadre de ce stage est le fait qu'elle soit paramétrique et que le contenu harmonique puisse être contrôlé via les partiels. De ce fait, plusieurs possibilités de contrôle s'ouvrent par la modulation de F_0, du timbre entre autres.

2.4 Description sonore

La description sonore est l'extraction de descripteurs audio à partir de l'étude d'un son. Un descripteur audio est une valeur constante ou variable dans le temps qui décrit le signal.

Peeters, par exemple, présente une grande gamme de descripteurs sonores en [16].

Dans sa thèse, Schwarz fait une distinction entre trois types de descripteurs. Premièrement, les descripteurs de bas niveau (Low level descriptors ou LLD) sont extraits de la source sonore en utilisant des méthodes de traitement du signal. Deuxièmement, les descripteurs de haut niveau (High-level descriptors ou HLD) sont ceux qui ont un sens musical. Enfin, les descripteurs perceptifs, se situent à un niveau intermédiaire et sont déduits des LLD pour leur donner un sens perceptif [25].

Dans ce stage les descripteurs audio jouent un rôle fondamental au niveau du mapping. L'idée est d'apprendre et de comparer par l'exemple l'évolution temporelle de ces signaux à ceux décrivant le geste. Le choix de ces descripteurs doit donc correspondre à la réaction attendue du système.

En particulier, ce stage vise à articuler description sonore et synthèse avec une structure modulaire. Il s'agit de séparer parties harmoniques et bruitées afin de permettre un contrôle expressif indépendant pour chacune de ces composantes, tout en garantissant la conservation d'attaques. Nous nous intéresserons particulièrement aux partiels sonores, descripteurs de l'évolution harmonique du son.

2.5 Technologies

L'équipe ISMM et autres équipes de l'IRCAM ont développés au cours des années plusieurs technologies pour la synthèse et la description audio et geste temps réel. Le synthétiseur conçu gravite autour de toutes ces technologies. Ces technologies seront utilisées au cours du stage :

- SDIF - Sound Description Interchange Format : format pour le stockage et la communication de descripteurs sonores.
- Pipo (Plug-in, Plug-out) : API C++ pour la définition de blocs de traitement audio temps réel.
- Mubu : Conteneur multi-buffer pour Max Msp qui permet de stocker des descriptions de sons et de gestes[15].
- MO : Musical Objects sont des objets qui permettent de capter, analyser et enregistrer des gestes via des capteurs intertiaux (gyroscopes, accéléromètres) [23].

3 Problématique et formalisation du problème

Comme expliqué précédemment le but de ce stage a été de développer un moteur de synthèse en lien avec le système d'apprentissage par démonstration présenté par Françoise [8] et expliqué en section 2.2.2.

En pratique, l'utilisateur apprends au système un geste en simultané avec un son. Ce système permet de démontrer les liens geste-son par l'exemple, un utilisateur peut donc apprendre au système différentes performances que le système est capable de suivre et re-synthétiser en cohérence avec la performance gestuelle.

L'objectif de ce stage est d'améliorer la qualité de la synthèse en créant un synthétiseur conçu pour être contrôlé par le système de mapping par démonstration. Le synthétiseur est orienté vers la resynthèse de vocalisations composées de contenu harmonique, bruité et transitoire.

Premièrement, ce synthétiseur se veut paramétrique afin d'étendre les possibilités de contrôle de la synthèse par le geste. Le but est de contrôler le synthétiseur par des paramètres sonores. Ceux-ci seront estimés en temps réel par des HMM en utilisant le système de mapping par démonstration.

Deuxièmement, le synthétiseur doit pouvoir réaliser de la synthèse en temps réel par rapport à l'analyse des données gestuelles fournissant des stratégies de contrôle complémentaires sur différents aspects du son. Celui-ci doit, par exemple, permettre de contrôler la temporalité du son par le geste.

Troisièmement, le synthétiseur doit être capable de conserver le lien geste-son. En effet, lors de la resynthèse, la concordance geste-son démontrée doit être respectée, de ce fait, les caractéristiques sonores qui créent l'identité du son, comme les attaques, doivent être préservées. Le synthétiseur doit donc pouvoir tout en modulant la temporalité du son conserver ces "zones caractéristiques" du lien geste-son.

4 Moteur de synthèse développé

Cette partie se concentre dans la présentation du moteur de synthèse développé. Chaque sous partie a été réalisée avec l'idée de répondre à un but précis du problème présenté en section 2.5.

4.1 Synthèse granulaire à conservation d'attaques

4.1.1 Position du problème

Les premiers travaux de resynthèse sonore en utilisant le système de mapping par démonstration [8] ont été de contrôler un moteur de synthèse granulaire par un index temporel pour la localisation des grains sonores dans le son fourni en exemple.

Comme décrit dans l'état de l'art, la synthèse granulaire se base sur la superposition de grains sonores de quelques dizaines de millisecondes (Figure 3).

L'avantage de l'utilisation de la synthèse granulaire dans ce cadre est principalement de pouvoir contrôler la temporalité de la resynthèse. Cependant, comme expliqué précédemment, celle-ci est orientée vers la synthèse de textures sonores du fait de la superposition de grains et du non respect de la phase du signal original.

De ce fait, la synthèse de vocalisations composées de transitoires, de contenu bruité et harmonique n'est pas satisfaisante. En effet, quand un grain de la synthèse est composé d'une partie de transitoire le son créer est désagréable à l'écoute. Par ailleurs, comme expliqué précédemment les transitoires sont à la base de la structure du son et donc du lien geste-son démontré par l'utilisateur.

De ce fait, un premier développement est la création d'une synthèse granulaire à conservation d'attaques. Les développements de ce moteur granulaire ont été faits sous Max par contrôle temps réel du synthétiseur `mubu.granular~` développé par l'équipe ISMM.

`mubu.granular~` présente les paramètres habituels de la synthèse granulaire, l'idée des développements est de les modifier en temps réel par rapport au signal resynthétisé. Le contrôle des enveloppes des grains nous intéresse particulièrement. Par ailleurs, nous voulons pouvoir contrôler la durée de chaque grain et la période de déclenchement afin de pouvoir reproduire les attaques du son d'origine.

Les premiers développements ont été de créer un synthétiseur granulaire à conservation d'attaques. Celui-ci adapte les paramètres de synthèse par rapport au signal sonore à resynthétiser afin de jouer les transitoires et les attaques sans modifications et ainsi préserver les caractéristiques structurelles de l'exemple sonore. D'autre part, ceci permet d'éviter de prendre des grains contenant des transitoires dans la synthèse granulaire ce qui génère un son désagréable du fait de la répétition trop rapide de fragments de transitoires.

4.1.2 Discrimination du signal pour la synthèse granulaire à conservation d'attaques

Pour ce faire la première étape se passe en temps différé et est une détection des positions et des durées des attaques dans l'exemple sonore.

Détection des attaques
Afin de détecter la position des onsets dans le signal nous réalisons les opérations de la figure 4 sur le signal $x(t)$:

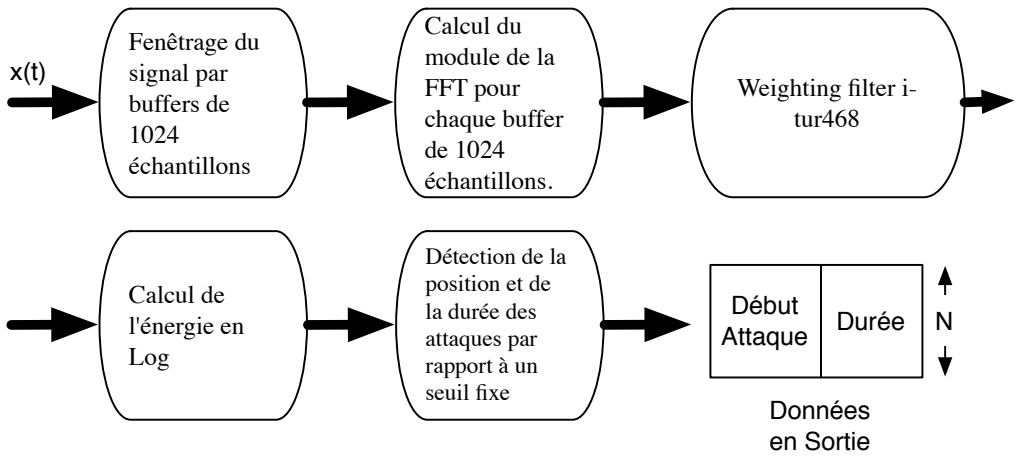


FIGURE 4 – Description de l'algorithme pour la detection de la position et de la durée des attaques dans le signal $x(t)$.

La détection de la position et de la durée des attaques par l'algorithme de la figure 4 se fait par l'étude de la sonie. Plus celle-ci augmente plus il y a d'énergie dans le spectre. Les transitoires et attaques sont caractérisés par la présence d'énergie sur tout le spectre. Quand l'évolution de la sonie dépasse un seuil à l'instant t nous considérons que le signal $x(t)$ est composé d'une attaque. Tant que la sonie reste au-dessus du seuil on considère que t est à l'intérieur de la même attaque, nous en déduisons la position et la durée.

Discrimination du signal pour la synthèse granulaire
À partir de ces marqueurs nous pouvons déduire les parties du signal qui peuvent être utilisées pour une synthèse granulaire classique avec plusieurs grains superposés et les parties qui doivent être jouées comme dans le son original.

Nous définissons donc pour un signal $x(t)$ des "zones de transitoires" et des "zones de textures" où la synthèse granulaire classique peut se faire. Ainsi, un index temporel t est situé dans une zone de transitoire si $\exists k \in [0, K]$ tel que,

$$pa_k - d_u < t < pa_k + da_k \tag{9}$$

où K est le nombre d'attaques dans le signal, d_u est la longueur du grain fixé pour la synthèse par l'utilisateur et da_k et pa_k sont respectivement la durée et la position de début de l'attaque k.

Le paramètre d_u peut varier au cours du temps car il est contrôlable par l'utilisateur du synthétiseur. De ce fait, les bornes de définition des zones "transitoires" ou "textures" peuvent elles-mêmes varier en temps réel. La définition de ces bornes se fait donc à chaque redéfinition de d_u.

Par ailleurs, la borne inférieure de l'inéquation $(pa_k - lg)$ permet d'éviter qu'aucune partie d'un grain sonore de longueur d_u soit à l'intérieur d'une attaque. En effet, même si la position initiale d'un grain est située en dehors d'une attaque, la fin elle peut contenir des fragments de transitoire. Ceci permet de s'assurer qu'il n'y aura pas de granulation sur une attaque. La figure 5 montre cette catégorisation par des marqueurs.

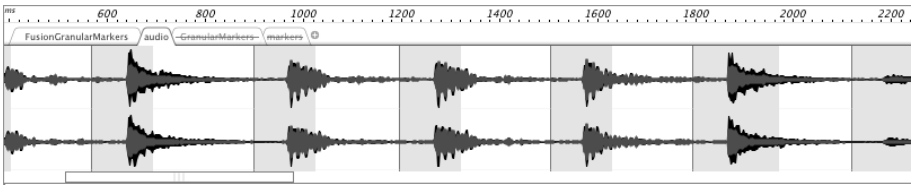


FIGURE 5 – Exemple de zones de textures (blanc) et transitoires (gris) pour un grain de 70 ms sur un son de batterie.

Le moteur de synthèse intercale donc une synthèse granulaire "classique" avec des transitoires joués une seule fois, sans superposition. Afin de pouvoir jouer chaque transitoire tel que le son original nous réalisons un changement temps réels des enveloppes et paramètres du grain comme dans la figure 6

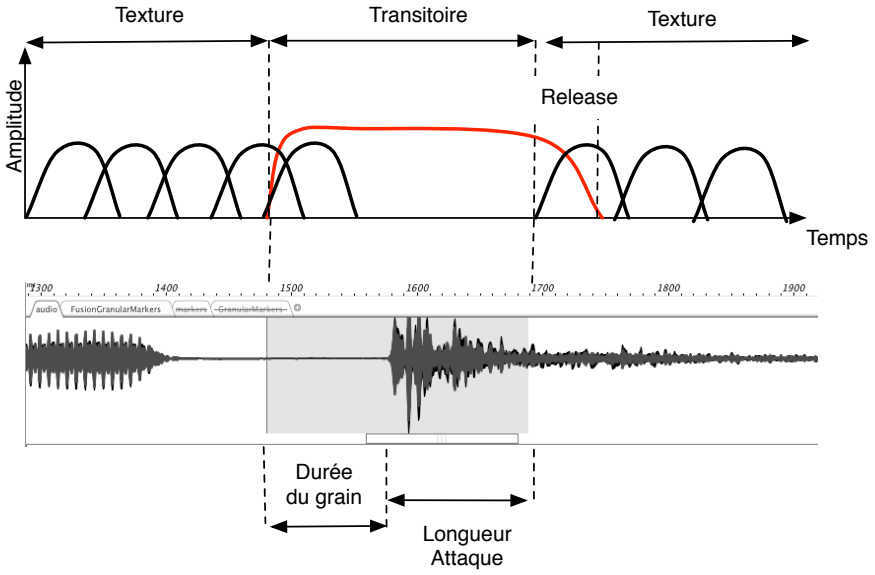


FIGURE 6 – Exemple de la superposition de grains dans la synthèse granulaire à conservation d'attaques.

4.2 Structure et fonctionnement du synthétiseur granulaire à conservation d'attaques

Afin de modéliser le fonctionnement présenté en figure 6, le synthétiseur se comporte comme l'automate à états fini en figure 7, les paramètres sont décrits dans le tableau 1 :

TABLE 1 – Paramètres du synthétiseur granulaire à conservation d 'attaques

Paramètres de l'automate du moteur granulaire	
r	Release de la synthèse granulaire
d	durée du grain de la synthèse granulaire
p	Periode de la synthèse granulaire
a	Attaque de la synthèse granulaire
K	Nombre de zones "non granulables"
da_k	durée de la kième zone "non granulable"
pa_k	position de la kième zone "non granulable"
r_u	Release donnée par l'utilisateur
d_u	Durée du grain donnée par l'utilisateur
p_u	Période donnée par l'utilisateur
a_u	Attaque donnée par l'utilisateur

La figure 7 montre le fonctionnement à respecter afin de moduler les paramètres de la synthèse granulaire pour la conservation des attaques. Tout d'abord, pour une position t particulière nous analysons si celle-ci se situe à l'intérieur d'un attaque par analyse de l'inéquation 9. Si c'est pas le cas, la synthèse granulaire se fait en utilisant les paramètres fixés par l'utilisateur.

Si l'index t est à l'intérieur d'une attaque, Il faut changer les paramètres des enveloppes et positionner l'index temporel au début de cette attaque puis le jouer. À la fin de l'attaque il faut positionner l'index temporel de la synthèse granulaire à la fin de l'attaque jouée et revenir aux paramètres fixés par l'utilisateur pour la synthèse de textures.

4.3 Le moteur de synthèse additive

4.3.1 Position du problème

La synthèse granulaire à conservation d'attaques développée et présentée dans la partie précédente permet de conserver les parties percussives et bruitées d'un son. En général, la synthèse granulaire s'utilise pour la création de textures du fait de la superposition d'atomes sonores. Cependant, cette synthèse ne permet pas de resynthétiser des contenus harmoniques. En effet, la superposition des grains ne respecte pas la phase ce qui résulte en des parties harmoniques plus bruitées. Les travaux réalisés autour de la synthèse additive s'orientent dans cette direction : améliorer la resynthèse et le contrôle du contenu harmonique.

Par ailleurs, une autre motivation de l'utilisation de la synthèse additive est le fait que celle-ci soit paramétrique : la génération sonore se fait entièrement par le contrôle de paramètres de partiels.

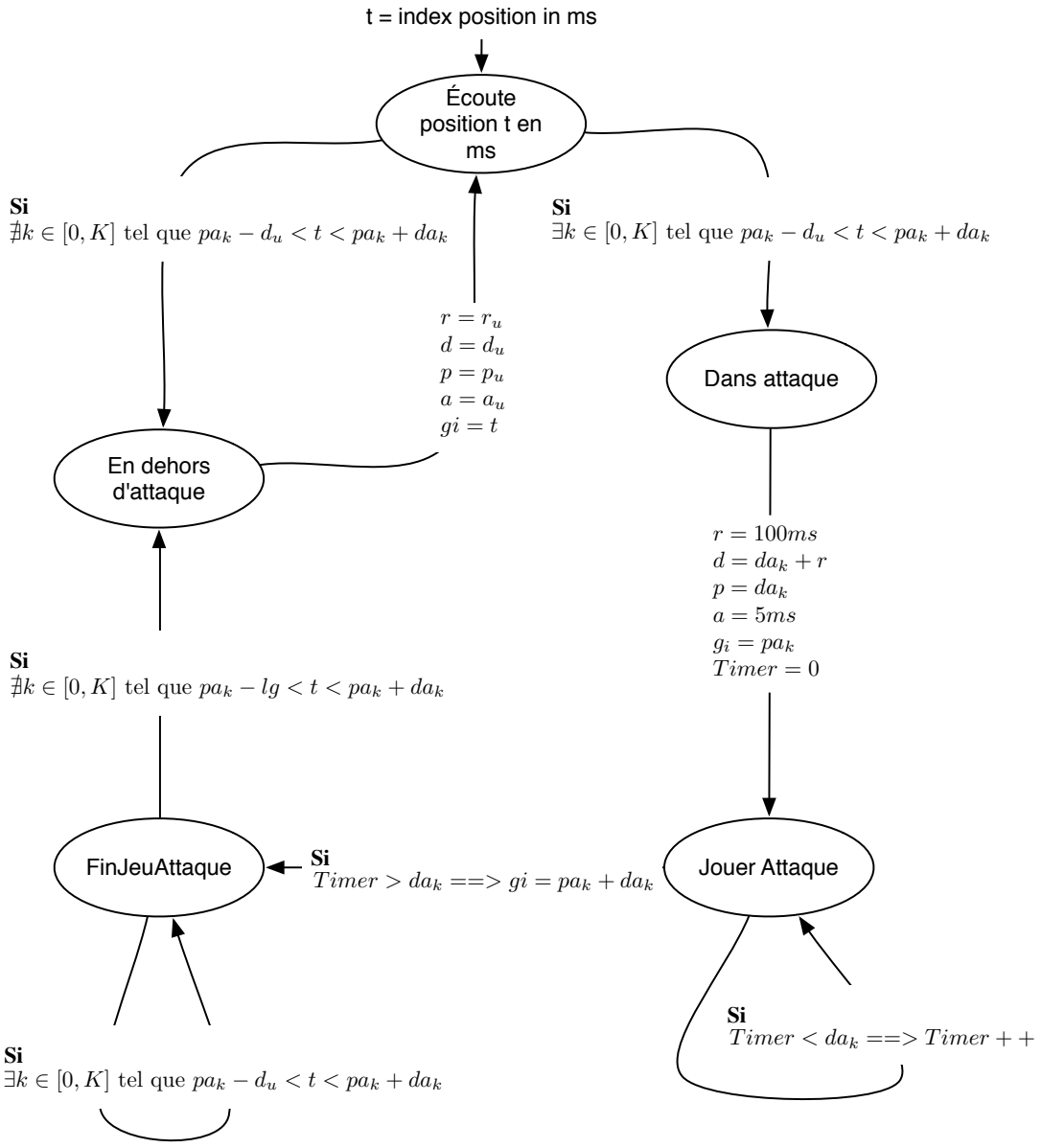


FIGURE 7 – Automate de fonctionnement du moteur de synthèse granulaire à conservation d'attaques.

L'avantage de l'utilisation de la synthèse additive dans le cadre de ce stage est donc d'une part que celle-ci est paramétrique et d'autre part qu'elle permet de synthétiser le contenu harmonique. Celle-ci étant paramétrique, il est possible d'apprendre et d'inférer directement les paramètres de contrôle de la synthèse par les HMM multimodaux du système d'apprentissage par démonstration.

Comme expliqué dans l'état de l'art, l'algorithme utilisé est celui présenté par Rodet en [21]. Dans l'étape d'analyse, le signal sonore est divisé en une partie harmonique (partiels caractérisés par l'amplitude la fréquence et la phase) et un signal résiduel contenant les parties bruités et percussives du son.

4.3.2 Intégration du système d'analyse-synthèse additive

Le synthétiseur utilisé est `mubu.additive~` implémenté sous Max par l'équipe ISMM qui permet de faire la synthèse d'un ensemble de partiels et d'un contenu résiduel en parallèle. Celui-ci est contrôlable via des fréquences et amplitudes de partiels.

Par ailleurs, l'équipe Analyse-Synthèse de l'IRCAM a développé un outil d'extraction et stockage de partiels et autres caractéristiques du son appelé PM2. Ce logiciel réalise l'extraction et suivi de partiels du son par l'utilisation des algorithmes présentés par Röbel [20].

Afin d'intégrer ces deux technologies dans un seul synthétiseur, les outils d'analyse et de synthèse additive ont été intégrés dans l'environnement Max : le suivi de partiel réalisé par PM2 est contrôlé dynamiquement depuis Max et communique avec le synthétiseur `mubu.additive~`.

Ensuite, le synthétiseur a été intégré dans un environnement capable d'enregistrer une performance gestuelle et sonore, l'analyser puis l'apprendre en utilisant les HMM multimodaux du système d'apprentissage par démonstration. Les données des paramètres sonores sont ensuite estimés en temps réel par rapport à une performance gestuelle afin de resynthétiser le son d'exemple par le geste. L'interface utilisateur de ce synthétiseur se trouve en annexe G

4.3.3 Synthèse des résidus

Le résidus est la partie du signal restante de l'analyse des partiels, celle-ci est représentée par la différence entre le signal d'origine et l'information de l'extraction des partiels. Les résidus sont un bruit coloré, extrait du signal original afin d'avoir le même contenu avec les partiels et résidus qu'avec le son original. Ils représentent donc tous ce qui n'est pas contenu dans l'extraction des partiels. La partie résiduelle est, suite à l'analyse, sous forme d'un fichier son

4.4 Le moteur de synthèse hybride additive-granulaire

4.4.1 Motivations

Le résultat de la synthèse des résidus est présenté sous forme d'un fichier audio, les possibilités de contrôle sont donc faibles. Le but du développement du synthétiseur hybride est d'apporter un contrôle distinct sur les parties harmoniques et bruitées, d'une part, par le contrôle paramétrique de la synthèse additive, d'autre part, par l'utilisation de la synthèse granulaire avec conservation d'attaques pour la partie résiduelle.

L'idée de réaliser une synthèse hybride composée des deux moteurs de synthèse développés est donc de pouvoir resynthétiser tant la partie harmonique d'un son que la partie percussive et bruitée en gardant les différentes méthodes de contrôle et les avantages de chacune des synthèse.

4.4.2 Structure du synthétiseur hybride

La synthèse granulaire à conservation d'attaques est la plus appropriée pour la synthèse de textures sonores et de contenus percussifs. Or, pour la synthèse additive + résidus de `mubu.additive~` la catégorisation du signal en une partie harmonique et une partie bruité est déjà réalisée. De ce fait, l'idée principale du synthétiseur hybride est d'utiliser le synthétiseur granulaire à conservation d'attaques pour la resynthèse de la partie résiduelle du son. La partie harmonique est synthétisée en utilisant `mubu.additive~` par le contrôle de partiels.

Cette structure permet une distribution des tâches de synthèse par la discrimination du contenu audio faite dans l'étape d'analyse, la figure 8 montre la structure de ce synthétiseur.

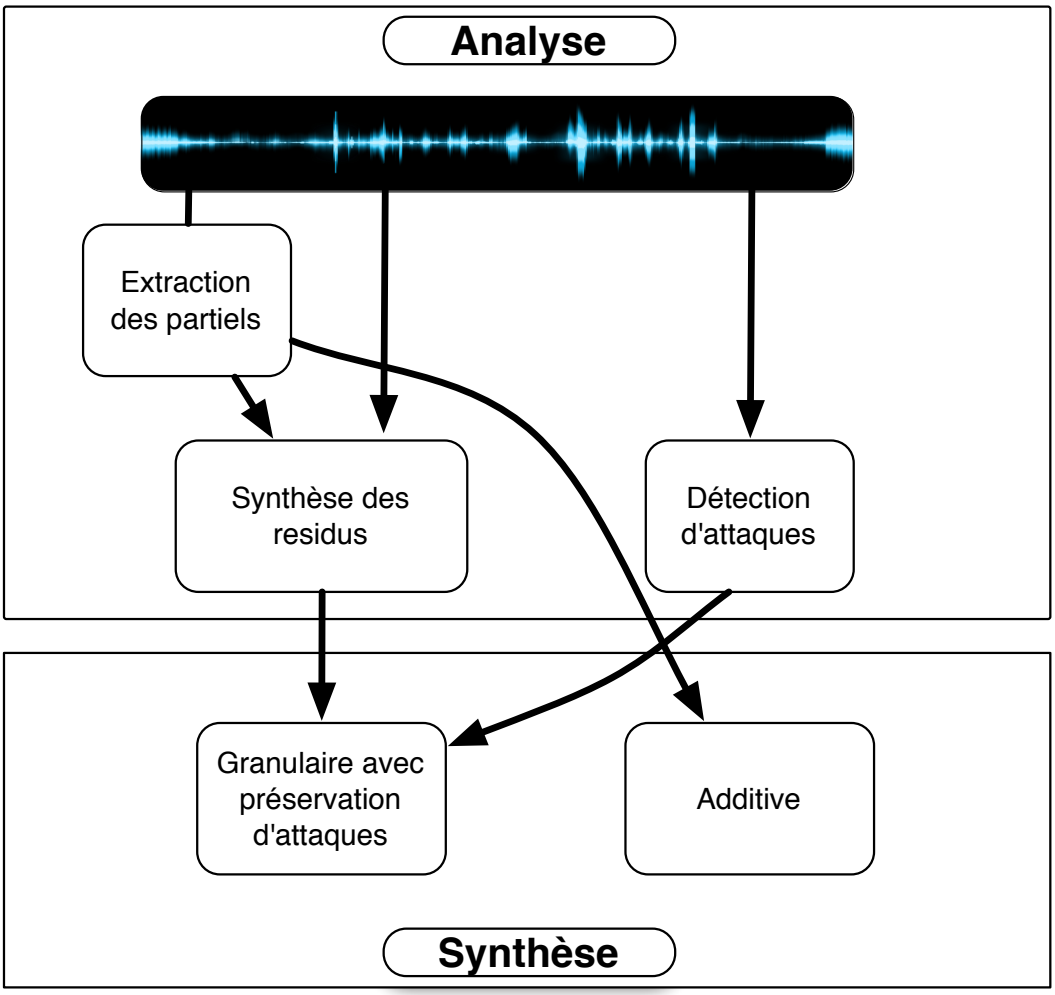


FIGURE 8 – Structure du synthétiseur hybride

5 Contrôle du moteur de synthèse par le système d'apprentissage mouvement-son par démonstration

5.1 Le but

Un des buts des développements autour du synthétiseur hybride présenté ci-dessus était de réaliser un moteur paramétrique afin de pouvoir exploiter les possibilités du système de mapping par démonstration. Nous cherchons à apprendre un mapping continu entre paramètres du mouvement et paramètres de contrôle de la synthèse sonore. Il s'agit donc d'une régression dont les paramètres évoluent dynamiquement selon un modèle de Markov caché.

5.2 Contrôle de la synthèse sonore par HMM

5.2.1 Estimation par HMM des fréquences et amplitudes de partiels

Une première stratégie de contrôle du système de mapping par démonstration consiste à contrôler la synthèse directement depuis les paramètres sonores. Dans ce cas, l'apprentissage est réalisé entre les paramètres décrivant le geste (par exemple l'accélération ou la vitesse 2D sur une tablette graphique), et les fréquences et amplitudes des partiels extraits de l'analyse. Lors de la phase de performance, les fréquences et amplitudes de partiels sont estimées en continu à partir du mouvement en entrée du système. Par ailleurs, la synthèse de la partie résiduelle est réalisée par le moteur de synthèse granulaire à conservation d'attaques. Celui-ci est contrôlé à partir de l'estimation de l'avancement temporel à l'intérieur du geste.

5.2.2 Estimation par HMM de la fréquence fondamentale et amplitudes de partiels

Pour des bruits ou transitoires le suivi de partiel n'est pas stable, il y a des incohérences entre les ordres des partiels à des instants successifs, ce qui crée des artefacts au moment de la synthèse.

De ce fait, afin d'éviter d'apprendre sur des partiels bruités, une autre méthode de contrôle est proposée. Tout d'abord, nous posons la contrainte que la partie additive ne réalise que la synthèse de contenu harmonique. Ainsi, il est possible d'éviter d'apprendre sur les parties bruitées des partiels.

L'idée de cette deuxième méthode est donc d'apprendre seulement sur la fréquence fondamentale et sur les amplitudes de partiels. La série harmonique des partiels peut être recréée par multiples du f_0 inféré. Les partiels déduit à partir du f_0 auront pour amplitude l'amplitude estimée à l'ordre correspondant.

6 Validation du système

6.1 Resynthèse pure - Test du moteur de synthèse

Cette partie vise à étudier la performance du synthétiseur hybride pour la resynthèse et pas pour le contrôle gestuel. Dans ce cadre là, la synthèse hybride implémentée diffère de la synthèse additive et résidus par l'utilisant de la synthèse granulaire avec conservation d'attaques pour la partie résiduelle. De ce fait, nous nous concentrons sur la validation de la conservation d'attaques et de la modulation temporelle du son.

À titre d'exemple, les figures 9a, 9c et 9b présentent une forme d'onde et ses resynthèses huit fois plus lentes avec et sans conservation d'attaques. Nous remarquons sur celles-ci la conservation des transitoires du son.

6.2 Évaluation de la régression sur les amplitudes et fréquences de partiels

6.2.1 Synthèse de contenu harmonique

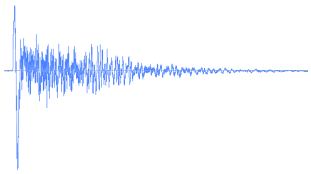
Erreur de la régression sur les amplitudes et les fréquences des partiels
Le but de cette partie est de mesurer la qualité du contrôle de la synthèse par HMM par le modèle d'apprentissage par démonstration sur un son d'exemple ayant principalement du contenu harmonique. Les HMM estiment les valeurs de fréquences et amplitudes de partiels.

La figure 10 présente un exemple d'estimation des partiels par HMM. Cette figure présente les partiels originaux et estimés, l'erreur moyenne de la régression puis la vocalisation utilisée. La vocalisation, est un son harmonique avec variation de hauteur. Afin d'évaluer la qualité de la synthèse dans cet example, la resynthèse est effectuée avec les même données gestuelles que l'exemple d'apprentissage.

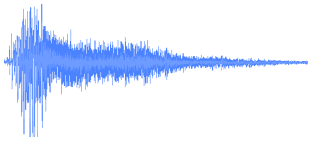
La figure 10 montre que les partiels originaux évoluent plus rapidement que les fréquences estimés par HMM. La régression joue un rôle de filtre passe bas et garde l'évolution à grands traits des partiels en perdant ainsi les mouvements rapides. La synthèse additive a comme but dans le synthétiseur hybride de resynthétiser le contenu harmonique et donc les sinusoïdes pseudo-stationnaires du son. Le suivi lent de ses paramètres est donc satisfaisant pour la synthèse de contenu harmonique.

La figure 11 présente la même expérience que la figure 10 mais avec les amplitudes de partiels. La resynthèse est effectuée avec les mêmes données gestuelles que l'exemple d'apprentissage.

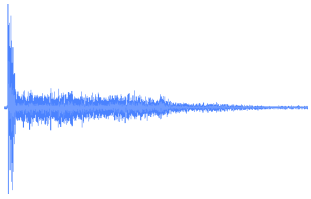
Les figures 10 et 11 montrent que la régression agit comme un filtre passe bas et que l'évolution des données estimés est plus lente que les données originales. L'évolution lente des paramètres estimés a pour conséquence l'apparition d'artefacts au niveau sonore que nous détaillerons en section 6.2.3.



(a) Forme d'onde du son original composer d'une attaque.



(b) Forme d'onde de la synthèse du son de la figure 9a sans conservation d'attaques. Resynthèse huit fois plus lente que le son original



(c) Forme d'onde de la synthèse du son de la figure 9a avec conservation d'attaques. Resynthèse huit fois plus lente que le son original

FIGURE 9 – Exemple de conservation d'attaques par la synthèse granualaire développée

Variabilité de l'erreur par rapport aux paramètres du HMM

La qualité de la régression est liée aux paramètres du HMM. Notamment au nombre d'états du HMM et à la *variance offset*. La *variance offset* représente un seuil minimal et absolu qu'on ajoute à la variance. Ceci afin que les gaussiennes soit plus ou moins larges afin d'être plus souple dans la reconnaissance. De ce fait, des gaussiennes trop larges ne permettent pas une estimation précise des paramètres. Par ailleurs, des gaussiennes trop proches, ne permettent pas de suivre le geste.

Les figures 12 et 13 présentent les différentes estimations pour un même geste en variant d'une part, le nombre d'états du HMM et d'autre part la *variance offset*.

La figure 12 montre que plus le nombre d'états est élevé plus la régression est précise, cependant, le nombre d'états du HMM est fortement lié au temps d'entraînement, ainsi, plus le nombre d'états est élevé plus l'entraînement est long. Cependant, le temps d'entraînement n'est pas le seul désavantage d'avoir un grand nombre d'états. En effet, avec beaucoup d'états le système fait du sur-apprentissage, il est donc plus proche de l'exemple d'origine comme le montre la figure 12 mais perd la capacité à suivre des gestes semblables au geste d'origine, la performance gestuelle doit alors suivre plus précisément l'exemple de l'apprentissage.

La figure 13 montre l'influence de la variation de la *variance offset* pour les paramètres

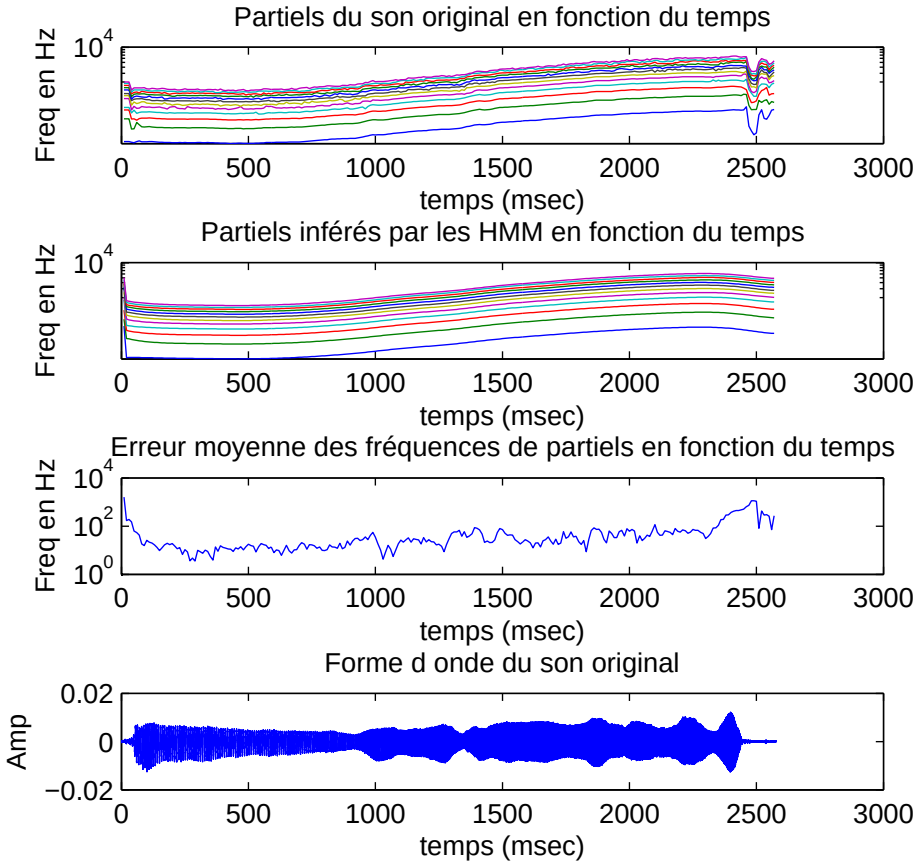


FIGURE 10 – Exemple d'estimation de fréquences de partiels par HMM et calcul d'erreur pour une vocalisation harmonique.

du premier partiel d'un son composé de deux vocalisations et un silence. Des courbes plus détaillées sur ces variations se trouvent en annexe B.

En ce qui concerne la *variance offset*, comme le montre la figure 13, il est nécessaire de trouver un compromis afin que les gaussiennes permettent de faire une estimation précise des paramètres sonores.

Par ailleurs, la courbe 12 montre que le système estime des amplitudes négatives en fonction du choix des paramètres. Pour le contrôle du synthétiseur ces amplitudes sont considérées comme étant à zéro.

6.2.2 Resynthèse de bruit et de transitoires

La synthèse du résidu se fait par la partie granulaire avec conservation d'attaques. Celle-ci est contrôlée par un indice temporel estimé par le HMM. À chaque nouvelle observation,

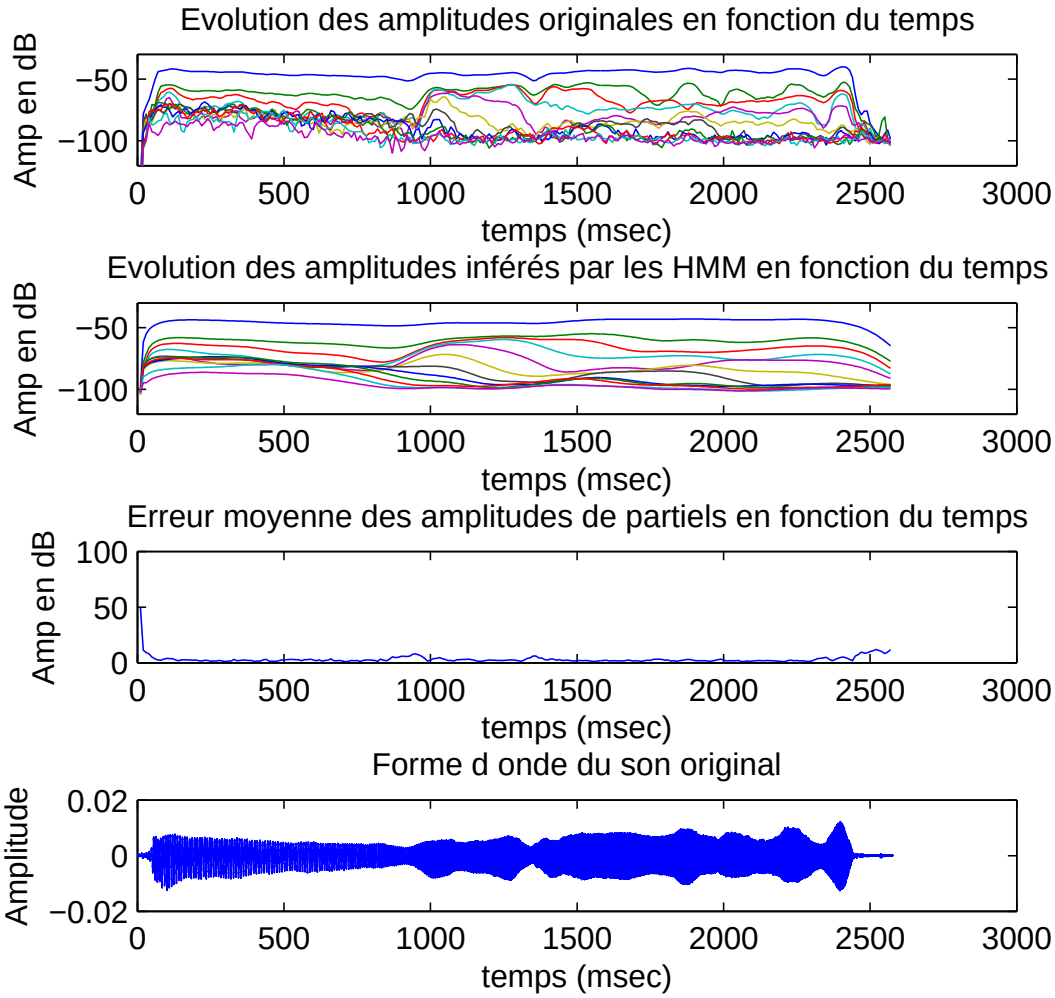


FIGURE 11 – Exemple d'estimation des amplitudes de partiels pour une vocalisation harmonique par le système d'apprentissage mouvement-son par démonstration.

le modèle estime par un algorithme "forward" la distribution de probabilité sur les états. L'avancement temporel relatif peut alors être estimé comme l'espérance mathématique de la distribution de probabilité sur les états.

Le synthétiseur hybride ne prévoit pas de resynthétiser les résidus (contenu bruité et transitoire) par la partie additive, de ce fait, la régression sur les partiels ne prévoit pas de contrôler la resynthèse de contenu bruité. Nous verrons cependant, que des artefacts sonores sont créés du fait de l'apprentissage sur des partiels modélisant du bruit.

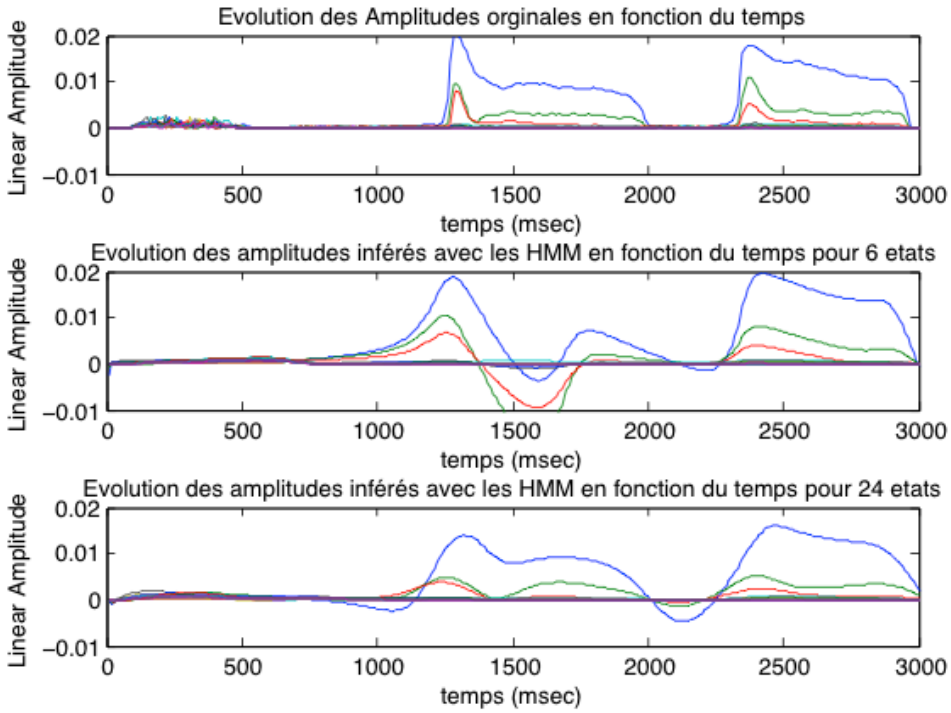


FIGURE 12 – Influence du nombre d'états du HMM sur la qualité de la régression.

6.2.3 Résultats de la régression sur les fréquences et amplitudes de partiels

La régression sur les partiels se comporte tant pour les amplitudes comme pour les fréquences comme un filtre passe bas. Par ailleurs, la vitesse d'évolution et les paramètres du HMM influencent la qualité de la régression. De ce fait, une première limite du système est le suivi de paramètres variant trop rapidement dans le temps. Le modèle n'estime pas les changements abrupts de paramètres du fait de l'évolution lente des signaux estimés. Pour illustrer ceci l'annexe A présente les fréquences et amplitudes des partiels originaux et estimés par le HMM pour un son composé de deux vocalisations harmoniques séparées par un silence. La figure 14 synthétise et met en évidence ce phénomène.

La figure 14 montre à partir de 1200 ms le comportement de la régression lorsque les partiels ne suivent pas des sinusoïdes pseudo-stationnaires mais ne représentent que du bruit. Dans ce cas, du fait du filtrage passe-bas les partiels estimés suivent une trajectoire qui ne représente pas le contenu spectral du son.

Le problème se pose lors du passage d'un son bruité donc sans partiels vers un son harmonique avec des amplitudes audibles. En effet, lors d'un tel passage, les fréquences de partiels se stabilisent lentement de même que les amplitudes augmentent. Ceci a pour conséquence un effet de "glitch" dans le passage d'un moment bruité du son à un moment harmonique ou d'un silence à une partie harmonique.

La méthode de régression sur les fréquences et amplitudes de partiels est donc fidèle au son original si celui-ci ne comporte pas de changements abrupts des paramètres sonores.

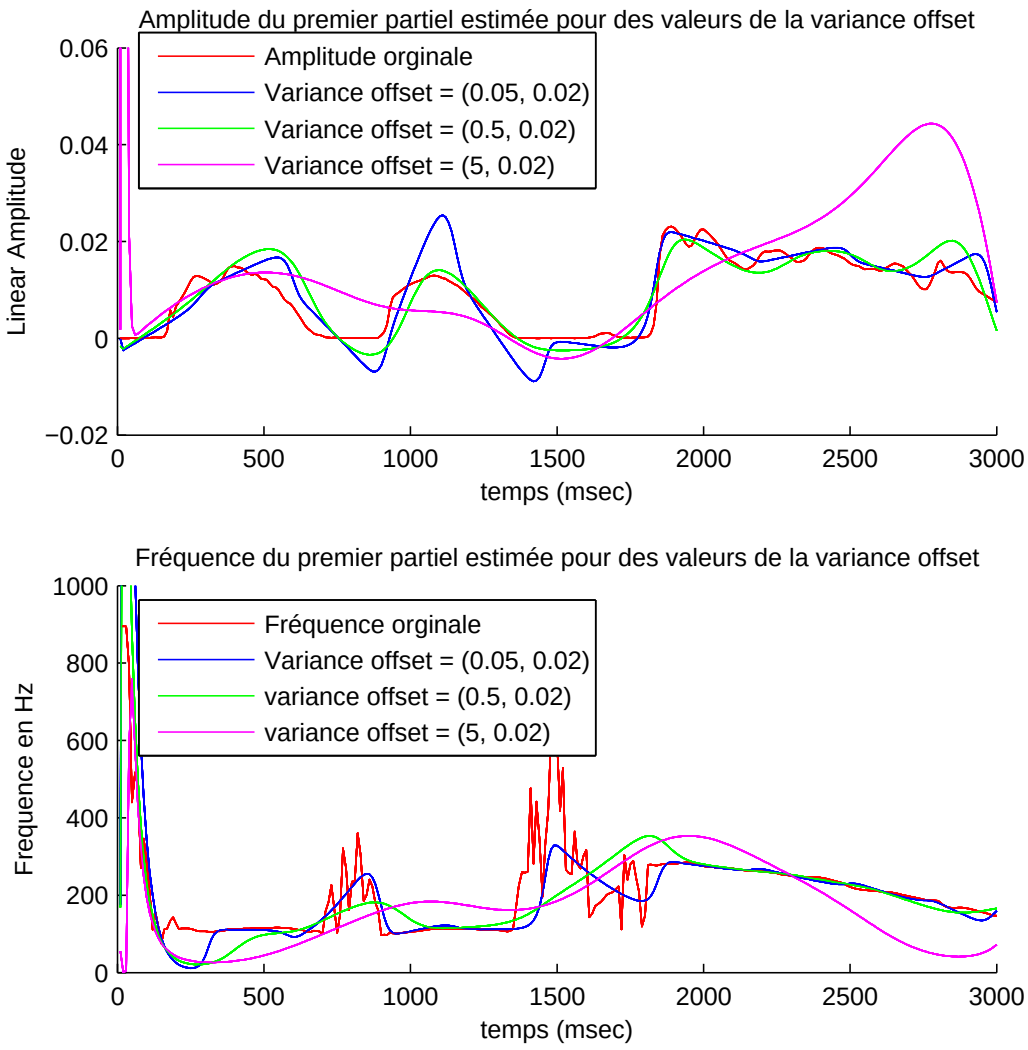


FIGURE 13 – Exemple de l'influence de la variation de la *variance offset* sur l'amplitude du premier partiel (haut) et sur sa fréquence (bas) pour vingt-quatre états

Pour cette première méthode le problème du synthétiseur hybride repose sur la concordance des deux synthèses. En effet, chacune des deux synthèses resynthétise proprement le contenu pour laquelle elle a été faite. Cependant, les artefacts introduits dans les fréquences de partiels par la régression lors de la resynthèse de matières sonores pas harmoniques sont facilement perceptibles. Afin de combler ceci, nous proposons une méthode basée sur l'estimation sur la fréquence fondamentale du son.

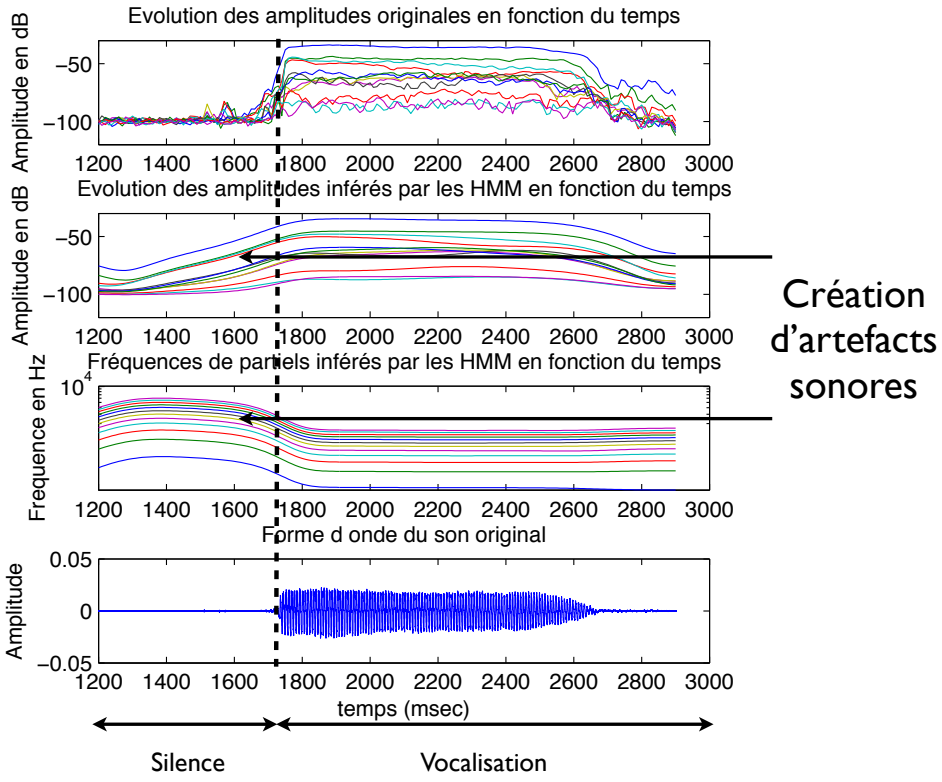


FIGURE 14 – Mise en evidence de la génération d'artefacts par la resynthèse de paramètres variant rapidement dans le temps.

6.3 Performances et erreurs de la régression sur le f_0 et les amplitudes de partiels

Cette deuxième méthode cherche principalement à supprimer les artefacts sonores introduits par la régression sur les fréquences de partiels à cause de l'apprentissage sur des partiels bruités. La méthode proposée est de recréer en temps réel la série harmonique suite à l'inférence de f_0. Nous supposons donc que la synthèse additive, contrôlée par les HMM ne synthétise que du contenu harmonique.

Les paramètres d'apprentissage utilisés pour ce deuxième modèle sont N amplitudes de partiels et un f_0. La figure 15 montre la fréquence fondamentale originale et estimée pour les mêmes données lors de l'apprentissage et lors du jeu. Celle-ci montre la série harmonique générée avec laquelle est contrôlé le synthétiseur hybride.

La régression sur les amplitudes de partiels se comporte de façon similaire à celle observée pour la première méthode (Figure 11) celle-ci n'étais pas à l'origine de la création des artefacts sonores.

En utilisant cette deuxième méthode les fréquences de partiels sont plus stables dans le temps et assurent l'harmonicité de la resynthèse. Celle-ci ne contient plus les artefacts

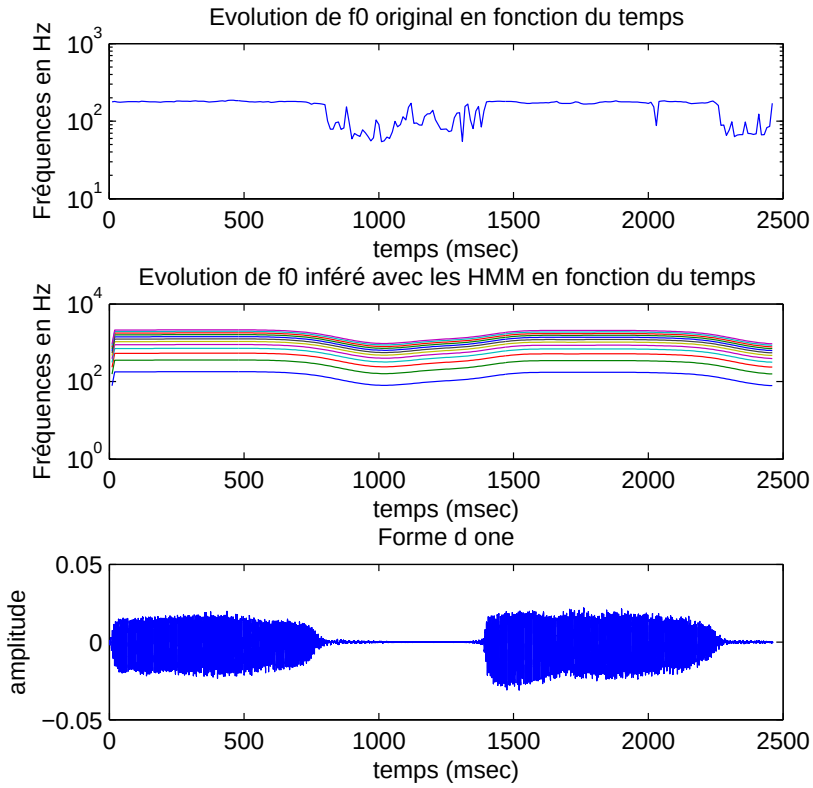


FIGURE 15 – Résultats de la régression sur f0 sur un son composer de contenu harmonique et de bruit. F0 inféré par HMM (bas) et f0 original (haut)

sonores remarqués pour la première méthode.

Avantages et inconvénients

Comme montré dans l'étude sur la régression, les artefacts introduits dans le son ne sont pas dus au moteur de synthèse mais à la régression elle-même. Cette deuxième méthode a été introduite pour résoudre les problèmes d'apprentissage sur des partiels bruités. En tant que synthèse la première méthode est optimale car elle ne crée pas une série artificielle de partiels harmoniques.

Cependant, pour le contrôle par le système d'apprentissage par démonstration, la méthode par f_0 est plus fiable car elle n'introduit pas d'artefacts lors de la variation rapide de paramètres sonores. Les artefacts remarqués avec le première méthode ne sont plus présents.

Le système perd cependant en réalisme de voix à cause de la création artificielle de la série harmonique des partiels. En effet, le timbre du son resynthétisé devient moins naturel. La voix du locuteur reste tout de même reconnaissable.

6.4 Contrôle gestuel

Cette partie cherche à évaluer les synthèses développées dans le cadre du contrôle gestuel. Pour ce faire, un test perceptif a été organisé. La suite présente une description de l'expérience, le protocole expérimental et les résultats obtenus.

6.4.1 Expérience

Cette partie décrit l'expérience visant à évaluer qualitativement les méthodes de synthèse et de contrôle. Le périphérique utilisé pour la captation du geste est une tablette à deux dimensions contrôlée par un stylo.

Quatre conditions ont été sélectionnées, elles sont présentées dans le tableau 2 :

TABLE 2 – Synthèses et méthodes de contrôle utilisés pour le test perceptif

Condition	**Synthèse sonore**	**Stratégie de contrôle**
1	Synthèse granulaire	Indice temporel
2	Synthèse granulaire avec conservation d'attaques	Indice temporel
3	Synthèse hybride	Régression sur les fréquences et amplitudes de partiels + indice temporel
4	Synthèse hybride	Régression sur la fréquence fondamentale et amplitudes de partiels + indice temporel

L'idée de cette expérience est de noter la performance de chaque synthèse par rapport à différents paramètres du son et du geste. Le questionnaire est composé de deux parties principales.

Dans la première partie, le participant doit rejouer un geste pour contrôler la synthèse d'un son. Le son et le geste sont pré-enregistrés. Ils doivent ensuite noter chaque synthèse sur une échelle de Likert à cinq points pour différentes questions :

— Qualité de la synthèse (Absence de sons indésirables et d'artefacts)
— Respect du son original
— Contrôle par le geste et lien geste-son

Cette expérience se fait trois fois, les différents gestes et sons sont les mêmes pour les différents sujets. Les sons choisis mettent en valeur les différences entre les différentes synthèses (vocalisations harmoniques, vocalisations avec attaques). Les gestes utilisés sont présentés en figure 16.

La deuxième partie de l'expérience est la démonstration en simultané d'un son et d'un geste par le sujet. Le participant doit alors rejouer le son par le geste et donner une note sur le respect de la relation geste-son démontrée et sur le contrôle du son par le geste.

Protocole expérimental

L'expérience se déroule de la même façon pour les différents participants. Premièrement, une explication de l'expérience au sujet et de la notation sur les synthèses est faite. Ensuite, le sujet commence la comparaison entre les synthèses. Celle-ci se fait au même niveau sonore,

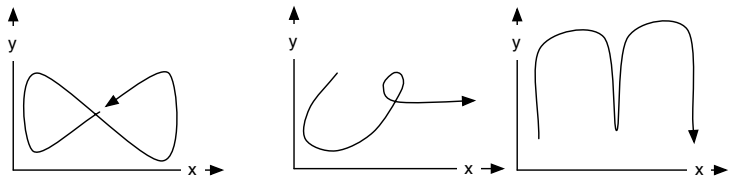


FIGURE 16 – Gestes de la base utilisés pour l'expérience.

les candidats sont libres d'écouter le son original et les différentes resynthèses à tout moment. Aucune précision n'est donnée en cours d'expérience sur les différences entre les synthèses.

Pour chaque question les sujets sont invités à :

— Écouter le son original

— Contrôler les quatre différentes synthèses par le geste

— Donner une note à chaque synthèse

Les sujets qui ont passé le test sont des personnes travaillant à l'IRCAM dans différents domaines des technologies musicales et audio. La moyenne d'âge est de 24 ans. Chaque expérience dure environ vingt minutes. Dix sujets ont été interrogés, pour un total de 130 notes pour chaque synthèse.

Le protocole expérimental et le questionnaire de l'expérience se trouvent en annexes C et D.

6.4.2 Analyse Anova pour l'analyse statistique des données

L'analyse de la variance aussi appelée Anova est un test statistique permettant d'étudier la distribution d'une variable aléatoire. Dans ce cas nous cherchons à étudier si les données recueillies par le test perceptif sont significatives ou non. Pour l'analyse Anova nous supposons que les populations suivent des distributions normales, qu'elles ont une variance égale et que les notes entre les synthèses sont indépendantes. Le modèle est décrit par l'équation :

$$y_{ij} = \alpha_i + \epsilon_{ij} \tag{10}$$

où α_i est la moyenne de la distribution et ϵ_{ij} est l'erreur. La matrice y_{ij} est composée par exemple de toutes les réponses sur la qualité sonore sur les lignes i et pour chaque synthèse sur les colonnes j.

La moyenne pour chaque distribution s'écrit,

$$\overline{y_{i.}} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \sim \mathcal{N}\left(\mu_i, \frac{\sigma^2}{n_i}\right) \tag{11}$$

où σ^2 est la variance et μ_i la moyenne.

La somme des carrés des écarts peut être calculée par la formule :

$$SCE_{\text{total}} = SCE_{\text{facteur}} + SCE_{\text{residu}} \tag{12}$$

Le paramètre SCE_{facteur}, aussi appelée "variabilité inter-classe" plus bas dénominé "SS columns" est donné par :

$$SCE_{\text{facteur}} = \sum_{i=1}^{p} n_i (\overline{y_i} - \overline{y})^2 \tag{13}$$

la part de SCE_{total} qui ne peut être expliquée par le modèle (erreur ou résidus) est donné par :

$$SCE_{\text{residu}} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \overline{y_i})^2 \tag{14}$$

Par ailleurs, les DDL (degrés de liberté) sont aussi calculés. Ceux-ci représentent le nombre de variables aléatoires qui ne peuvent être déterminées ou fixées par une équation (notamment les équations des tests statistiques).

$$\begin{aligned} DDL_{\text{facteur}} &= p - 1 \\ DDL_{\text{residu}} &= n - p \end{aligned} \tag{15}$$

où n est le nombre de colonnes et p est le nombre total d'individus.

Ces valeurs permettent de calculer la variable F. Si celle-ci est supérieure au seuil de rejet alors il existe une différence statistiquement significative entre les distributions. Sinon les résultats ne sont pas significatifs et nous ne pouvons pas tirer des conclusions sur les échantillons. F se calcul par :

$$F = \frac{\frac{SCE_{\text{facteur}}}{DDL_{\text{facteur}}}}{\frac{SCE_{\text{total}}}{DDL_{\text{total}}}} \tag{16}$$

La valeur p montre explicitement si F est supérieur ou non au seuil de rejet, en pratique, si $p < 0.05$ les valeurs sont significatives. La suite présente les analyse Anova ainsi que les moyennes des tests perceptifs réalisés.

6.4.3 Résultats

Suite à l'analyse Anova des données trois ensembles de réponses sont significatifs, celles concernant la qualité du son (présence d' artefacts, sons indésirables), le respect du son original et les tests sur des sons harmoniques. Ces résultats sont présentés en figure 17, 18 et en annexe E.1.

La figure 17 montre la note moyenne et l'écart type pour les données recueillies pour les trois questions concernant la qualité du son. Cette question cherchait à mesurer la perceptibilité d'artefacts ou d'éléments indésirables dans le son resynthétiser lors du contrôle par le geste. Les données des sons utilisés pour ces résultats sont des vocalisations harmoniques avec des attaques. Ces résultats sont significatifs. En effet, $p = 0.008$ pour le test Anova. Par ailleurs, un T-test présenté dans le tableau 3 a été réalisé afin de savoir si les expériences ne reflètent pas simplement des mesures aléatoire. Selon le tableau 3, les différences se situent principalement entre les synthèses granulaires et les synthèses hybrides. Il y a donc une différence significative entre les synthèses hybrides développés et les synthèses granulaires. Nous pouvons conclure qu'au niveau de la qualité sonore les synthèses avec les partiels ont un meilleur rendu sonore. Selon les sujets, la synthèse avec le meilleur rendu sonore est la synthèse hybride contrôlée par estimation de la fréquence fondamentale et les amplitudes.

Il est important de remarquer que les T-test pour la qualité sonore et pour le respect du son original montre les mêmes résultats, de ce fait, ils sont tous les deux regroupés dans le même tableau 3.

D'autre part, les résultats sur le respect du son original sont aussi significatifs ($p = 0.0006$). La figure 18 montre des résultats similaires à ceux sur la qualité sonore. Les T-tests sont présentés dans le tableau 3

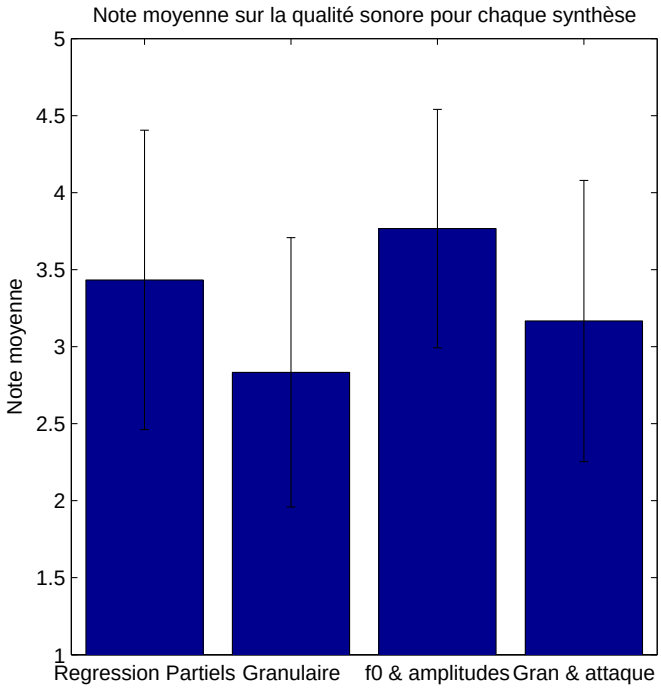


FIGURE 17 – Note moyenne sur la qualité sonore (artefacts, sons sindésirables) pour chaque synthèse

TABLE 3 – T-test pour les données sur la qualité sonore et le respect du son original. (• signifie qu'il y a une différence qui n'est pas due à l'aléatoire).

	Granulaire	Granulaire & attaques
Régression Partiels	•	
Régression sur f_0	•	•

Encore une fois, les différences sont marqués entre les synthèses hybrides et les synthèses granulaires. Les deux synthèses hybrides sont considérées comme plus respectueuses du son original, la synthèse contrôlée par f0 et amplitudes est, selon les sujets, encore une fois la plus performantes.

Les résultats non significatifs de cette enquête sont principalement les résultats sur le lien geste-son et sur les notes sur les sons composés d'attaques. La figure 29 présentée en annexe E.2 montre les résultats sur le lien geste-son par rapport aux différentes synthèses, résultats qui ne sont pas significatifs.

On peut cependant dire que pour les différents sujets il n'y a pas de différences importantes en ce qui concerne le lien geste-son entre les synthèses. Ceci rejette l'hypothèse que la conservation des éléments percussifs dans le son est fondamentale dans la conservation de ce lien. Ces résultats peuvent être dus au fait que les sujets n'étaient pas au courant de

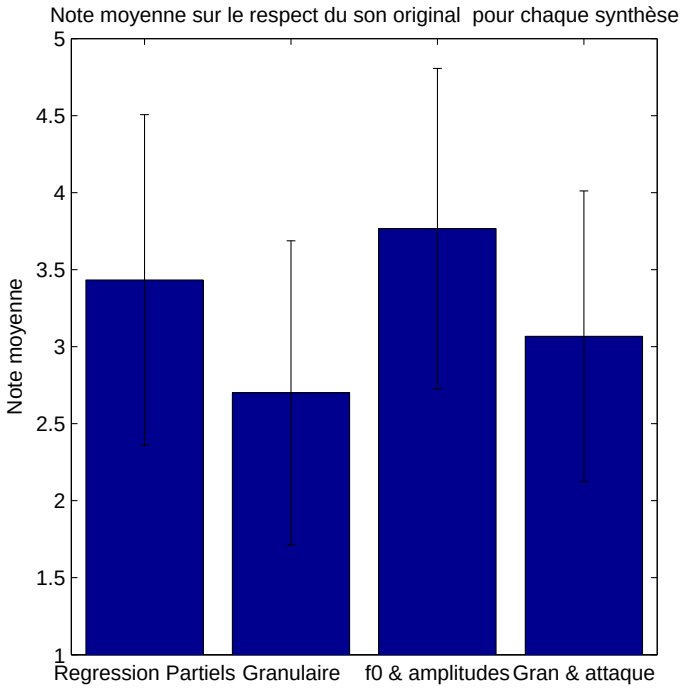


FIGURE 18 – Note moyenne sur le respect du son original pour chaque synthèse

cette différence et qu'elle n'a pas été remarquée par un grand nombre. Cependant, la figure 19 montre une tendance intéressante sur laquelle nous remarquons que la conservation d'attaques est bien perceptible et appréciée par les sujets. Ces résultats montrent cependant seulement des tendances et ne sont pas significatifs.

Cette enquête reste une mesure qualitative des performances des synthèses afin de remarquer la différence perceptive entre les synthèses.

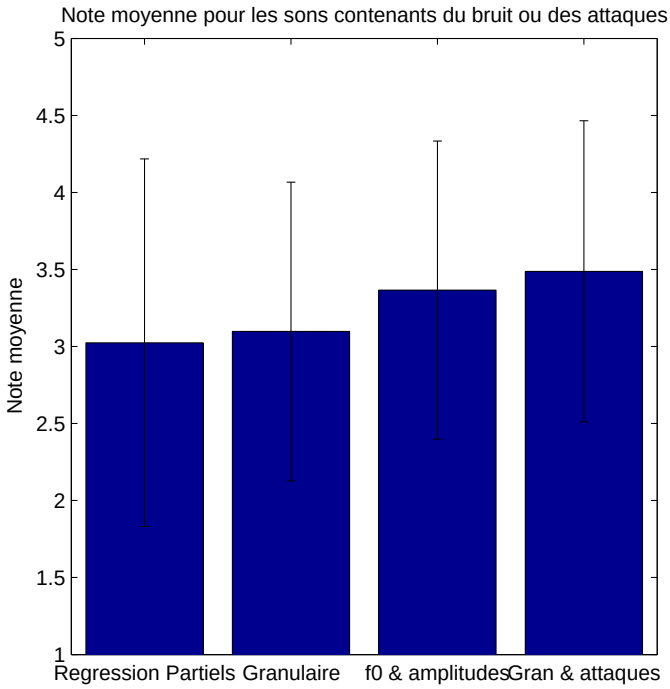


FIGURE 19 – Note moyenne pour les sons composés d'attaques

7 Extensions du contrôle par le geste du moteur de synthèse

Le moteur de synthèse est donc contrôlable par le geste cependant, l'expressivité du système reste limitée. En effet, dans un but artistique il serait intéressant d'avoir un contrôle plus large au niveau sonore que la seule modulation temporelle d'un ou de plusieurs gestes. Dans cette optique, les travaux qui ont suivi le développement du moteur de synthèse ont visé à générer un espace sonore continu, contrôlable par le geste et appris par démonstration.

Pour ce faire il faut d'une part définir et générer un "son intermédiaire" à partir de sons de référence. D'autre part, il faut pouvoir définir un geste intermédiaire afin de le reconnaître et le suivre.

Cette partie présente une méthode d'interpolation paramétrique pour la création d'un son intermédiaire et développe la question du geste intermédiaire.

7.1 La création d'un son "intermédiaire"

Nous définissons un son intermédiaire comme un son composé des caractéristiques de deux sons de base. Dans le cadre de l'utilisation du moteur de synthèse la génération d'un son intermédiaire à partir de deux exemples se fait par l'interpolation des paramètres de synthèse, notamment des amplitudes et des fréquences de partiels.

7.1.1 Création d'un son intermédiaire par HMM

Une façon de générer un son intermédiaire à partir d'exemples sonores est l'utilisation directe du système de mapping par démonstration. En effet, le système basé sur un HMM multimodal modélise conjointement des performances sonores et gestuelles. Le modèle implémente à chaque état une régression entre les paramètres gestuels et les paramètres sonores. Le système peut donc, dans une certaine mesure, capturer les corrélations entre mouvement et son sur différentes variations gestuelles en lien avec des variations sonores. Lorsqu'un nouveau geste est joué en entrée, le modèle estime continûment les paramètres sonores associés, ce qui devrait permettre d'interpoler les paramètres sonores en fonction des variations gestuelles en entrée. Cette méthode a été explorée au cours de ce stage sans résultats satisfaisants, la méthode utilisée est en section 7.1.2.

7.1.2 Création d'un son intermédiaire par interpolation paramétrique

Pour chaque son de référence, nous disposons après l'analyse de deux matrices : une matrice composée des amplitudes A_i de partiels et une deuxième composée des fréquences F_i de partiels où i est l'index du son de référence. Chacune de ces matrices est de taille $M \times N$ où N est le nombre de partiels utilisés pour la synthèse et M le nombre d'échantillons de l'analyse. La fréquence d'échantillonnage est de 100 Hz.

L'interpolation paramétrique vise à synthétiser un son qui soit composé de caractéristiques de deux sons de référence pour lesquels les paramètres de synthèse, comme les partiels, sont connus. L'interpolation de N buffers d'exemples sonores est définie par l'interpolation linéaire des paramètres des partiels, comme indiqué en équation 17 :

$$\begin{aligned} A_{total} &= \sum_{i=1}^{N} w_i A_i \\ F_{total} &= \sum_{i=1}^{N} w_i F_i \end{aligned} \tag{17}$$

où w_i sont les poids de chaque buffer et A_{total} et F_{total} sont les amplitudes et fréquences résultantes de l'interpolation.

Chaque valeur du vecteur de poids w_i est une valeur de 0 à 1 qui représente l'importance du buffer i dans le résultat de l'interpolation. Les poids sont normalisés, de sorte que $\sum_{1}^{N} w_i = 1$. Plus le poids du buffer i est important plus ces paramètres seront importants dans l'interpolation.

7.2 Contrôle du son intermédiaire par le geste

Il existe donc plusieurs méthodes de créer et générer des sons intermédiaires à partir d'exemples en utilisant le moteur de synthèse développé. Il s'agit maintenant de contrôler cette synthèse par le geste afin de créer un espace de trajectoires avec, comme résultat sonore, un son qui varie d'un exemple vers l'autre.

7.2.1 Le geste intermédiaire

Le but de l'étude du geste intermédiaire est principalement de pouvoir contrôler l'interpolation présentée précédemment par le geste. Cependant, plusieurs problématiques émergent de cette question. En effet, le geste intermédiaire ne peut pas être défini facilement. Le but de ces études est de définir un geste intermédiaire qui puisse être décrit par un sujet sans qu'il soit démontré explicitement.

Ainsi, il s'agit de définir une trajectoire qui soit le résultat d'un mélange entre deux trajectoires. Suivant cette définition, il est difficile de dire quel est le geste intermédiaire pour n'importe quel geste. De ce fait, nous nous limitons à des gestes d'exemples simples pour lesquels la définition du geste intermédiaire est évidente comme en figure 20. Le périphérique de captation gestuelles est ici une tablette 2D, les paramètres sont les axes x et y. Nous étudions donc le cas de trajectoires d'exemples simples qui partent d'un même point, ont une même durée et une même direction.

7.2.2 La vraisemblance pour le contrôle de l'interpolation

Le modèle de mapping par démonstration assimile des trajectoires gestuelles jouées à des trajectoires apprises et les compare pour générer les paramètres de synthèse. Lorsque le système apprend plus d'un exemple la comparaison entre les gestes permet de calculer la vraisemblance à chaque geste. La vraisemblance est le paramètre du modèle symbolisant la

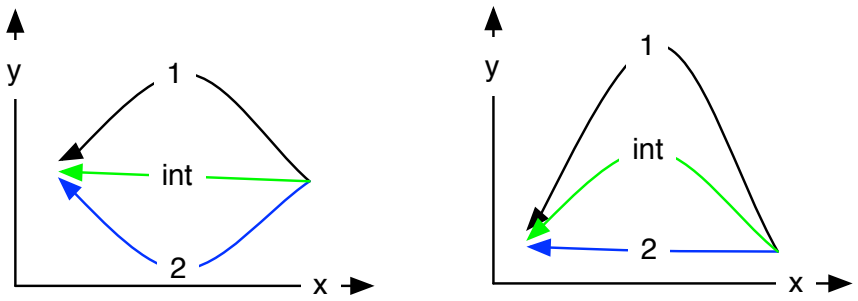


FIGURE 20 – Exemples de gestes intermédiaires (int) à partir de deux exemples (1, 2).

ressemblance du geste joué en temps réel aux gestes de la base. Plus la vraisemblance pour un geste est élevée plus le geste joué ressemble au geste appris.

Un première stratégie consiste à utiliser le paramètre de vraisemblance pour contrôler l'interpolation sonore, auquel cas la variation de la vraisemblance fait varier les poids w_i de chaque buffer.

Cependant, le problème de l'utilisation de la vraisemblance est que celle-ci ne reflète pas des trajectoires intermédiaires comme on les décrirai intuitivement (Figure 20) de ce fait, un mapping direct entre les poids des buffer pour l'interpolation et la vraisemblance n'as pas de sens au niveau perceptif. En effet, dans le cas du modèle de reconnaissance gestuelle utilisé, les probabilités sont normalisées globalement à l'ensemble des gestes, ce qui implique un contraste important des vraisemblances. Il est donc difficile d'obtenir de manière stable des vraisemblances représentant des variations intermédiaires entre les gestes d'exemple.

7.2.3 La géométrie pour la définition et l'apprentissage

Le problème du contrôle de l'interpolation réside donc dans la reconnaissance du geste intermédiaire par le système d'apprentissage par démonstration. En effet, la vraisemblance ne reflète pas l'espace intermédiaire pour les gestes simples considérés.

De ce fait, afin que le système puisse d'une part reconnaître et d'autre part suivre le geste intermédiaire il est nécessaire que le système l'apprenne.

Nous utilisons donc la géométrie pour calculer le geste intermédiaire pour ensuite l'apprendre au modèle. Ainsi, pour deux gestes x et y composés de N échantillons, le geste intermédiaire est une moyenne entre les échantillons des gestes d'exemple $g(n) = \frac{y(n)+x(n)}{2}$, $n \in [1; N]$. Cette définition simple a pour avantage de suivre l'intuition cartésienne définie en figure 20.

Nous calculons donc la trajectoire intermédiaire en temps différé. Ainsi, le système peut être entraîné avec cette trajectoire. Les poids des buffers w_i pour l'interpolation sont modulés en temps réel par rapport au geste reconnu par le système.

Ce système a été implanté dans un patch max, qui génère après l'enregistrement de deux gestes et leur vocalisation un ensemble de gestes intermédiaires. Ainsi, le modèle d'apprentissage par démonstration peut s'entraîner tant sur les gestes d'exemple que sur les gestes calculés. De ce fait, le système est capable de suivre le geste intermédiaire ainsi que les gestes d'exemple et contrôler l'interpolation par poids de buffers expliquée en section 7.1.2.

Cette solution est satisfaisante mais pose le problème de la perte de la définition continue du geste intermédiaire, une motivation pour les recherches futures.

8 Conclusion et perspectives

8.1 Travail accompli

Pendant ce stage a été développé un moteur de synthèse hybride et paramétrique qui combine synthèse additive et granulaire avec conservation d'attaques dans le but d'être contrôlé par le système d'apprentissage par démonstration.

Dans un premier temps, une extension de la synthèse granulaire a été faite pour pouvoir resynthétiser les attaques du son original tout en synthétisant le reste du signal comme la synthèse granulaire classique par rapport à une durée de grain et une enveloppe d'amplitude.

D'autre part, un travail d'intégration analyse-resynthèse a été réalisé pour créer un synthétiseur à synthèse additive capable de réaliser l'analyse et la synthèse dans un même module.

Par la suite les deux synthétiseurs sont utilisés en parallèle en tant que synthétiseur hybride capable de resynthétiser tant des contenus harmoniques comme des contenus transitoires et bruités. Dans la phase d'analyse le signal à resynthétiser est séparé en deux, une partie harmonique et une partie résiduelle. La synthèse additive resynthétise les partiels du son, la partie granulaire avec conservation d'attaques les résidus.

Le but de ces travaux était de créer un moteur de synthèse paramétrique pour lequel il soit possible d'utiliser des paramètres sonores pour le contrôle de la resynthèse. Ceci, afin d'explorer les possibilités de contrôle par le geste en utilisant la méthode d'apprentissage par démonstration.

Une intégration du synthétiseur au modèle de suivi gestuel a donc été faite. Dans un premier temps le synthétiseur est contrôlé par l'estimation de fréquences et amplitudes de partiels par HMM.

Cette méthode est satisfaisante pour la modulation temporelle et l'interpolation paramétrique. Cependant, lors de l'estimation des partiels par HMM, la régression des fréquences introduit des artefacts indésirables à cause de l'apprentissage sur des données variant rapidement dans le temps. Notamment, l'apprentissage sur des silences ou du bruit et non-pas des sinusoïdes.

De ce fait, nous proposons une méthode adaptée aux variations paramétriques lentes de la régression, celle-ci repose sur une inférence de la fréquence fondamentale et des amplitudes de partiels. La série harmonique de partiels est recréée par les multiples de f_0, les amplitudes estimées sont couplées aux fréquences de partiels calculées afin de recréer le timbre de la vocalisation.

Par ailleurs, des tests perceptifs ont été réalisés afin de comparer les différentes méthodes de synthèse par le contrôle gestuel. Les résultats statistiquement significatifs de l'étude montrent que la méthode par f_0 a été préférée par les participants en ce qui concerne la qualité sonore et le respect du son original.

Une des motivations du stage était la création d'un espace sonore continu contrôlable par des trajectoires gestuelles définies par l'exemple. Ceci nécessite d'une part un travail sur

la création d'un son intermédiaire et d'autre part une méthode de création, reconnaissance et suivi du geste intermédiaire.

Pour la création du son intermédiaire, la solution adoptée consiste à combiner les paramètres sonores des exemples fournis par l'utilisateur. Ceci se fait par interpolation de paramètres de synthèse : amplitude, fréquences et résidus. Au niveau sonore l'interpolation est satisfaisante pour des matières sonores adaptées.

La solution de contrôle par le geste trouvée est la synthèse d'un geste intermédiaire par moyenne cartésienne de trajectoires. L'entraînement du HMM se fait alors avec les gestes d'exemples et le geste intermédiaire. Du fait de sa définition dans le plan, la création de gestes intermédiaires est évidente pour des gestes simples commençant au même endroit et avec une même direction.

8.2 Perspectives

8.2.1 Extensions du système d'apprentissage par démonstration

Le système d'apprentissage par démonstration avec la synthèse hybride peut être étendu par le geste ou par le son. D'une part, il serait intéressant de pouvoir commencer à jouer le geste à n'importe quel instant t de celui-ci. En plus, pour étendre le contrôle sur le son, le système doit être capable de suivre un geste dans la temporalité à laquelle il a été démontré mais aussi inversement.

Pour ce faire, la technique pour le calcul du geste intermédiaire, dans laquelle les gestes et paramètres sonores sont calculés en temps différé, peut être utilisée. Il serait donc possible de faire une catégorisation du geste appris en plusieurs étapes significatives du lien geste-son et les apprendre au modèle. Enfin, en ce qui concerne le contrôle par le geste il serait intéressant que les micro-variations gestuelles soient liées au son. Nous pouvons alors envisager deux traitements gestuels séparés. Un pour le suivi gestuel et un autre pour la reconnaissance des micro-variations. En utilisant la synthèse hybride développée il est simple de contrôler les paramètres sonores par rapport à ces micro variations. La figure 21 montre ces extensions.

Par ailleurs, un problème rencontré sur le travail réalisé sur le geste et son intermédiaire est la difficulté d'avoir une définition gestuelle et sonore communes. Il serait utile de se poser la question de définition d'une typologie de geste pour la définition et synthèse de gestes.

Au niveau sonore, la synthèse paramétrique développée peut être étendue par le contrôle avec des enveloppes spectrales et par variations micro temporelles des paramètres comme expliqué en figure 21. Ceci afin d'avoir un modèle paramétrique pour la partie harmonique et pour les résidus.

8.2.2 Expressivité sonore par le geste

Le système d'apprentissage par démonstration est riche en possibilités de mapping et donc en applications, la liberté de pouvoir choisir les paramètres en entrée et en sortie permet d'explorer des mappings uniques. Au cours du stage deux possibilités ont été explorées, d'une part, l'apprentissage sur la voix pour le contrôle sonore qui ouvre des possibilités de réharmonisation par apprentissage. D'autre part, la création d'un espace aléatoire de sons par l'exemple où les partiels estimés ne contrôlent plus le synthétiseur hybride mais des échantillons de batterie. La particularité de cette deuxième méthode est la création d'environnements sonores qui n'ont à priori pas de lien logique mais qui peuvent être appris par l'utilisateur. Ainsi, le système se concentre sur la capacité de l'utilisateur à apprendre.

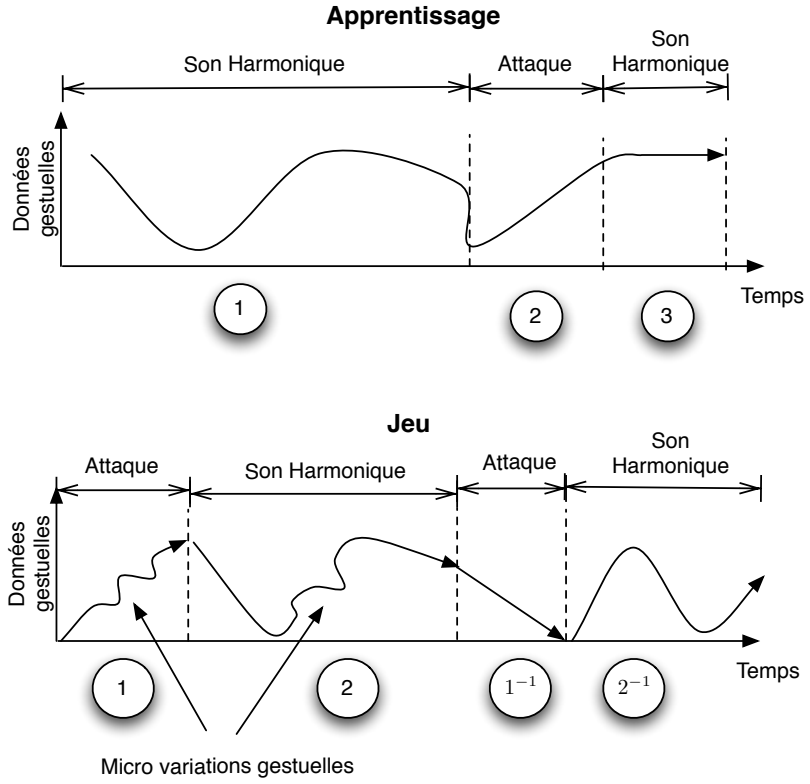


FIGURE 21 – Extensions pour le contrôle du système d'apprentissage par démonstration.

En effet, une des limitations du système d'apprentissage par démonstration est que le contrôle sonore est trop direct. L'idée serait d'étendre les capacités de contrôle et donc de créer une boucle de rétroaction plus présente entre le système et l'utilisateur afin de renforcer l'interaction.

Références

[1] D. Arfib et al. “Strategies of mapping between gesture data and synthesis model parameters using perceptual spaces”. In : *Organised Sound* 7.02 (jan. 2003), p. 127–144. issn : 1355-7718.

[2] JR Beltrán et F Beltrán. “Additive synthesis based on the continuous wavelet transform : A sinusoidal plus transient model”. In : *DAFX03* 3 (2003), p. 1–6.

[3] Frédéric Bevilacqua et Norbert Schnell. “Online gesture analysis and control of audio processing”. In : *Musical Robots and Interactive Multimodal Systems* (2011), p. 127–142.

[4] Frédéric Bevilacqua et al. “Continuous Realtime Gesture Following and Recognition”. In : *GW'09 Proceedings of the 8th international conference on Gesture in Embodied Communication and Human-Computer Interaction* (2010), p. 73–84.

[5] B. Caramiaux et al. “Mapping Through Listening”. In : *Computer Music Journal*. 2006, p. 1–30.

[6] A Cont. “Realtime audio to score alignement for polyphonic music instruments using sparse non-negative constraints and hierarchical HMMS”. In : *Acoustics, Speech and Signal Processing, 2006. ...* (2006).

[7] Arshia Cont. “ANTESCOFO : Anticipatory Synchronization and Control of Interactive Parameters in Computer Music.” In : *International Computer Music Conference (ICMC)* (2008).

[8] Jules Françoise. “Gesture – Sound Mapping by Demonstration in Interactive Music Systems”. In : *Training* (2013), p. 3–6.

[9] Jules Françoise, Norbert Schnell et Frédéric Bevilacqua. “A multimodal probabilistic model for gesture-based control of sound synthesis”. In : *Proceedings of the 21st ACM international conference on Multimedia - MM '13* (2013), p. 705–708.

[10] Jules Françoise et al. “Probabilistic Models for Designing Motion and Sound Relationships”. In : *New Interfaces for Musical Expression* (2014).

[11] RI Godø y et AR Jensenius. “Body movement in music information retrieval”. In : *10th International Society for Music Information Retrieval Conference* April (2009), p. 45–50.

[12] Andy Hunt et al. “Towards a Model for Instrumental Mapping in Expert Musical Interaction University of York Analysis-Synthesis Team”. In : *Proc. of the Int. Computer Music Conf. (ICMC'2000), ICMA : Berlin, Germany.* (2000), pp. 209–12.

[13] Julius O Smith Iii et Scott N Levine. “A Sines + Transients + Noise Audio Representation for Data Compression and Time Pitch Scale Modifications”. In : *Audio Engineering Society Convention 105* (1998).

[14] Frédéric Bevilacqua Jules Françoise, Baptiste Caramiaux. “A hierarchical approach for the design of gesture-to-sound mappings”. In : *SMC'12* (2012), p. 1 –8.

[15] Max Reference Pages. “MuBu for Max”. In : (2012).

[16] G Peeters. “A large set of audio features for sound description (similarity classification) in the CUIDADO project”. In : *Vasa* (2004).

[17] Valery A Petrushin. “Hidden Markov Models : Fundamentals and Applications Part 1 : Markov Chains and Mixture Models”. In : *Online Symposium for Electronics Engineer 2000* (2000).

[18] Valery A Petrushin. “Hidden Markov Models : Fundamentals and Applications Part 2 : Discrete and Continuous Hidden Markov Models”. In : *Online Symposium for Electronics Engineer 2000* (2000).

[19] L Rabiner. “A tutorial on hidden Markov models and selected applications in speech recognition”. In : *Proceedings of the IEEE* (1989), p. 257 –286.

[20] A Röbel. “Frequency slope estimation and its application for non-stationary sinusoidal parameter estimation”. In : *Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-07)* Bordeaux, France, September 10-15, 2007 (2007), p. 1–8.

[21] X Rodet et P Depalle. “Spectral Envelpes and Inverse FFT synthesis”. In : *Audio Engineering Society Convention 93* 3393 (1992).

[22] Xavier Rodet. “Musical Sound Signal Analysis/Synthesis : Sinusoidal+Residual and Elementary Waveform Models”. In : *TFTS'97* (1997), p. 1–11.

[23] Norbert Schnell, Frederic Bevilacqua et N Rasamimanana. “Playing the" MO"–Gestural Control and Re-Embodiment of Recorded Sound and Music”. In : *Proc. of NIME* June (2011), p. 535–536.

[24] Diemo Schwarz. “Concatenative sound synthesis : The early years”. In : *Journal of New Music Research* 35.1 (mar. 2006), p. 3–22. issn : 0929-8215.

[25] Diemo Schwarz. “Data-driven concatenative sound synthesis”. Thèse de doct. 2004.

[26] Diemo Schwarz, Norbert Schnell et Sebastien Gulluni. *Scalability in content-based navigation of sound databases*. 2009, p. 1–4.

[27] Diemo Schwarz et al. “Real-time corpus-based concatenative synthesis with catart”. In : *Proc. of the 9th Int. Conference on Digital Audio Effects (DAFx-06)* September (2006), p. 1–7.

[28] Xavier Serra. “A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition”. Thèse de doct. 1989.

Annexes

A Exemple d'estimation de paramètres pour un son avec changement rapide de paramètres , composé d'une vocalisation, d'un silence puis d'une vocalisation

A.1 Estimation de fréquences de partiels pour un son composé d'une vocalisation d'un silences puis d'une vocalisation

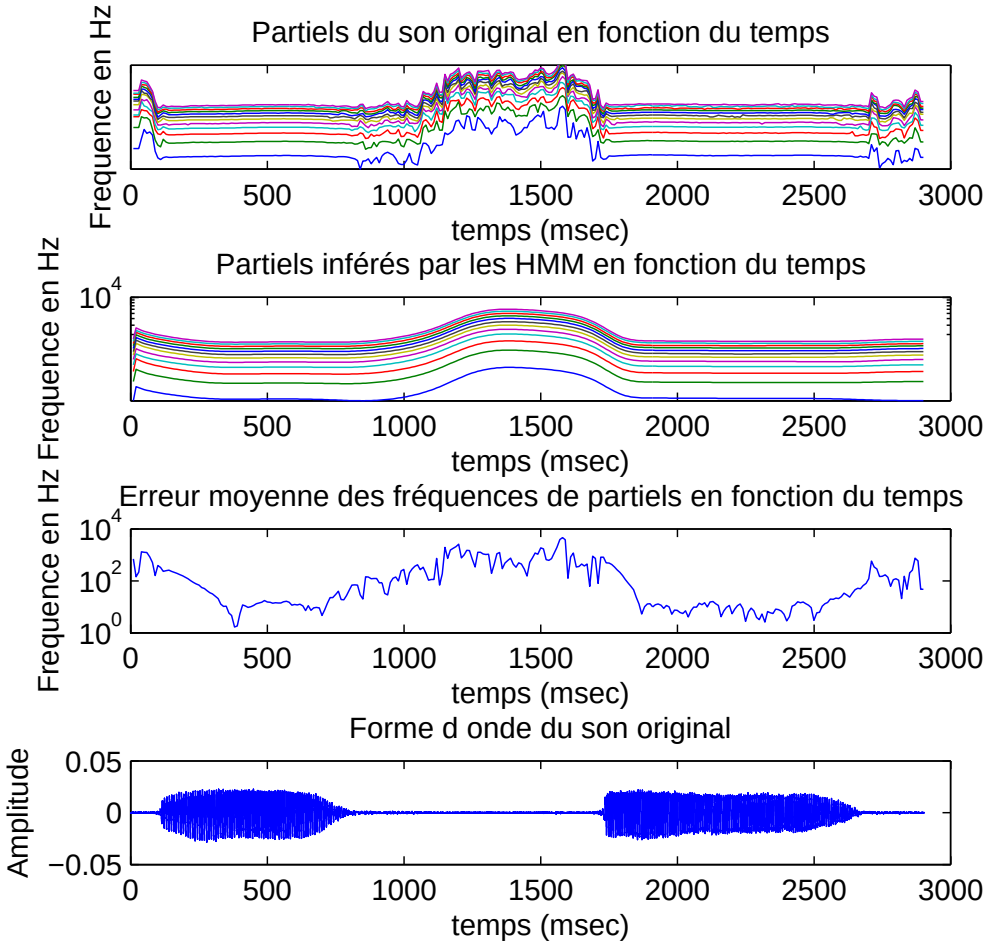


FIGURE 22 – Estimation de fréquences de partiels pour un son composé d'une vocalisation d'un silence puis d'une vocalisation.

A.2 Estimation des amplitudes de partiels pour un son composé d'une vocalisation d'un silence puis d'une vocalisation

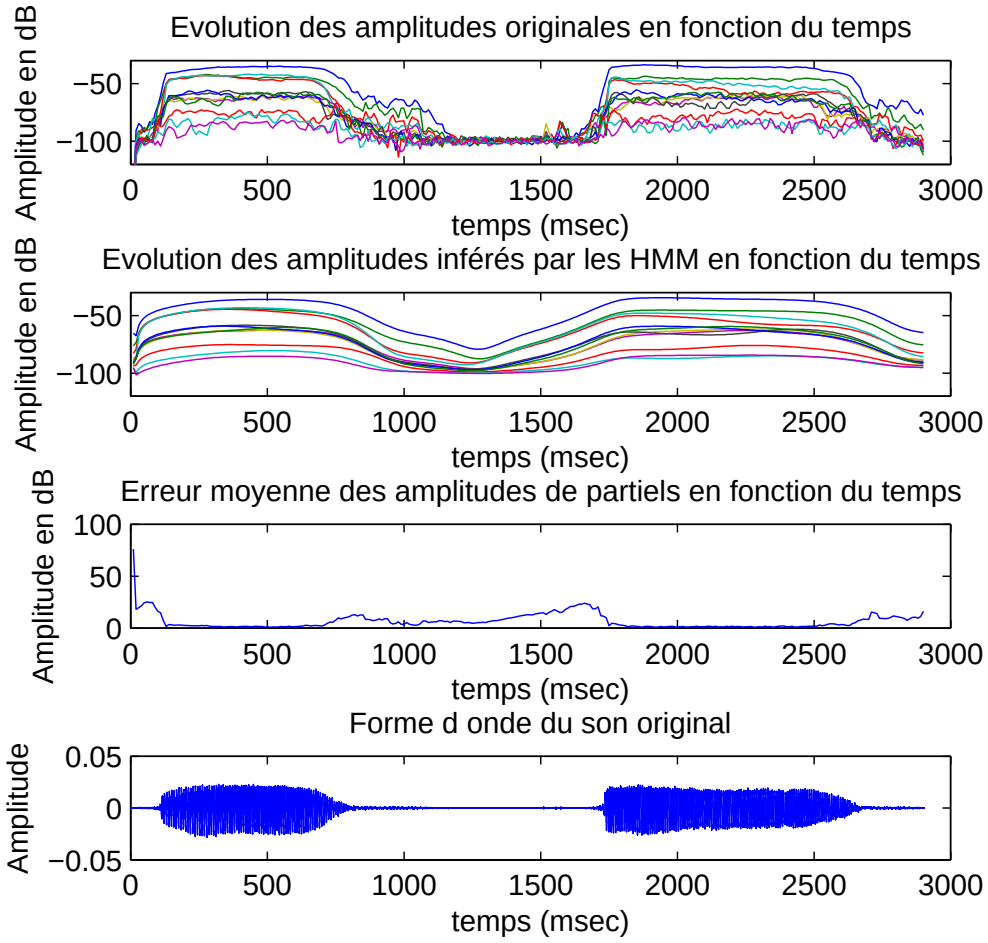


FIGURE 23 – Estimation d'amplitudes de partiels pour un son composé d'une vocalisation d'un silence puis d'une vocalisation.

B Influence de la "variance offset" sur la qualité de la régression les amplitudes de partiels

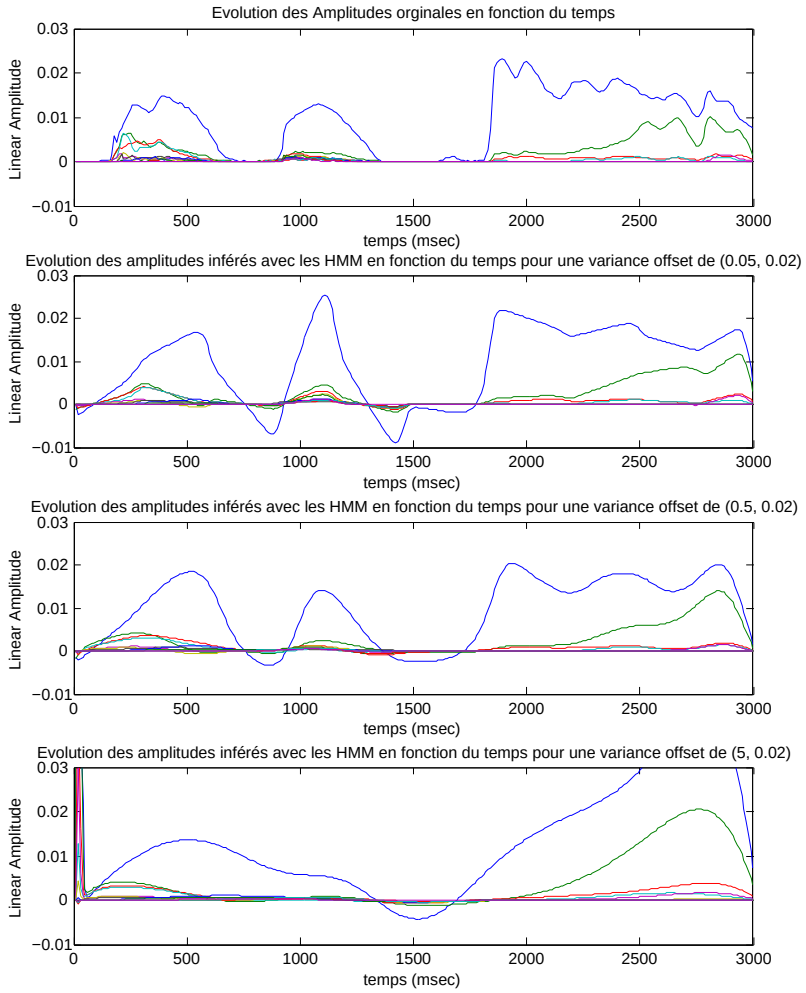


FIGURE 24 – Influence de la "variance offset" sur la qualité de la régression pour les amplitudes de partiels.

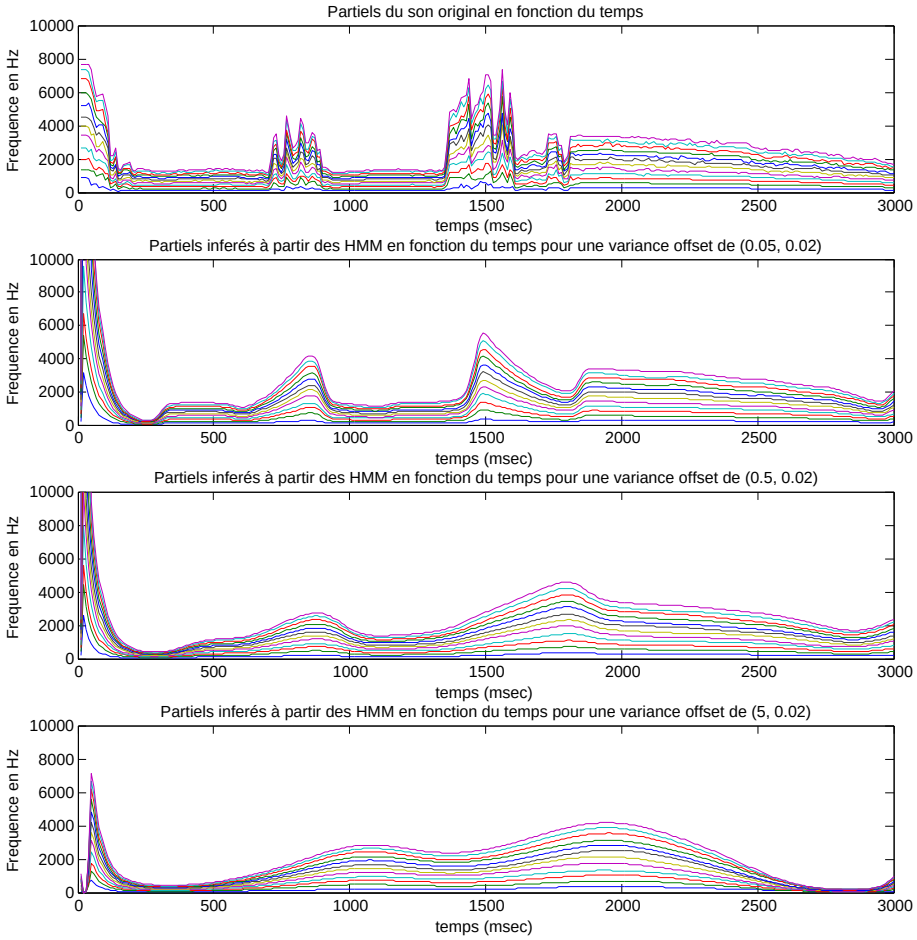


FIGURE 25 – Influence de la "variance offset" sur la qualité de la régression pour les fréquences de partiels.

C Questionnaire du test perceptif

Pour chaque question donner une note de un à cinq à chaque synthèse (1 mauvais, 5 très bon)

1- Exemple Geste – Son

1) Noter de un à cinq la qualité de chaque synthèse (Sons indésirables, artefacts).

A	B	C	D

2) Est-ce que la synthèse respecte le son original ? (1 non – 5 oui)

A	B	C	D

3) Est-ce que le contrôle par le geste est intéressant? (1–Pas lié - 5 très lié)

A	B	C	D

2- Exemple Geste – Son

1) Noter de un à cinq la qualité de chaque synthèse. (1 mauvais, 5 très bon)

A	B	C	D

2) Est-ce que la synthèse respecte le son original ? (1 non – 5 oui)

A	B	C	D

3) Est-ce que le contrôle par le geste est intéressant? (1–Pas lié - 5 très lié)

A	B	C	D

3 - Exemple Geste – Son

1) Noter de un à cinq la qualité de chaque synthèse. (1 mauvais, 5 très bon)

A	B	C	D

2) Est-ce que la synthèse respecte le son original ? (1 non – 5 oui)

A	B	C	D

3) Est-ce que le contrôle par le geste est intéressant? (1–Pas lié - 5 très lié)

A	B	C	D

4 - Exemple donné par l'utilisateur

1) Est-ce que la synthèse respecte le lien geste-son que vous avez démontrer?

A	B	C	D

2) Est-ce que la reproduction du geste vous permet de contrôler le son et pas seulement de le rejouer?

A	B	C	D

3) Décrivez en deux mots le son fait :

-

5 - Exemple donné par l'utilisateur

1) Quand vous refaite le geste, est-ce que la synthèse respecte le lien geste-son que vous avez démontrer?

A	B	C	D

2) Est-ce que la reproduction du geste vous permet de contrôler le son et pas seulement de le rejouer?

A	B	C	D

3) Décrivez en deux mots le son fait

-

7) Avez vous des commentaires sur le système? Qu'est ce qui vous semble intéressant ? Est-ce que vous trouvez les différences entre les synthèses significative

D Protocole Expérimental

— 1. Explication au sujet du déroulement de l'expérience et de la méthode de notation de chaque synthèse

- On cherche à étudier la qualité des différentes synthèses développés
- On cherche à étudier le lien entre geste et sons des différentes synthèses
- Vous allez devoir noter chaque synthèse de un à cinq par rapport à la question, un étant la plus mauvaise note et cinq la meilleure.
- Dans un premier temps vous allez refaire des gestes d'une base de données
- Dans un deuxième temps vous pourrez démontrer vos propres gestes en lien avec des vocalisations

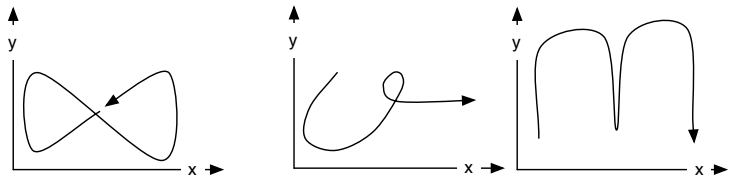


FIGURE 26 – Gestes utilisés pour l'expérience.

— 2. Explication du déroulement de la notation pour les trois premiers gestes

- Démonstration du geste de la base
- Écoute du son original
- Apprentissage du geste de la base
- Écoute des quatre synthèses
- Notation et écoute
- Refaire les étapes précédentes pour toutes les questions du geste
- Changement de geste

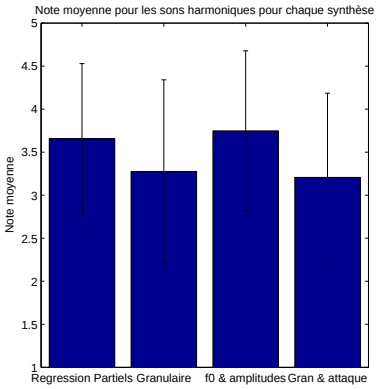
— 3. Démonstration d'un lien geste/ son : Explication au sujet de la méthode pour apprendre un geste en lien avec sa vocalisation

— 4. Pour les deux gestes appris :

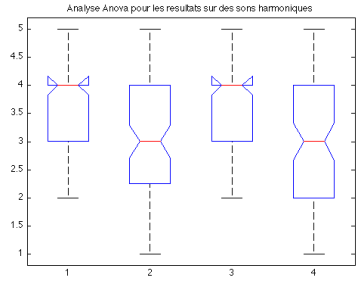
- Écoute des quatre synthèses
- Notation et écoute
- Refaire les étapes précédentes pour les trois questions du geste

E Résultats des tests perceptifs

E.1 Résultats pour les sons harmoniques



(a) Note moyenne pour les sons harmoniques



(b) Analyse anova des données pour les sons harmoniques

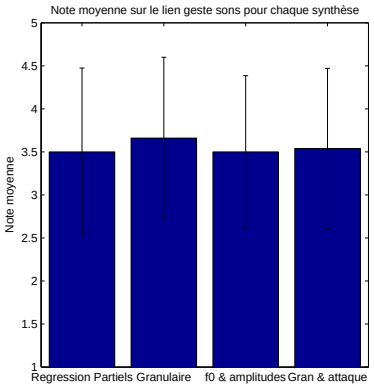
FIGURE 27 – Analyse des données pour les sons harmoniques

ANOVA Table

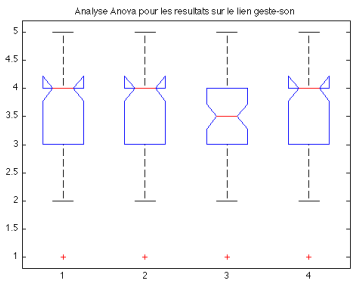
Source	SS	df	MS	F	Prob>F
Columns	18.966	3	6.32184	6.8	0.0002
Error	319.747	344	0.9295		
Total	338.713	347			

FIGURE 28 – Valeurs de l'analyse Anova pour les questions concernant les sons harmoniques.

E.2 Résultats pour les lien geste-son



(a) Note moyenne sur le lien geste son pour chaque synthèse



(b) Analyse Anova des données sur le lien geste son pour chaque synthèse

FIGURE 29 – Analyse des données sur le lien geste-son

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	0.86	3	0.28667	0.33	0.8042
Error	170.64	196	0.87061		
Total	171.5	199			

FIGURE 30 – Valeurs de l'analyse anova pour les questions concernant le lien geste-son.

F Interface utilisateur du synthétiseur granulaire à conservation d'attaques

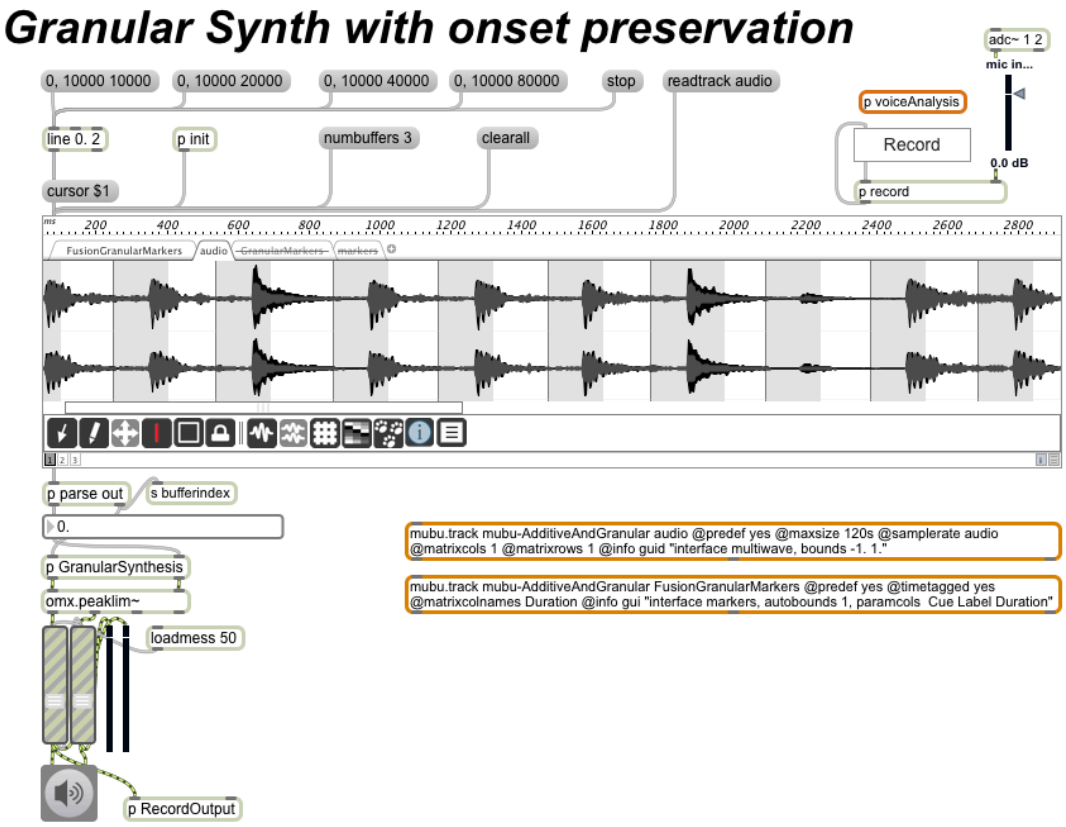


FIGURE 31 – Interface utilisateur du synthétiseur granulaire à conservation d'attaques.

G Interface utilisateur du synthétiseur hybride additif-granulaire

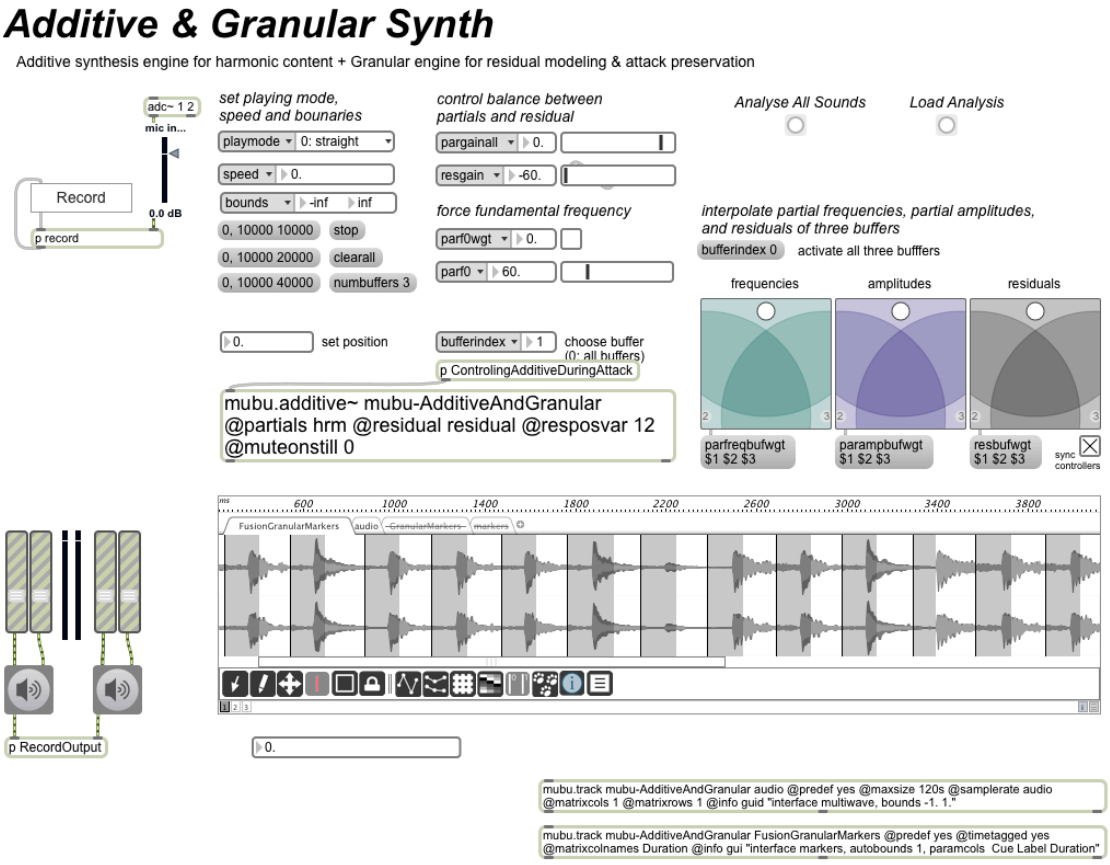


FIGURE 32 – Interface utilisateur du synthétiseur granulaire à conservation d'attaques.

H Interface utilisateur du patch utilisé pour réaliser les tests perceptifs

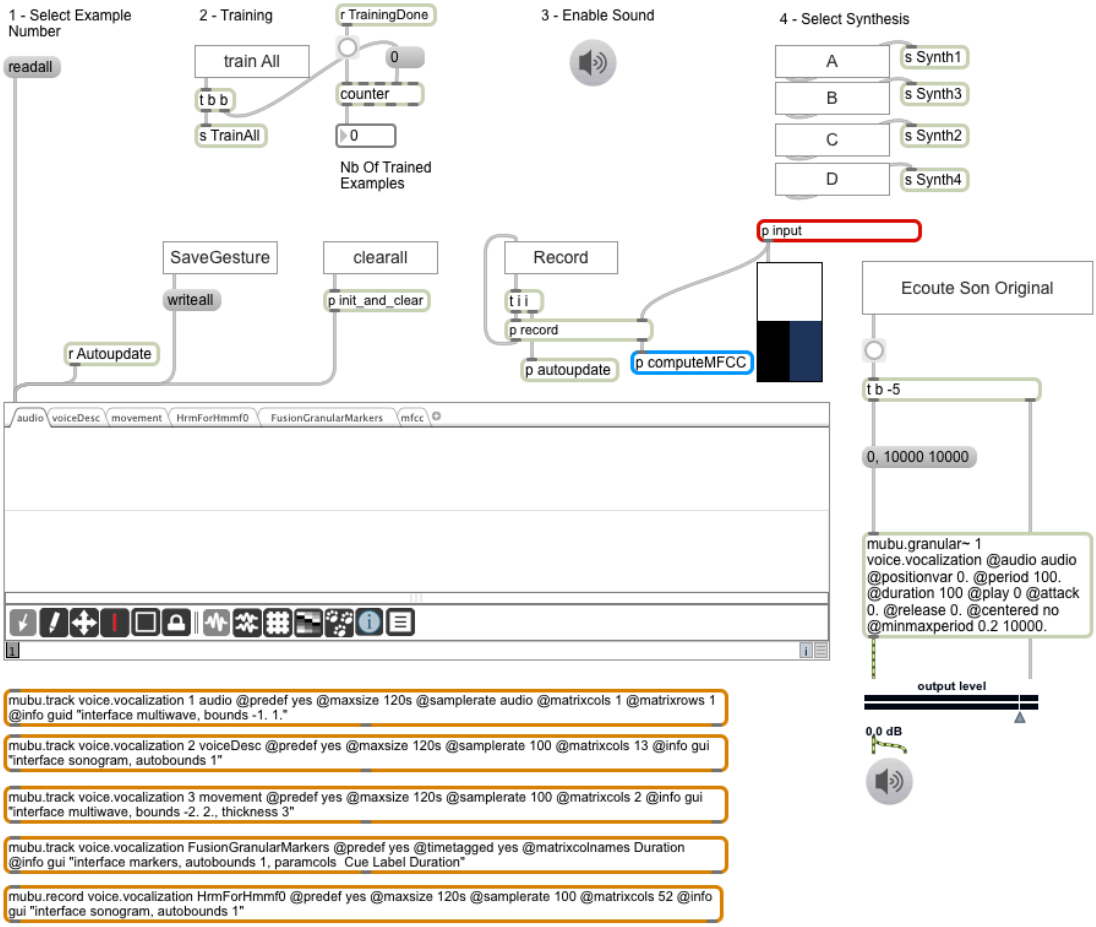


FIGURE 33 – Interface utilisateur du patch de régression sur les partiels.

Table des figures