# Vokinesis : instrument de contrôle suprasegmental de la synthèse vocale

Samuel Delalez

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir

# Vokinesis : instrument de contrôle suprasegmental de la synthèse vocale

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Paris-Sud

École doctorale n°580, Sciences et Technologies de l'Information et de la Communication, STIC
Spécialité de doctorat: Informatique

Thèse présentée et soutenue à Orsay, le 28 Novembre 2017, par

**Samuel Delalez**

Composition du Jury :

| | |
|---|---|
| **Wendy Mackay**<br>Directeur de Recherche INRIA, Université Paris-Sud (LRI) | Président |
| **Philippe Depalle**<br>Professeur, McGill University (Schulich School of Music) | Rapporteur |
| **Nathalie Henrich Bernardoni**<br>Directeur de Recherche CNRS, Université de Grenoble (GIPSA-LAB) | Rapporteur |
| **Boris Doval**<br>Maître de Conférence, Sorbonne Université (IJLRA)<br>(Institut Jean le Rond ∂ 'Alembert) | Examinateur |
| **Christophe d'Alessandro**<br>Directeur de Recherche CNRS, Sorbonne Université<br>(Institut Jean le Rond ∂ 'Alembert) | Directeur de thèse |

**Résumé**

Ce travail s'inscrit dans le domaine du contrôle performatif de la synthèse vocale, et plus particulièrement de la modification temps-réel de signaux de voix pré-enregistrés. Dans un contexte où de tels systèmes n'étaient en mesure de modifier que des paramètres de hauteur, de durée et de qualité vocale, nos travaux sont centrés sur la question de la modification performative du rythme de la voix. Une grande partie de ce travail de thèse a été consacrée au développement de Vokinesis, un logiciel de modification performative de signaux de voix pré-enregistrés. Il a été développé selon ces objectifs : permettre le contrôle du rythme de la voix, avoir un système modulaire, utilisable en situation de concert ainsi que pour des applications de recherche. Son développement a nécessité une réflexion sur la nature du rythme vocal et sur la façon dont il doit être contrôlé. Il est alors apparu que l'unité rythmique inter-linguistique de base pour la production du rythme vocal est de l'ordre de la syllabe, mais que les règles de syllabification sont trop variables d'un langage à l'autre pour permettre de définir un motif rythmique inter-linguistique invariant. Nous avons alors pu montrer que le séquencement précis et expressif du rythme vocal nécessite le contrôle de deux phases, qui assemblées forment un groupe rythmique : le *noyau* et la *liaison* rythmiques. Nous avons mis en place plusieurs méthodes de contrôle rythmique que nous avons testées avec différentes interfaces de contrôle. Une évaluation objective a permis de valider l'une de nos méthodes du point de vue de la précision du contrôle rythmique. De nouvelles stratégies de contrôle de la hauteur et de paramètres de qualité vocale avec une tablette graphique ont été mises en place. Une réflexion sur la pertinence de cette interface au regard de l'essor des nouvelles interfaces musicales continues nous a laissé penser que la tablette est la mieux adaptée au contrôle expressif de l'intonation et de la mélodie monophonique, mais que les PMC (Polyphonic Multidimensional Controllers) sont mieux adaptés au contrôle polyphonique. Le développement de Vokinesis a également nécessité la mise en place de la méthode de traitement de signal VoPTiQ (Voice Pitch, Time and Quality modification), combinant une adaptation de l'algorithme RT-PSOLA et des techniques particulières de filtrage pour les modulations de qualité vocale. L'utilisation musicale de Vokinesis a été évaluée avec succès dans le cadre de représentations publiques du Chorus Digitalis, pour du chant de type variété ou musique contemporaine. L'utilisation dans un cadre de musique électronique a également été explorée par l'interfaçage du lociciel de création musicale Ableton Live. Les perspectives d'application sont multiples : études scientifiques (recherches en prosodie, en parole expressive, en neurosciences...), productions sonores et musicales, pédagogie des langues, thérapies vocales.

**Mots clés :** synthèse vocale, interactions Humain-Machine, informatique musicale, prosodie, traitement du signal vocal

**Abstract**

This work belongs to the field of performative control of voice synthesis, and more particularly of real-time modification of pre-recorded voice signals. In a context where such systems were only capable of modifying parameters such as pitch, duration and voice quality, our work focuses on the question of performative modification of voice rhythm. One significant part of this thesis has been devoted to the development of Vokinesis, a program for performative modification of pre-recorded voice. It has been developed under these goals : to allow for voice rhythm control, to obtain a modular system, usable in public performances situations as well as for research applications. To achieve this development, a reflexion about the nature of voice rhythm and how it should be controlled has been carried out. It appeared that the basic inter-linguistic rhythmic unit is syllable-sized, but that syllabification rules are too language-dependant to provide an invariant inter-linguistic rhythmic pattern. We showed that accurate and expressive sequencing of voice rhythm is performed by controlling the timing of two phases, which together form a rhythmic group : the rhythmic *nucleus* and the rhythmic *link*. We developed several rhythm control methods, tested with several control interfaces. An objective evaluation showed that one of our methods allows for very accurate control of rhythm. New strategies for voice pitch and quality control with a graphic tablet have been established. A reflexion about the relevance of graphic tablets for pitch control, regarding the rise of new continuous musical interfaces, has left us think that they best fit expressive monophonic intonation and melody control, but that PMC (Polyphonic Multidimensional controllers) are better for polyphonic control. The development of Vokinesis also required the implementation of the VoPTiQ (Voice Pitch, Time and Quality modification) signal processing method, which combines an adaptation of the RT-PSOLA algorithm and some specific filtering techniques for voice quality modulations. The use of Vokinesis as a musical instrument has been successfully evaluated in public representations of the Chorus Digitalis ensemble, for various singing styles (from pop to contemporary music). Its use for electronic music has also been explored by interfacing the digital audio workstation Ableton Live with Vokinesis. Application perspectives are diverse : scientific studies (research in prosody, expressive speech, neurosciences...), sound and music production, language learning and teaching, speech therapies.

**Keywords :** voice synthesis, Human-Computer interactions, sound and music computing, prosody, vocal signal processing

*Je dédie cette thèse à Sacha*

*Les paroles que nous venons de prononcer,*
*Le temps, dans son vol,*
*Les a déjà emportées, et rien ne revient.*

-

Horace, *Odes.*
Vers restitués en français par Carlo Rovelli dans *l'Ordre du Temps*,
à partir de la traduction de Giulio Galetto dans *In questo breve cerchio.*

# Remerciements

Mes premiers remerciements se tournent naturellement vers Christophe d'Alessandro, qui m'a offert ces trois années de recherches passionnantes, agrémentées de conversations toujours enrichissantes aussi bien du point de vue culturel que scientifique. Il a su m'aider à surmonter les moments les plus difficiles, et a ainsi été un élément clé dans l'accomplissement de cette thèse. J'ai également beaucoup apprécié la confiance qu'il a pu porter à l'égard de mes initiatives dans les différentes représentations musicales qu'ont nécessité ces travaux.

Je souhaite également remercier Boris Doval qui, comme je le lui ai déjà dit, est sans doute l'un des meilleurs enseignants que j'ai pu rencontrer lors de ma scolarité. Ses cours sur l'analyse / synthèse de la parole que j'ai suivis en Master 2 ont su me passionner, à un point tel que je conclue aujourd'hui l'écriture d'une thèse dans le domaine. J'étais très heureux qu'il ait pu faire partie de mon jury de thèse.

Je remercie également tous les autres membres du jury, Wendy Mackay, Nathalie Henrich Bernardoni et Philippe Depalle, de m'avoir fait l'honneur de leur présence, d'être venu de si loin pour certains, mais surtout pour leurs judicieux commentaires et conseils indispensables à la finalisation de cette thèse.

Merci aux nombreux membres du Chorus Digitalis, Christophe, Boris, Lionel, Olivier, Annelies, Hélène, Victor, Robert, Michael (le ramasseur de balles !), pour leurs participations aux représentations dont les préparations ont été extrêmement bénéfiques au bon développement de Vokinesis, mais également pour les bons moments passés en votre compagnie.

Je tiens à saluer les collègues qui ont partagé des bureaux, des couloirs, des repas, des coups à boire... Salut donc aux gens du LAM : Camille, Hugo (le doctorant), Arthur, Augustin, Louis, Hugo (le stagiaire), Jean-Loïc, Hughes, Michèle, Claudia, Laurent, René. Salut aux membres de la défunte équipe AA du LIMSI, dont j'ai soutenu la dernière thèse : Olivier, David (les deux !), Lionel, Trang, Bart, Peter, Brian, Albert, Marc, Areti, Laurent, Justin. Salut aux secrétaires du LIMSI super sympas, et à la direction, super sympa ! Un merci particulier à Brian et David pour leurs retours extrêmement bénéfiques lors de mes répétitions de soutenance. Vous avez été indispensables à ma réussite.

Un immense merci à ma sœur, mon père et ma mère (et à ceux qui les accompagnent !), qui m'ont toujours suivi dans mes décisions, qu'elles soient bonnes ou mauvaises, et qui m'ont ainsi épaulé dans les moments où le soutien était nécessaire. Je ne remercierai jamais assez mes parents de m'avoir permis de suivre mes études jusqu'au bout, sans jamais douter. Je tiens aussi à remercier Gérard et Diloue, qui n'ont jamais cessé de croire en moi.

Je tiens également à remercier chaleureusement tous mes amis, qui m'ont toujours poussé à continuer, et qui m'ont permis de m'évader, loin du monde scientifique. Que l'on se soit beaucoup vus, ou juste aperçus, cela aura toujours été un grand plaisir ! Merci donc à Nico, Quentin, Jano, Marie, Maureen, Antoine, Juju, Jason, Valoche, Georgie, Élie, Hugo V., Wilson, Jules, Augustin, Hugo T., Yoyo, Wilou, Pouny, Sandra, Crisp'X, Diana, Franky, Pablo, et j'espère ne pas trop en oublier... Un grand merci également à tous les membres de Zärhza : Chipou, Timo, Tonio, Dussan, (et un coucou à Vianney !). Cette année passée avec vous m'a fait un bien fou. Merci aux présents le jour de ma soutenance : Michèle, Jean-Sylvain, Olivier, David, Valentin et Luc. C'était très important que vous soyez là. Et merci à tous ceux qui sont venus boire des coups après : Olivier, Marie, Charlene, Hugo et Jason. Love to all of these people ! (c'est plus simple à dire en anglais...)

Enfin, je tiens à remercier les Monty Python pour avoir écrit leur sketch « the argument », et aux personnes qui ont eu l'idée de le faire interpréter par Dectalk Express et Intex Talker. La vidéo qui en résulte [1] a été l'objet de nombreuses pauses réparatrices, indispensables à la rédaction de ce manuscrit. Pour les mêmes raisons, merci à Michael Jackson.

1. https://www.youtube.com/watch?v=WjMwGWdqHVQ

# Terminologie

| | |
|---|---|
| $d_o(\gamma)$ | Durée du $\gamma^e$ groupe rythmique original |
| $d_s(\gamma)$ | Durée du $\gamma^e$ groupe rythmique de synthèse en mode *Sync Speed* |
| $D(t)$ | Fonction de déformation de durée |
| $FCP$ | Frame Control Point (point de contrôle du cadre rythmique) |
| $F/C$ | Frame/Content |
| $i$ | indice des périodes du signal original |
| $j$ | indice des périodes du signal de synthèse |
| $LN$ | Liaison / Noyau |
| $n$ | indice temporel des signaux discrets |
| $N$ | taille du signal original (en échantillons) |
| $NLN$ | Noyau / Liaison / Noyau |
| $N_w(i) = 2 \times P(i)$ | taille de la fenêtre d'analyse de la $i^e$ période originale |
| $P_c$ | position d'un contrôleur [0,1] |
| $P(i)$ | durée de la $i^e$ période originale |
| $P'(j)$ | durée de la $j^e$ période de synthèse |
| $P_{end}(\gamma)$ | FCP de fin de boucle en mode *Loop* |
| $P_l(\gamma)$ | FCP de liaison du $\gamma^e$groupe rythmique |
| $P_n(\gamma)$ | FCP de noyau du $\gamma^e$groupe rythmique |
| $P_{start}(\gamma)$ | FCP de début de boucle en mode *Loop* |
| $t_o(i)$ | $i^e$ périodique original |
| $t_s(j)$ | $j^e$ marqueur périodique de synthèse synthèse |
| $v_{max}$ | vitesse maximale en mode *Speed* |
| $v_{min}$ | vitesse minimale en mode *Speed* |

| | |
|---|---|
| $v_s$ | vitesse de lecture *Speed* |
| $w(i,n)$ | fenêtre d'analyse de la $i^e$ période originale |
| $x(n)$ | signal original |
| $x_w(i,n)$ | $i^e$ signal à court-terme original, fenêtré autour de la $i^e$ période originale |
| $y(j,n)$ | $j^e$ signal à court-terme de synthèse |
| $y(n)$ | signal de synthèse final |
| $\alpha$ | facteur d'interpolation périodique |
| $\tau(j)$ | instant cible dans le signal original, obtenu à la $j^e$ période de synthèse |
| $\gamma$ | indice de groupe rythmique |
| $\tau_{loop}$ | instant cible en mode *Loop* |

# Exemples audio et vidéo

Les exemples audio et vidéo mentionnés dans ce manuscrit peuvent être téléchargés à l'adresse suivante :

www.kepstral-audio.com/download/ThesisDelalezMedia.zip

# Table des matières

Chapitre 1

# Introduction

**Sommaire**



## 1.1 Contexte et problématiques

La production vocale est le résultat d'interactions complexes d'un ensemble d'organes internes à l'appareil phonatoire. La production de consonnes et de voyelles nécessite une maîtrise experte des différents articulateurs, qui s'acquière au bout de plusieurs années d'apprentissage. L'*externalisation vocale*, ou le *contrôle performatif de la voix*, consiste à utiliser des techniques de contrôle temps-réel d'une voix synthétique par des membres externes à l'appareil phonatoire. Des gestes externes permettent alors d'imiter les gestes internes. Il peut s'agir de contrôler des paramètres articulatoires, acoustiques, phonétiques, ou temporels de voix synthétisées par des systèmes mécaniques, électroniques ou informatiques, grâce à des interfaces de contrôle diverses et variées (claviers, gants de données, tablettes graphiques...) Cette thèse pose la question de l'externalisation du contrôle rythmique d'un texte parlé ou chanté.

La première application pratique de l'externalisation vocale qui vienne à l'esprit est sans doute la dernière étape que connaîtra ce champs de recherche : le dispositif de remplacement complet de la voix. Un système qui puisse permettre à des personnes ayant perdu l'usage de la parole de communiquer verbalement sans avoir à écrire le texte à l'avance – mais nous en sommes encore loin. D'un point de vue plus réaliste par rapport aux avancées actuelles, l'externalisation vocale peut être utilisée à des fins musicales, thérapeutiques, pédagogiques ou encore scientifiques

(la séparation des éléments de contrôle et de production du son permet l'étude indépendante des ces deux fonctionnalités, chose bien entendu impossible dans le cas de la voix réelle, comme dans celui de la plupart des instruments acoustiques [Wanderley & Depalle 2004]).

Si l'on souhaite un jour atteindre l'application de remplacement de la parole, il est primordial de chercher des techniques de contrôle qui ne demandent pas un temps d'apprentissage aussi long que celui de la parole naturelle ou du violon. Depuis le début du XX$^{e}$ siècle, un certain nombre de systèmes de synthèse performative a vu le jour. Certains permettent de contrôler les phonèmes de façon partielle ou complètes : le *Voder* [Dudley 1939], le *Glove Talk* [Fels & Hinton 1998], le *SPASM* [Cook 1993], le *HandSketch* [D'Alessandro & Dutoit 2007] et le *Cantor Digitalis* [Feugère *et al.* 2017]. Ce sont tous des systèmes de synthèse pure : aucun son pré-enregistré n'est utilisé pour produire le signal vocal. le Voder, le Glove Talk et le SPASM permettent de produire des voyelles et des consonnes, mais ils demandent un très long temps d'apprentissage pour des résultats au naturel peu convainquant. Le Cantor Digitalis et le HandSketch offrent une synthèse de bonne qualité, mais ils ne permettent de produire que des voyelles.

Certains systèmes, tels que le *MAGE* [Astrinaki *et al.* 2012] ou *Calliphony* [Le Beux *et al.* 2007], fonctionnent à partir de modification de signaux de voix préparés. Le *MAGE* permet de contrôler la vitesse de lecture, la hauteur, et certains paramètres de qualité vocale de signaux obtenus à partir de synthèse Text-To-Speech par HMM, et Calliphony permet de modifier en temps-réel la vitesse de lecture et la hauteur de signaux de voix pré-enregistrés. Cela permet d'obtenir des signaux re-synthétisés au naturel bien supérieur à ceux des systèmes de synthèse pure. En effet, toutes nos voix de synthèse quotidiennes (transports, GPS, répondeurs...) sont aujourd'hui obtenues à partir de bases de données de signaux de voix naturelle, et il devient difficile de discerner les voix qui sont synthétiques de celles qui ne le sont pas. Cependant, les systèmes actuels de modification performative de signaux de voix préparés n'offrent pas une assez grande liberté de contrôle temporel : seule la vitesse de lecture est modifiable.

Lors de ce travail de thèse, nous avons développé Vokinesis (du latin *vox* : la voix, et du grec *kinesis* : le mouvement), un système d'externalisation vocale fondé sur la modification temps-réel de signaux pré-enregistrés. Il permet un contrôle temps-réel de la hauteur, de la force de voix, de la tension vocale et de la taille du conduit vocal, mais également du rythme, moins complexe à contrôler que les phonèmes, mais offrant plus de liberté q'un contrôle de la vitesse de lecture. Le développement de ce logiciel ainsi que les réflexions sur la nature du rythme vocal et sur les méthodes et interfaces permettant de le contrôler ont constitué le centre de ce travail de thèse. Sa réalisation a soulevé des questions sur la précision et le pouvoir expressif des différentes méthodes et interfaces de contrôle de la hauteur et du rythme de la voix, mais également sur des problèmes de traitement temps-réel du signal vocal. La maturité qu'a atteinte le logiciel nous a permis de mettre à l'épreuve ses capacités à être utilisé dans différents styles musicaux (variété, musique contemporaine, musique électro), et nous avons mis en place plusieurs représentations publiques en faisant

usage avec l'ensemble de voix de synthèse Chorus Digitalis.

La suite de ce chapitre a pour objectif de fournir une revue des précédents systèmes de contrôle performatif de la voix de synthèse. Nous y proposons une adaptation de l'historique proposé par [Perrotin 2015], en mettant l'accent sur deux époques distinctes : l'*avant* et l'*après* tablette graphique. En effet, la première période, qui s'étend de la fin du XVIII$^{e}$ à la fin du XX$^{e}$ siècle, n'a vu apparaître que quatre systèmes principaux de contrôle performatif de voix synthétiques. La seconde période, qui a débuté au début de ce siècle, a quant-à elle vu apparaître au moins cinq systèmes de ce type. Cela souligne bien le succès que connaît la tablette graphique pour une utilisation dans le cadre du contrôle performatif de la voix. En plus de cette réorganisation, nous proposons une légère mise à jour de cette revue en présentant deux nouveaux systèmes de synthèse performative.

## 1.2 Systèmes d'externalisation vocale ne faisant pas usage d'une tablette graphique

### 1.2.1 Machine de Von Kempelen

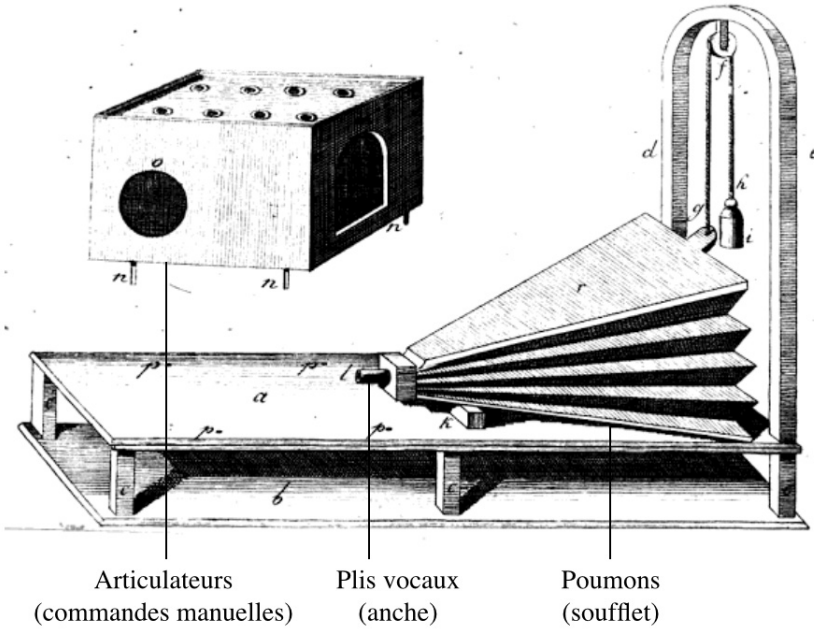


FIGURE 1.1 – *Machine parlante de Von Kempelen* [Von Kempelen 1791]. *La boite des articulateurs doit être reliée à la sortie du soufflet. L'anche, qui modélise les plis vocaux, se trouve à l'intérieur de cette boîte.*

Le premier système de contrôle performatif de la voix artificielle, la machine parlante de Von Kempelen, est un système mécanique rendu publique en 1791 à travers le livre [Von Kempelen 1791]. C'est un modèle mécanique de l'appareil vocal qui permet de produire une voix artificielle contrôlée par des gestes (FIGURE 1.1).

Les poumons sont modélisés par un soufflet, les plis vocaux par une anche et les articulateurs par un jeu d'actionneurs, non détaillés sur la figure. Le contrôle s'effectue avec un bras posé sur le soufflet et les deux mains à l'intérieur de la boite. Le soufflet, pressé par le bras, crée un flux d'air, et l'anche vibre. Les deux mains agissent ensuite sur les actionneurs pour moduler le son de l'anche et produire des phonèmes. Ce système a démontré son succès pour la prononciation de mots complexes, faisant intervenir des groupes de plusieurs consonnes consécutives, mais les utilisateurs ne semblent pas dépasser la longueur du mot. Nous terminerons sa présentation par une citation de son auteur :

*« Je ne donne pas [...] la machine parlante [...] comme un ouvrage bien achevé, et qui imite parfaitement la parole, mais j'ose me flatter, sans trop d'amour propre, que toute imparfaite qu'elle est, elle donne du moins de bons principes pour en construire une plus parfaite. [...] Puisse-t-il à la fin de ce siècle si fertile en découverte, se trouver une main de maître, qui porte cette découverte [...] au plus haut degré de perfection. »*

Plus de deux siècles plus tard, ce degré n'a sans doute pas encore été atteint. Nous montrerons par la suite que le vœu de Von Kempelen est aujourd'hui encore entendu, bien qu'en raison de la maîtrise de l'électricité, la voie choisie pour tenter de l'exaucer soit devenue bien différente de celle qu'il aurait sans doute imaginée.

### 1.2.2 Voder

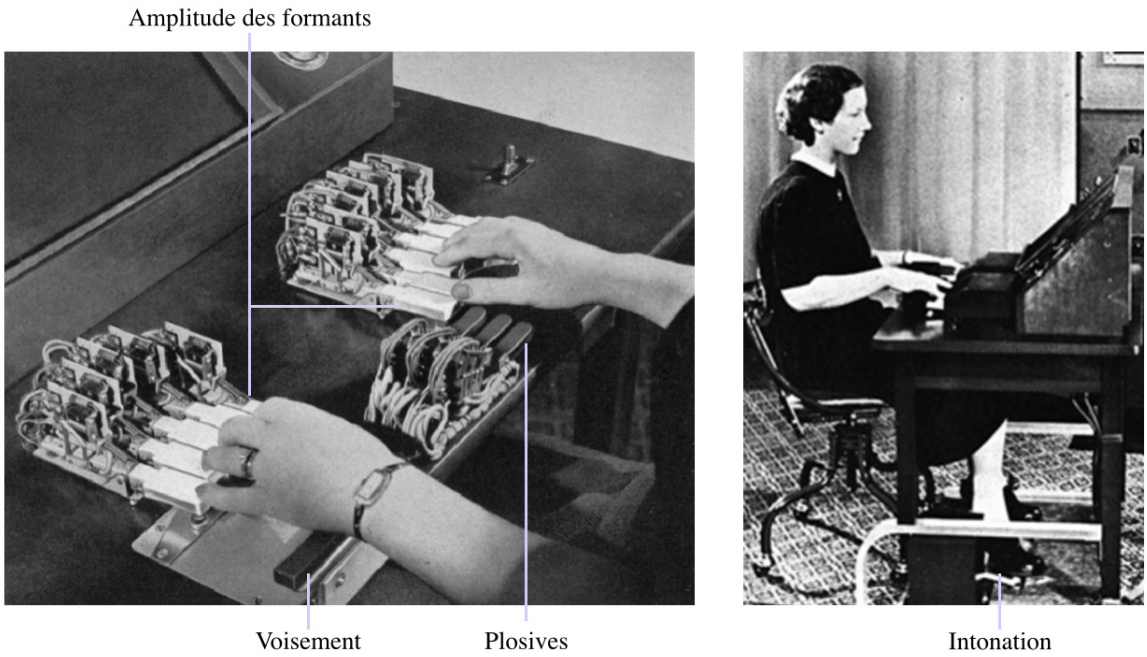


Figure 1.2 – *Utilisation du Voder* [Dudley *et al.* 1939] *: les mains contrôle l'articulation et le voisement, le pied contrôle la hauteur.*

Il a fallu attendre le XX$^{e}$ siècle et une excellente maîtrise de l'électricité avant que ne fût inventée une nouvelle machine parlante. Le Voder (Voice Operation DEmonstrator) a été développé en 1939 par [Dudley *et al.* 1939] dans les laboratoires Bell. Son principe de contrôle est représenté figure 1.2. C'est un synthétiseur à formants qui possède une source périodique pour les sons voisés, une source bruitée pour les sons fricatifs, et un jeu de 10 filtres passe-bande à fréquence de coupure et à largeur de bande fixes, qui servent à modéliser les formants. L'opératrice peut contrôler l'état de voisement grâce à la barre qui se trouve sous son poignet gauche. Si la barre est relâchée, la source non-voisée est active, et son amplitude sera diminuée au fur et à mesure que la pression sera augmentée. Inversement, l'amplitude de la source voisée augmentera avec la pression. L'articulation de la plupart des phonèmes se contrôle avec les touches blanches : chaque touche permet de contrôler l'amplitude d'un des dix filtres passe-bande. Si toutes ces touches sont relâchées, l'amplitude de chaque filtre est nulle, et aucun son n'est produit. Les trois touches noires permettent de déclencher les plosives. Enfin, l'intonation est contrôlée par une pédale. Selon l'auteur, il a fallu aux opératrices environ un an d'entraînement soutenu pour être en mesure d'avoir une conversation simple dont l'intelligibilité et le naturel laissaient souvent à désirer. La vidéo[1] fournit une démonstration de ce système.

### 1.2.3 SPASM

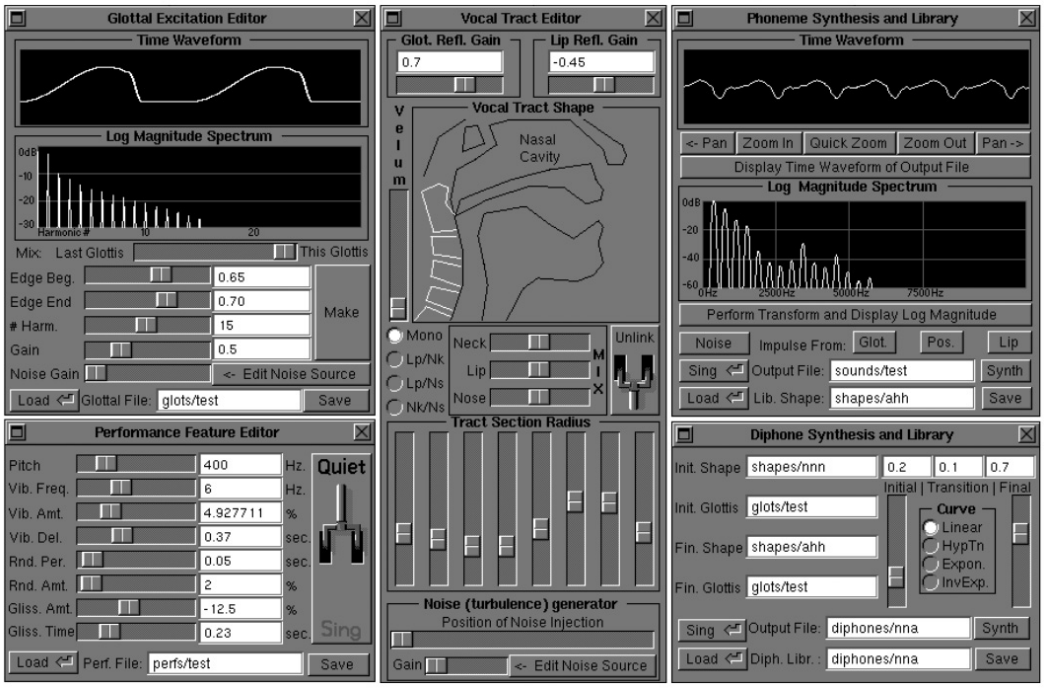


FIGURE 1.3 – *Fenêtres principale du SPASM* [Cook 1991]

1. https://youtu.be/0rAyrmm7vv0

Plus de cinquante ans, et des capacités de calcul informatique assez puissantes, furent nécessaires avant de voir apparaître le système *SPASM* (Singing Physical Articulary Synthesis Model) [Cook 1993]. Comme son nom l'indique, le système de synthèse de ce logiciel est basé sur un modèle articulatoire. Les paramètres contrôlables en temps-réel sont la hauteur, le vibrato (fréquence, amplitude et taux d'aléa), l'amplitude de la source de bruit, sa position d'injection dans le conduit vocal, l'aire de chacune des huit sections qui composent le conduit vocal et l'ouverture du voile du palais. Tous ces paramètres peuvent être retrouvés dans la Figure 1.3, qui présente les fenêtres principales du logiciel. Il est clair que la seule utilisation de la souris et du clavier de l'ordinateur n'offre pas assez de liberté pour un contrôle temps-réel simultané de tous ces paramètres. Pour ce qui concerne l'articulation, l'auteur propose une solution qui consiste à interpoler jusqu'à six différentes formes du conduit vocal avec un seul paramètre. Ensuite, il est possible d'assigner un contrôleur MIDI à n'importe quel paramètre contrôlable en temps-réel. Par défaut, le logiciel assigne la hauteur aux notes d'un clavier MIDI, l'effort vocal à l'aftertouch, et l'interpolation des formes du conduit vocal à la roue de modulation (une explication du fonctionnement du protocole et des interfaces MIDI sera fournie dans la section 4.2). L'auteur a par la suite cherché à développer de nouvelles interfaces mieux adaptées au contrôle expressif de l'instrument vocal [Cook 2005], avec comme idée principale de lier l'effort vocal à la force d'un souffle, et de trouver des manières de contrôler la hauteur et l'articulation de manière continue. Ainsi apparurent les Squeezevoxen (des accordéons augmentés), le VOMID (*Voice-Oriented Melodica Interface Device*, un mélodica augmenté) et le COWE (*Controller, One With Everything*, une interface composée d'un capteur de souffle, de capteurs de pression, de boutons et d'accéléromètres). Des exemples audio de signaux synthétisés par SPASM peuvent être écoutés sur cette page [2]

### 1.2.4 Glove Talk

le Glove Talk est un système de contrôle temps-réel d'un synthétiseur à formants [Rye & Holmes 1982]. La voix de synthèse est contrôlée par une pédale d'expression et par des gants augmentés de capteurs de pression, d'un accéléromètre et d'un gyromètre. Le principe consiste à apprendre à l'ordinateur, à travers un réseau de neurones, une association de gestes et de sons. Dans sa première version [Fels & Hinton 1993], les gestes permettent de contrôler les mots d'un dictionnaire. Dans sa seconde version [Fels & Hinton 1998], les gestes permettent de contrôler des phonèmes, et donc de prononcer n'importe quelle phrase, offrant ainsi la possibilité d'avoir une véritable conversation. La Figure 1.4 montre son auteur en train de parler à l'aide de ce système. La hauteur du signal de synthèse est contrôlée par la hauteur de sa main droite. Cette même main permet également de contrôler l'articulation des voyelles de manière continue grâce à sa position dans le plan horizontal. Les consonnes sont contrôlés par les deux mains : différentes positions du pouce sur les autres doigts sélectionnent différentes consonnes. La force de voix est contrôlée

2. https://www.cs.princeton.edu/~prc/SingingSynth.html

FIGURE 1.4 – *Utilisation du Glove Talk par S. Fels.*

par une pédale d'expression. Selon l'auteur une centaine d'heures d'entraînement sont nécessaires avant d'être capable de prononcer un discours intelligible. La vidéo[3] le confirme, et présente même une étape de conversation, mais elle montre tout de même que le naturel de la synthèse est assez peu convainquant, et que les gestes à accomplir sont assez complexes.

### 1.2.5 Miku Stomp : contrôle du rythme vocal avec une guitare


FIGURE 1.5 – *Korg Miku Pedal : contrôle à la guitare du rythme syllabique et de la hauteur de la voix de Hatsune Miku, le personnage de Vocaloid.*

La *Miku Stomp*, présentée FIGURE 1.5, est une pédale d'effet. Elle transforme

3. https://youtu.be/hJpGkroFP3o

le signal audio émis par une guitare en un signal vocal calculé en temps-réel par le synthétiseur Vocaloid [Kenmochi & Ohshita 2007]. La voix contrôlé est celle de la célébrité virtuelle japonaise Hatsune Miku [4]. À notre connaissance, c'est le seul système de synthèse performative permettant de contrôler le rythme vocal : les attaques de notes déclenchent des syllabes. Malheureusement, aucune publication scientifique n'en fait l'objet. Cependant, quelques testeurs de pédales d'effets nous donnent leur avis sur son fonctionnement dans les vidéos A [5] et B [6]. Même si leur manière d'exprimer le problème est différente (la vidéo A est plus diplomate que la B), les deux semblent d'accord pour dire que le contrôle de la mélodie fonctionne très bien, mais que le contrôle du rythme pose parfois problème.

Notez que depuis que la tablette graphique est utilisée dans un cadre musical (vers le début de ce siècle), la Miku Stomp est le seul système de contrôle performatif de la synthèse vocale à ne pas en faire usage.

## 1.3 Systèmes d'externalisation vocale faisant usage d'une tablette graphique

Quand l'utilisation des tablettes graphiques s'est développée, l'interface a fait consensus pour le contrôle de la synthèse vocale : la plupart des systèmes de contrôle performatif de la voix qui ont vu le jour depuis le début de ce siècle en font usage. [Wright *et al.* 1997] fournit une évaluation positive d'une tablette Wacom pour un contrôle musical, car elle offre un grand nombre de degrés de liberté. En effet, le stylet d'une tablette Wacom permet en contrôle en 5 dimensions : les données émises correspondent à sa position et au degré d'inclinaison dans le plan $(x, y)$ (4 dimensions), mais également à la pression qui y est appliquée (axe $z$). De plus, les stylets possèdent deux boutons poussoirs accessibles avec l'index.

La façon dont les différents paramètres vocaux sont contrôlés diffère selon le système. Pour ce qui est de l'articulation des voyelles, la plupart des systèmes utilisent un *plan vocalique* en deux dimensions, tel que celui présenté Figure 1.6 : les combinaisons de fréquences centrales des deux premiers formants F1 et F2 peuvent à elles seules représenter la totalité des voyelles non-nasales.

Le premier système de synthèse vocale faisant usage d'une tablette graphique permettait de contrôler le synthétiseur CHANT [Wanderley *et al.* 2000]. La hauteur et l'effort vocal étaient contrôlés par un capteur de position et de pression, et les voyelles par la position $(x, y)$ du stylet sur la tablette graphique, qui représentait alors le plan vocalique.

Dans le système *Voicer* [7], présenté en haut de la Figure 1.7, [Kessous 2004a, Kessous 2004b] a mis en place une méthode de contrôle circulaire de la hauteur sur une tablette graphique, permettant d'outrepasser les limitations spatiales pour

4. https://youtu.be/rL5YKZ9ecpg à partir de 12m50
5. https://youtu.be/a5tLniLHUgY à 5m45
6. https://youtu.be/aveUEZkcQno à 11m46
7. www.jmc.blueyeti.fr/Videos/Voicer.mpg

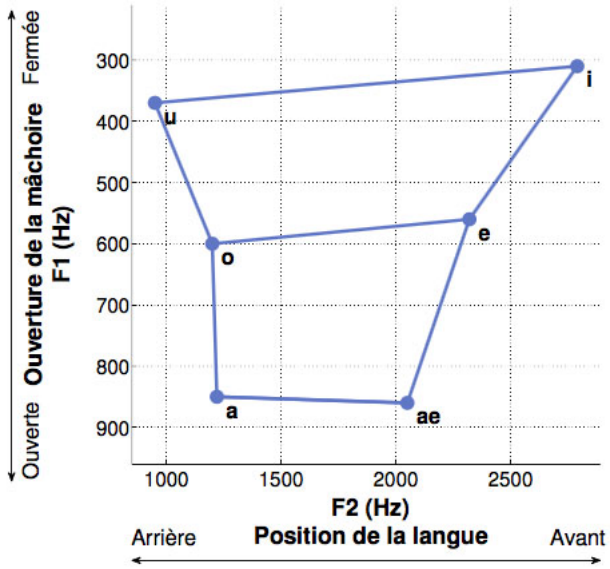


FIGURE 1.6 – *Représentation des voyelles extrêmes du français dans le plan des deux premiers formants. Image issue de* [Perrotin 2015].

le cas de variations mélodiques importants. Il a également testé la déportation du contrôle des voyelles sur un joystick (en haut à gauche dans la FIGURE 1.7).

Le *Speech Conductor Project* [d'Alessandro *et al.* 2005] a été l'événement fondateur des systèmes que nous allons présenter par la suite (en dehors du *Pink Trombone*). Il a engendré la version temps-réel du modèle de source glottique CALM [D'Alessandro *et al.* 2006a, d'Alessandro 2009], ainsi que le système RAMCESS [d'Alessandro *et al.* 2006b, Le Beux 2009], permettant de produire une voix synthétique de façon performative à partir de méthodes d'analyse/synthèse d'une base de données vocale.

L'espace de contrôle *HandSketch*[8], présenté au centre de la figure, est une adaptation de la tablette graphique, sur laquelle sont ajoutés un masque de contrôle angulaire, ainsi qu'un jeu de capteurs FSR. Cet espace a été utilisé pour le contrôle de plusieurs synthétiseurs, vocaux ou non. Dans tous ces synthétiseurs, la position angulaire du stylet contrôle la fréquence fondamentale. L'un des synthétiseurs vocaux à en faire usage était un synthétiseur à formants [D'Alessandro & Dutoit 2007, D'Alessandro & Dutoit 2009]. La position radiale du stylet permettait de contrôler des paramètres de qualité vocale, et les capteurs FSR permettaient de produire sauts de notes rapides, et d'articuler des phonèmes. Un autre synthétiseur vocal ayant fait usage de l'espace HandSketch est le MAGE [Astrinaki *et al.* 2012][9], un synthétiseur paramétrique basé sur une modélisation HMM d'un corpus de parole. Une fois un texte fourni au système, la position radiale du stylet permet d'en contrôler la vitesse de lecture, la pression l'intensité, et l'inclinaison la taille du conduit vocal.

8. https://youtu.be/EK1Q7X_c3Q8
9. https://youtu.be/W7OwfUOA_HM

Calliphony [Le Beux *et al.* 2007, Le Beux 2009] est un système de modification temps-réel de hauteur et de durée de signaux de voix pré-enregistrés, dont Vokinesis a hérité. La hauteur était contrôlée par la position du stylet sur l'axe $x$ de la tablette, et la vitesse de lecture sur l'axe $y$. Son fonctionnement sera présenté en détails dans la section 3.1.

Le Cantor Digitalis [Feugère *et al.* 2017] est un synthétiseur à formants, dont la surface de contrôle est représentée en bas à gauche de la figure. Ce masque est apposé à la tablette graphique de la même manière que celui du HandSketch. La position et la pression du stylet permettent de contrôler la hauteur et l'effort vocal, et la position du doigt dans l'espace vocalique (carré rouge), faisant usage de la fonction tactile d'une tablette Wacom Touch, permet de contrôler l'articulation des voyelles. [Feugere 2013] a également développé une version du Cantor Digitalis nommée *Digitartic*, qui permet d'articuler certaines consonnes avec une deuxième tablette graphique. La position du stylet sur l'axe $x$ définissait la consonne, et l'axe $y$ permettait de contrôler la transition Voyelle (stylet en bas) / Consonne (stylet en haut). Dans cette représentation [10], le second musicien en partant de la gauche contrôle le Digitartic, les autres le Cantor Digitalis.

Le dernier système de contrôle performatif de la synthèse vocale que nous souhaiterions présenter est le *Pink Trombone* de Neil Thapen, dont la surface de contrôle est présentée en bas à droite de la figure. Il n'est pas contrôlé par une tablette graphique, mais peut être contrôlé par une interface de type tablette tactile ou smartphone [11], ou encore à la souris [12]. Il permet de produire des voyelles et des consonnes à partir d'un modèle géométrique du conduit vocal manipulable directement avec les doigts. Bien qu'il ne permette pas de tenir une conversation, nous trouvons ce paradigme de contrôle intéressant, et estimons qu'il méritait d'être présenté dans cette section.

## 1.4 Bilan et contenu du manuscrit

L'historique des systèmes de synthèse performative que nous venons de présenter nous a permis de mettre en évidence le succès qu'ont connu les tablettes graphiques pour une utilisation dans le cadre du contrôle de la synthèse vocale : avant leur apparition, chaque nouveau système utilisait des méthodes de contrôle très différentes des précédents. Aujourd'hui, tous les nouveaux systèmes d'externalisation vocale font usage de la tablette graphique pour le contrôle de la hauteur (sauf pour la pédale Miku Stomp de Korg dont l'objectif est plutôt commercial). Parmi tous les systèmes d'externalisation vocale que nous avons présentés, nous pouvons distinguer ceux qui permettent un contrôle au niveau segmental (voyelles et consonnes) de ceux qui permettent un contrôle de la vitesse de lecture de signaux préparés. Nous avons pu voir à plusieurs reprises que le contrôle segmental est très difficile, et qu'il demande de longs temps d'apprentissage pour des résultats peu convaincants : les mains et

10. https://youtu.be/d4TV-IcK8c8 à 6m40
11. https://youtu.be/7LGnozlwU1o
12. https://dood.al/pinktrombone/

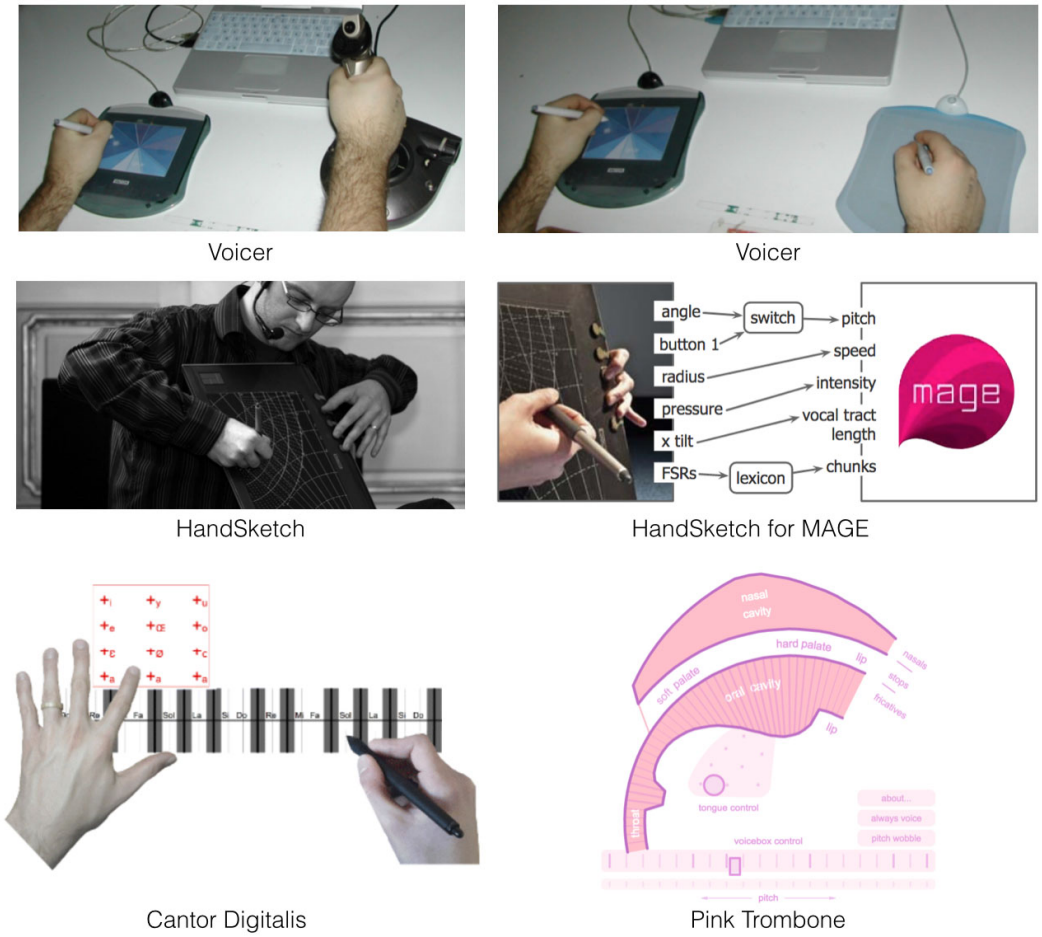


FIGURE 1.7 – *Systèmes de synthèse vocale contrôlés avec une tablette. Le Voicer* [Kessous 2004a, Kessous 2004b]*, le HandSketch* [D'Alessandro & Dutoit 2007, D'Alessandro & Dutoit 2009, Astrinaki *et al.* 2012] *et le Cantor Digitalis* [Feugère *et al.* 2017] *utilisent des tablettes graphiques équipées d'un stylet. Le Pink Trombone* [`https://dood.al/pinktrombone/`] *est contrôlable sur tablette et smartphone.*

les pieds reproduisent difficilement les mouvements coordonnés des différents articulateurs actifs lors de production vocale. Outre la pédale Miku, aucun système faisant usage de la tablette ne permet un contrôle du rythme vocal : le contrôle de la vitesse de lecture n'offre pas une liberté d'improvisation rythmique, ou le jeu synchrone entre plusieurs musiciens. La pédale Miku semble assez peu précise pour ce qui est du contrôle rythmique, et la guitare est sans doute moins adaptée que la tablette graphique pour le contrôle de la hauteur vocale.

Le Chapitre 2 présente d'abord une revue des principales méthodes de modification vocale. Il présente ensuite la méthode VoPTiQ (*Voice Pitch Time and Quality modification*) que nous avons mise en place. Cette méthode regroupe plusieurs techniques de traitement de signal, qui ensemble permettent de modifier en temps-réel la hauteur, la durée, la taille du conduit vocal, la force de voix et la tension vocale de signaux de voix pré-enregistrés.

La question du contrôle du rythme vocal sera traitée dans le Chapitre 3. Une réflexion sur la nature du rythme vocal y sera apportée, ce qui nous permettra de définir une structure rythmique inter-linguistique invariable pour le séquencement du rythme vocal. Nous y présenterons les différentes méthodes et interfaces de contrôle que nous avons pu explorer pour permettre le séquencement du rythme vocal.

Le Chapitre 4 traite la question de la pertinence de la tablette graphique pour le contrôle de la hauteur et des paramètres de qualité vocale, dans une époque où de nouvelles interfaces continues de contrôle musical voient régulièrement le jour.

Le Chapitre 5 présente le logiciel Vokinesis. Nous verrons comment il fonctionne d'un point de vue général, et nous détaillerons son architecture, afin de définir la façon dont les différents éléments du système communiquent les uns avec les autres. Les stratégies de mapping que nous avons mises en place seront présentées, ainsi que certains détails de programmation du cœur du système. La dernière section présentera le fonctionnement complet de Vokinesis.

Dans le Chapitre 6, nous présenterons les utilisations musicales que nous avons pu explorer avec Vokinesis. Une première section sera concentrée sur les différentes représentations publiques du Chorus Digitalis, un ensemble de voix synthétiques faisant usage de Vokinesis et du Cantor Digitalis. La seconde section présentera la façon dont Vokinesis peut être utilisé dans un cadre de composition assistée par ordinateur.

Enfin, le Chapitre 7 propose une conclusion et des perspectives d'applications.

Chapitre 2

# VoPTiQ : Voice Pitch, Time and Quality modification

## Sommaire



Afin de permettre la modification temps-réel de hauteur, de durée, et de paramètres de qualité vocale de signaux de voix, nous avons mis en place la méthode de traitement de signal VoPTiQ (Voice Pitch, Time and Quality modification), dont le schéma fonctionnel est représenté Figure 2.1. Cette méthode est d'abord composée d'une extensions de l'algorithme RT-PSOLA (VRT-PSOLA sur la figure), permettant un allongement la modification de durée, de hauteur, et de taille du conduit


FIGURE 2.1 – *Schéma de principe de VoPTiQ. Un signal d'entrée $x(n)$ est modifié par l'algorithme VRT-PSOLA, dont le résultat est filtré par des effets de modification de la tension et de l'effort vocal, pour obtenir le signal modifié $y(n)$.*

vocal. Des techniques de filtrage sont ensuite appliquées au signal $y_p(n)$ (en sortie de VRT-PSOLA sur la figure), qui permettent de simuler des effets d'atténuation ou d'augmentation de la tension et de l'effort vocal.

Le première section a pour objectif de présenter les principales techniques de modification de hauteur et de durée de la voix. Dans une seconde section, nous présenterons les détails de l'algorithme TD-PSOLA [Hamon *et al.* 1989, Moulines & Charpentier 1990, Moulines & Laroche 1995], l'ancêtre de son équivalent temps-réel RT-PSOLA [Le Beux *et al.* 2010]. La troisième section décrira l'algorithme VRT-PSOLA, une version améliorée de RT-PSOLA, notamment pour ce qui concerne l'allongement des parties non-voisées et la modification de la taille du conduit vocal. La dernière section présentera les techniques de filtrage que nous avons mises en place pour modifier les paramètres de tension et d'effort vocal.

## 2.1 Transformation d'un signal vocal

L'identité d'un signal vocal est constituée de deux paramètres acoustiques principaux. Sa *hauteur*, ou *fréquence fondamentale* ($f_0$), correspond à la fréquence de vibration des plis vocaux. Son *enveloppe spectrale* définit le *timbre* de la voix, qui varie selon la forme du conduit vocal. Les plis vocaux, en vibrant, créent une onde acoustique (l'*Onde de Débit Glottique*, ou ODG), qui, en se propageant dans le conduit vocal, subit une amplification de certaines bandes de fréquences, qui correspondent aux maximums spectraux du filtre permettant de modéliser le conduit vocal (ces maximums spectraux sont appelées *formants*). La fréquence centrale et la largeur de bande des formants varie selon la forme du conduit vocal. Lorsqu'un son est produit en l'absence de vibration des plis vocaux, on dit qu'il est *non-voisé*. La source sonore correspond alors à un bruit créé par l'écoulement turbulent de l'air dans une constriction étroite du conduit vocal. Certains sons voisés peuvent être le résultat d'un mélange de la source glottique et d'une source de bruit. Le modèle source/filtre de [Fant 1970], présenté dans la FIGURE 2.2, résume ce que nous venons d'énoncer : la source glottique et la source de bruit sont additionnées puis filtrées par les résonances du conduit vocal pour être transformées en signal vocal.

L'enjeu de la transformation vocale consiste à permettre la modification d'un paramètre de durée, de source ou d'enveloppe spectrale tout en conservant les caractéristiques des autres. Cette tâche n'est pas évidente. Par exemple, si un allongement de durée est effectué par un simple ralentissement du signal, alors sa fréquence

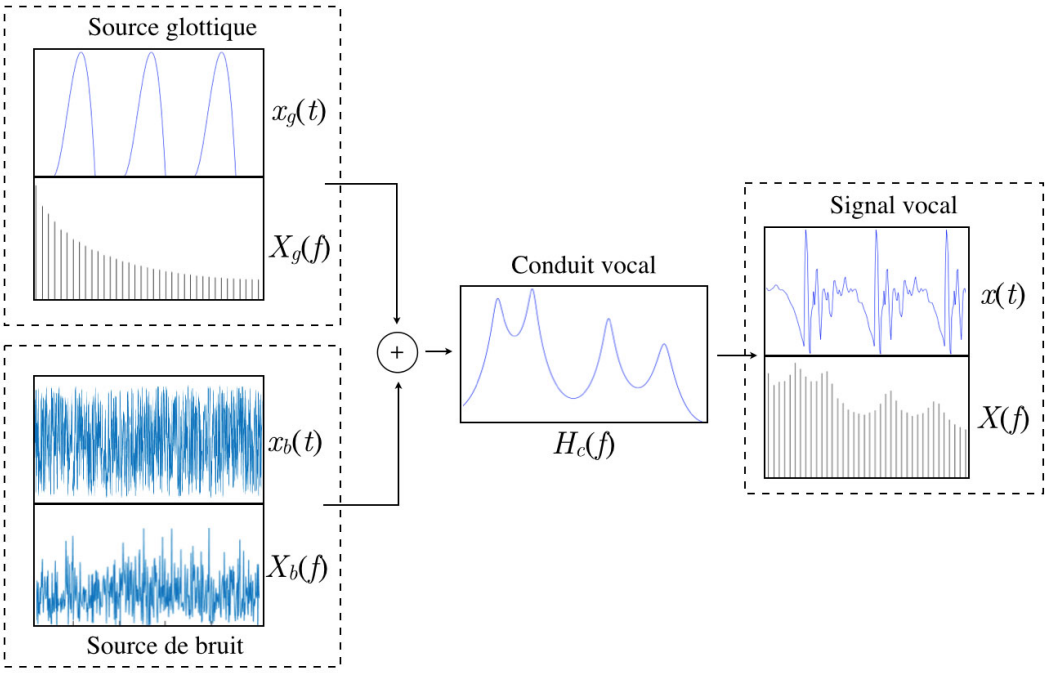


FIGURE 2.2 – *Modèle source/filtre de* [Fant 1970]. *Image adaptée de* [Henrich 2001].

fondamentale et la fréquence centrale de ses formants seront diminuées. Dans cette section, nous fournissons une revue des techniques de transformation vocale les plus répandues.

### 2.1.1 Vocoders

Le principe du vocoder consiste à décomposer un signal de parole original afin qu'il puisse être représenté par des signaux moins lourds à transmettre. Ces signaux représentatifs sont ensuite réutilisés dans une étape de re-synthèse pour reproduire le signal vocal original. Ce principe a d'abord été mis en place pour des applications de téléphonie, puis a plus tard été revisité pour des applications de transformation vocale. Nous présentons ci-dessous différentes techniques de vocoders.

#### 2.1.1.1 Vocoders à canaux

Le principe des vocoders à canaux consiste à extraire les informations de hauteur et d'enveloppe spectrale d'un signal d'entrée. Les paramètres ainsi obtenus peuvent ensuite être modifiés et utilisés pour obtenir un signal de synthèse, comme le montre la FIGURE 2.3.

Le premier vocoder à canaux (ou *channel vocoder*) a été introduit par [Dudley 1939]. C'était un système électronique qui transformait un signal d'entrée en 1 tension représentative de sa fréquence fondamentale et 10 autres d'un niveau d'énergie par bandes de fréquences. Ces tensions étaient ensuite réutilisée comme

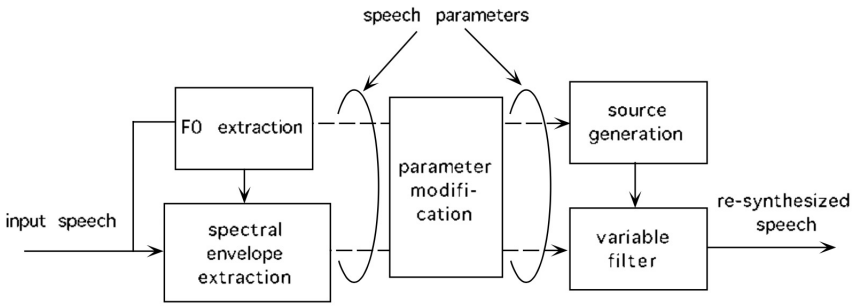


Figure 2.3 – *Principe du vocoder à canaux : les paramètres vocaux d'un signal d'entrée sont extraits et réutilisés pour la re-synthèse* [Kawahara *et al.* 1999]

commandes d'entrées d'un synthétiseur composé d'un oscillateur et de 10 filtres passe-bande.

L'estimation de l'enveloppe spectrale peut s'effectuer de différentes manières. Parmi elles, nous pouvons citer l'estimation par prédiction linéaire (*Linear Predictive Coding*, LPC) [Itakura 1970, Atal & Hanauer 1971], par paires de lignes spectrales (*Line Spectrum Pairs*, LSP) [Itakura 1975], ou par analyse cepstrale [Schafer & Rabiner 1970, Imai 1983, Tokuda *et al.* 1994].

D'après [Kawahara *et al.* 1999], la qualité de la synthèse de ce genre de systèmes diminue lorsque les paramètres de modification augmentent. La principale cause de cette perte de qualité serait due aux erreurs d'estimation spectrale, qui contiennent des interférences liées à la périodicité du signal original. Une autre cause de perte de qualité serait liée aux courbes d'analyse de la fréquence fondamentale discontinues. L'auteur propose d'éviter ces problèmes à travers un algorithme nommé STRAIGHT (*Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum*). En prenant en compte la fréquence fondamentale du signal original, cet algorithme permet de résoudre les problèmes d'analyse que nous venons de citer. Il est aujourd'hui très largement utilisé par la communauté en raison de la qualité des signaux de synthèse qu'il permet d'obtenir.

Très récemment est apparu le vocoder WORLD [Morise *et al.* 2016], qui, en plus des informations de fréquence fondamentale et d'enveloppe spectrale, extrait des paramètres d'apériodicité du signal original. Il semblerait que sa qualité soit supérieure à celle de STRAIGHT, ce qui en fait un vocoder très prometteur.

#### 2.1.1.2 Vocoder de phase

Cette section s'appuie en grande partie sur [Liuni & Röbel 2013] qui fournit un historique très complet de l'évolution du vocoder de phase. Le premier vocoder de phase a été introduit par [Flanagan & Golden 1966]. Le principe de l'analyse consiste à appliquer une STFT (*Short-Time Fourier Transform*, ou transformée de Fourier à court-terme) au signal original, permettant d'obtenir sa représentation fréquentielle sur un ensemble de trames temporelles consécutives (ou encore sa représentation temps-fréquence). La modification temporelle consiste à

modifier le nombre de trames d'analyse du signal original : plus le signal sera allongé, plus les trames d'analyse seront recouvertes les unes aux autres. La modification de hauteur consiste à étirer ou à comprimer chaque trame, puis à modifier la fréquence d'échantillonnage afin de conserver l'enveloppe spectrale d'origine. Lors d'éventuelles modifications de l'échelle temporelle, les phases des STFT doivent donc être adaptées pour assurer une cohérence dans l'ajout-recouvrement des composantes sinusoïdales. [Laroche & Dolson 1999] proposent également l'utilisation d'une technique de modification de hauteur basée sur le principe des modèles sinusoïdaux, que nous présenterons dans la section suivante. Un autre problème majeur des vocoders de phase est lié aux parties transitoires du signal original, où les méthodes de cohérence de phase entre les trames successives ne sont plus appropriées. [Bonada 2000, Duxbury *et al.* 2002, Röbel 2003] ont exploré différentes techniques de ré-initialisation de la phase des transitoires pour résoudre ces problèmes. Les dernières améliorations apportées aux vocoders de phase ont pour but d'obtenir une résolution optimale de leur représentation temps-fréquence. Certains utilisent des fenêtres d'analyse dont la taille est adaptée à chaque trame [Rudoy *et al.* 2010, Liuni *et al.* 2011b, Balazs *et al.* 2011], d'autres adaptent la taille de chaque bande de fréquence au sein d'une même trame [Evangelista *et al.* 2012], et d'autres encore combinent les deux approches [Jaillet & Torrésani 2007, Dörfler 2011, Liuni *et al.* 2011a].

### 2.1.2 Modèles sinusoïdaux

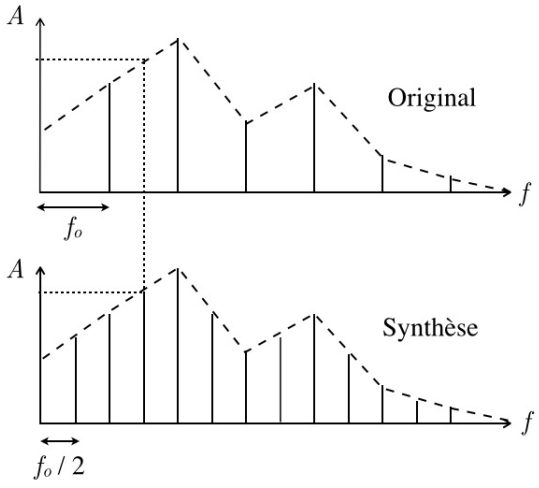


FIGURE 2.4 – *Décomposition sinusoïdale d'un signal original (en haut) et re-synthèse par interpolation des composantes sinusoïdales (en bas). la fréquence du signal de synthèse à été divisée par 2 par rapport à celle du signal original.*

Le principe des modèles sinusoïdaux, introduit par [McAulay & Quatieri 1986] et inspiré par le vocoder à canaux [Dudley 1939], est de décomposer un signal de parole en une somme de composantes sinusoïdales, définies par leur amplitude, leur

fréquence et leur phase, à partir de sa représentation temps-fréquence obtenue par STFT. La différence avec les vocoders de phase tient du fait que chacune des trames de la STFT se voit appliquer un algorithme de détection de pics permettant de déterminer les fréquences et amplitudes de leurs composantes sinusoïdales. Ceci est appliqué sur tout le signal de parole, qu'il soit voisé ou non. La Figure 2.4 montre comment s'effectue une transformation de la hauteur d'un signal original sans en affecter l'enveloppe spectrale, selon la méthode proposée par [Quatieri & McAulay 1986]. Les fréquences des harmoniques de synthèse sont calculées selon la fréquence de synthèse désirée, puis leurs phases et amplitudes sont déterminées par interpolation linéaire des phases et amplitudes originales.

D'après les auteurs, ce modèle ne se limite pas à la re-synthèse d'un locuteur unique. Il permettrait de re-synthétiser à la perfection des signaux multi-locuteurs, musicaux, bruités, d'animaux marins... Cependant, bien qu'il offre une représentation satisfaisante de la parole, il n'est pas tout à fait adapté à la modification temporelle de signaux non-voisés, car ils sont apériodiques. [Serra & Smith 1990, Stylianou 1996] ont alors proposé le modèle HNM (*Harmonic plus Noise Model*, ou modèle harmonique plus bruit), un modèle hybride qui fait intervenir une source de bruit pour la synthèse de sons non-voisés. Dans son modèle sinus + transitoire + bruit (*sine + transient + noise*), [Levine & Smith III 1998] propose de considérer, en plus des composantes harmoniques et aléatoires, les parties transitoires du signal original, afin de conserver leur qualité.

### 2.1.3 PSOLA

La méthode PSOLA (Pitch-Synchronous Overlap-Add), présentée en détails dans [Moulines & Charpentier 1990, Moulines & Laroche 1995], constitue la base des méthodes de transformation temps-réel que nous utilisons dans Vokinesis. Elle consiste à décomposer le signal original en trames d'analyse successives afin de les réorganiser en trames de synthèse pour les modifications de durée et de hauteur. La différence avec les vocoders de phase tient du fait que les trames d'analyse sont obtenues de façon synchrone à la fréquence fondamentale : chaque trame est représentative d'une période du signal original. La taille des trames d'analyse dépend de la technique utilisée, mais elle se définit par un multiple de la durée de la période qu'elle représente. Deux techniques principales de PSOLA peuvent être citées : TD-PSOLA (Time-Domain PSOLA) et FD-PSOLA (Frequency-Domaine PSOLA). La technique FD-PSOLA consiste à appliquer une méthode de type modèle sinusoïdal sur chaque trame d'analyse. Elle est adaptée à des applications de modification de hauteur et d'enveloppe spectrale, mais ne convient pas aux modifications de durée. La technique TD-PSOLA consiste simplement à réorganiser les trames d'analyse en trames de synthèse, sans modification préalable, et sans passer par le domaine fréquentiel. Une version temps-réel de TD-PSOLA, nommée RT-PSOLA (Real-Time PSOLA) a été proposée par [Le Beux *et al.* 2010]. La technique de modification vocale que nous utilisons dans Vokinesis est une adaptation de cette méthode, qui contient une amélioration du traitement des signaux non-voisés, et qui offre en plus

une possibilité d'effectuer des modifications d'enveloppe spectrale. Tout ceci sera présenté en détails plus loin.

L'avantage principal de PSOLA vient du fait qu'il est très peu coûteux en temps de calcul. Il s'implémente donc en temps-réel sans aucun problème. Son inconvénient vient de sa phase d'analyse synchrone à la fréquence fondamentale. En effet, cette phase nécessite une pré-analyse de périodicité du signal original afin de définir l'emplacement de chaque période pour le calcul de chaque trame d'analyse. La moindre erreur d'analyse de périodicité (omission ou ajout de périodes, mauvaise détection des parties voisées ou non-voisées) détériore la synthèse de façon notoire. Cette méthode nécessite donc parfois une correction manuelle des données d'analyse de périodicité.

Nous pouvons également citer la technique MBROLA [Dutoit *et al.* 1996] qui consiste à effectuer des modifications préalables de bases de données destinées à la synthèse par concaténation avec TD-PSOLA.

### 2.1.4 Expressivité et modification de la qualité vocale

Le naturel de la synthèse vocale peut être amélioré de manière significative si des modulations expressives y sont apportées [Umbert *et al.* 2015]. Outre les variation de fréquence fondamentale (vibrato, tremblements...) et de durée des phonèmes, qui peuvent être modifiées par les algorithmes que nous avons présentés plus haut, la modification de paramètres de qualité vocale peut jouer un rôle très important dans le contrôle expressif de la parole [Evrard 2015] et du chant [Umbert *et al.* 2015]. La qualité vocale correspond à des paramètres acoustiques différents de la fréquence fondamentale ou des phonèmes. Nous pouvons en citer quatre principaux :

- L'***effort vocal*** représente l'intensité d'une production vocale. Une augmentation de l'effort vocal provoque une augmentation de l'amplitude du signal vocal d'une part, mais également une augmentation de l'énergie dans les hautes fréquences [Sundberg & Rossing 1990, Titze & Sundberg 1992], qui peut être modélisée par un filtre de pente spectrale [Doval & d'Alessandro 1997, Doval *et al.* 2006]. [Perrotin & d'Alessandro 2016b] proposent une méthode d'augmentation de l'effort vocal basée sur l'ajout d'harmoniques en hautes fréquences par distorsion temporelle.
- La ***tension vocale*** représente la quantité de tension appliquée aux plis vocaux. D'après [Henrich 2001], une augmentation de la tension vocale a pour effet d'augmenter la fréquence centrale du *formant glottique*, qui correspond à la fréquence de coupure d'un filtre passe bas du deuxième ordre permettant de modéliser le spectre de l'onde de débit glottique.
- Le ***souffle*** correspond au niveau d'apériodicité contenu dans un signal vocal. Une voix chuchotée ne contient que du souffle : les plis vocaux ne vibrent pas. Une voix sensuelle est souvent caractérisée par un fort niveau de souffle, couplé à une légère vibration des plis vocaux [Evrard 2015]. L'utilisation d'un vocoder prenant en compte de paramètres d'apériodicité tel que WORLD [Morise *et al.* 2016] pourrait permettre de modifier le niveau de souffle.

- La ***taille du conduit vocal*** (distance entre les plis vocaux et les lèvres) peut être allongée ou rétrécie par un déplacement du larynx. En abaissant son larynx, un chanteur lyrique donne du coffre à sa voix. En effet, en terme de signal, un allongement du conduit vocal crée une diminution des fréquences centrales des formants. Un conduit vocal très grand donnera un timbre vocal très grave (voix de géant), et un conduit vocal très petit donnera un timbre très aigu (voix de souris). L'effet d'allongement du conduit vocal peut être simulé par un ré-échantillonnage du signal : un sous-échantillonnage augmentera les fréquences des formants, et un sur-échantillonnage les diminuera.

## 2.2 TD-PSOLA

Le fait d'accélérer ou de ralentir un signal original afin de modifier sa durée aura également un effet sur sa hauteur et sur son enveloppe spectrale, comme le montre la FIGURE 2.5 : une accélération augmentera les fréquences de ses harmoniques et de son enveloppe spectrale, et un décélération les diminuera.

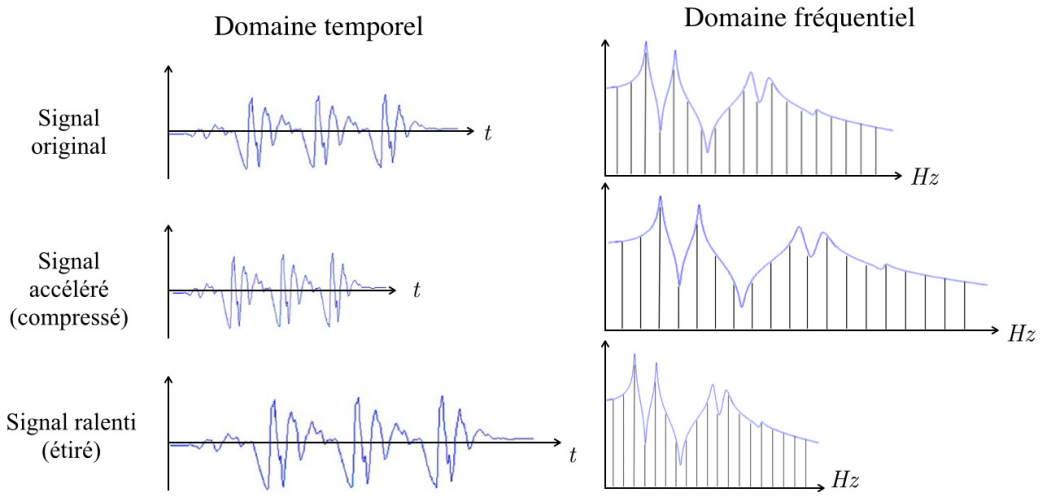


FIGURE 2.5 – *La modification du durée d'un signal original a pour effet de modifier les fréquences de ses harmoniques et de son enveloppe spectrale.*

Cette section résume une partie de [Moulines & Charpentier 1990, Moulines & Laroche 1995], qui décrit l'algorithme TD-PSOLA (Time Domain Pitch Synchronous Overlap and Add). Cet algorithme permet de modifier en temps différé et de façon indépendante la durée et la hauteur d'un signal de parole original, sans affecter son enveloppe spectrale. Son principe est représenté FIGURE 2.6. Il s'agit de réorganiser les périodes originales en périodes de synthèse. Pour augmenter/diminuer la durée, certaines périodes seront dupliquées/supprimées. Pour augmenter/diminuer la fréquence fondamentale, les périodes de synthèse seront rapprochées/écartées les unes des autres. Il faudra donc décomposer le signal original en trames d'analyse successives $x(i, n)$ (avec $i$ le numéro de période

du signal original) afin de les réorganiser en trames de synthèse $y(j,n)$ (avec $j$ le numéro de période de synthèse). Les trames d'analyse sont obtenues de façon synchrone à la fréquence fondamentale du signal original : chaque trame est représentative d'une de ses périodes. La taille des trames d'analyse correspond à deux fois la durée de la période qu'elle représente. Une trame d'analyse est obtenue en multipliant le signal original par une fenêtre de Hann.

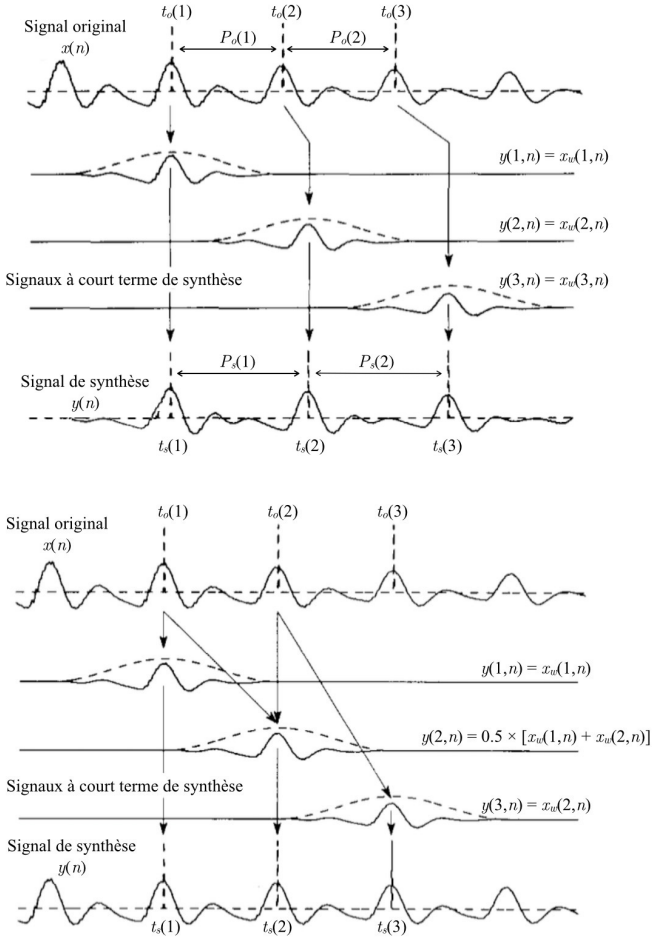


FIGURE 2.6 – *Modification en temps différé de la hauteur (en haut) et de la durée (en bas) d'un signal voisé par l'algorithme TD-PSOLA. Le repositionnement des périodes originales vers leurs trames de synthèse correspondants est représenté par les flèches. Le facteur de modification de hauteur est de 0.8, celui de durée est de 2. Figures adaptées de* [Moulines & Laroche 1995].

### 2.2.1 Préparation des données d'analyse de périodicité

Pour permettre le calcul des trames d'analyse synchrone à la fréquence fondamentale du signal original, il est nécessaire d'effectuer une phase de pré-analyse de périodicité. Tout d'abord, il convient de détecter les parties périodiques (ou voisées)

et les parties apériodiques (non-voisées) du signal original. En effet, nous verrons par la suite que le traitement ne s'effectue pas de la même manière selon l'état de voisement. Ceci fait, la détection de périodes peut être effectuée. Cette étape consiste à définir la position de chaque période par un point unique, noté $t_o(i)$ (voir la Figure 2.6). Les marqueurs périodiques consécutifs seront donc séparés par une durée $P_o(i)$ correspondant à la période originale dans les parties voisées. Les parties non-voisées étant apériodiques, les marqueurs apériodiques seront séparés d'une durée fixe (par exemple $P_o(i) = 5$ms). Les fenêtres d'analyse auront une taille $N_w(i) = 2P_o(i)$ : elles envelopperont une double période autour de $t_o(i)$.

Pour effectuer notre analyse de périodicité ainsi que la détection voisé/non-voisé, nous utilisons le logiciel Praat, spécialisé dans les applications d'analyse de la voix.

### 2.2.2 Calcul des trames d'analyse

Tout d'abord, le signal original $x(n)$ est décomposé en un ensemble de trames d'analyse $x_w(i,n)$ ($i$ correspond à l'indice de la période originale et $n$ à l'indice temporel des signaux discrets). Il s'agit de fenêtrer les double-périodes qui entourent les marqueurs de périodes originales $t_o(i)$, selon l'équation (2.1).

$$x_w(i,n) = w(i,n) \times x(i,n) \tag{2.1}$$

où $w(i,n)$ est une fenêtre de Hann centrée autour de $t_o(i)$ et dont la taille $N_w$ est égale à celle de $x(i,n)$, défini par l'équation (2.2) :

$$x(i,n) = x(n), \quad n \in [t_o(i-1), t_o(i+1)] \tag{2.2}$$

Le signal de synthèse sera le résultat de la réorganisation des trames $x_w(i,n)$. Cela consistera à déterminer l'emplacement des marqueurs périodiques de synthèse $t_s(j)$ (où $j$ correspond à l'indice des périodes de synthèse), et à recentrer les trames autour de chacun de ces marqueurs.

### 2.2.3 Déformation de l'échelle temporelle

La déformation temporelle d'un signal original consiste à modifier sa durée sans en affecter sa hauteur ou son enveloppe spectrale. Pour ce faire, il s'agit tout d'abord d'associer à chaque marqueur périodique original $t_o(i)$ un facteur de déformation temporelle $\delta(i)$, dont un exemple est représenté en bas de la Figure 2.7 : la durée de synthèse correspondra au double de la durée originale pour les périodes 1 et 2, et elle restera inchangée pour les autres périodes. La fonction de déformation temporelle $D(t)$ peut ensuite être déterminée grâce à l'équation (2.3) :

$$D(t) = \int_0^t \delta(t)dt. \tag{2.3}$$

Sur la Figure 2.8, $\delta(t)$ est constante. La fonction $D(t)$ est donc purement linéaire. Elle permet d'obtenir l'axe temporel de synthèse à partir de l'axe temporel original (flèches à gauche de la figure).

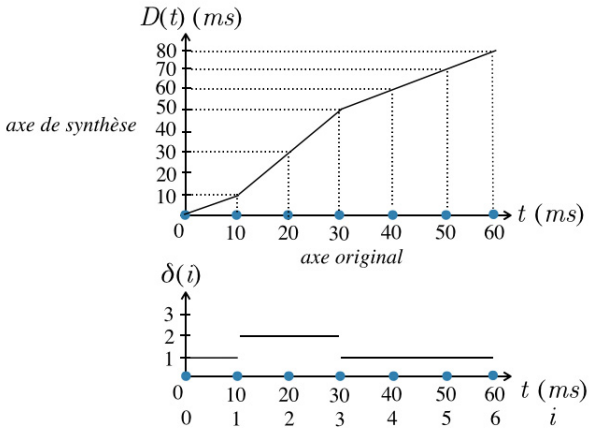


FIGURE 2.7 – *Exemple de fonction de déformation temporelle. $D(t)$ est obtenue par l'équation (2.3). Les points représentés sur les axes des abscisses représentent les marqueurs périodiques d'un signal original avec une fréquence fondamentale de 100Hz.*

La prochaine étape consiste à déterminer l'emplacement des marqueurs périodiques de synthèse $t_s(j)$ de façon à ce que l'équation (2.4) soit vérifiée :

$$t_s(j+1) = t_s(j) + P_s(t_s(j)) \tag{2.4}$$

où $P_s(t_s(j))$ correspond à la durée de la $j^e$ période de synthèse. L'évolution de la durée de chacune des périodes de synthèse doit suivre celle des périodes originales, de façon à ce que l'équation (2.5) soit respectée :

$$P_s(t) = P_o(D^{-1}(t)) \tag{2.5}$$

où $P_o(t)$ correspond à la durée de la période originale à l'instant $t$, et est déterminée par l'équation (2.6) :

$$P_o(t_o(i)) = t_o(i+1) - t_o(i) \tag{2.6}$$

Afin de déterminer la durée d'une période de synthèse, il est nécessaire d'utiliser un axe temporel virtuel permettant de remettre l'axe temporel de synthèse à l'échelle de l'axe temporel original. Ainsi sera obtenu un ensemble de marqueurs périodiques virtuels $t_v(j)$ tels que $t_s(j) = D(t_v(j))$ et $t_v(j) = D^{-1}(t_s(j))$, comme le montre la FIGURE 2.8. Le calcul de la durée d'une période de synthèse doit correspondre à la moyenne des périodes de l'axe original contenues dans une période de l'axe virtuel, ce qui se traduit par l'équation (2.7) :

$$P_s(t_s(j)) = \frac{1}{t_v(j+1) - t_v(j)} \int_{t_v(j)}^{t_v(j+1)} P_o(t)dt \tag{2.7}$$

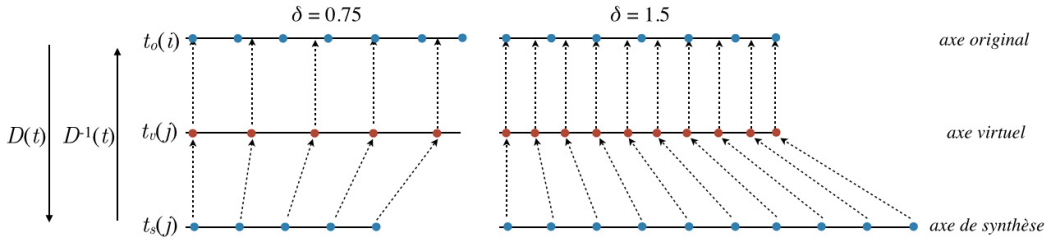


FIGURE 2.8 – *Exemple de déformation temporelle par un facteur* $\delta = 0.75$ *(à gauche) et* $\delta = 1.5$ *(à droite). Ici, la hauteur n'est pas modifiée, nous avons donc* $\phi = 1$.

### 2.2.4 Déformation de l'échelle mélodique

La déformation de l'échelle mélodique consiste à modifier la fréquence fondamentale d'un signal original sans en affecter sa durée ni son enveloppe spectrale. Ici, puisque la durée n'est pas modifiée, les positions des marqueurs de synthèse et virtuels sont identiques, comme le montre la FIGURE 2.9.

Le principe est d'associer à chaque période originale $P_o(t_o(i))$ un facteur de multiplication de la fréquence fondamentale $\phi(t_o(i))$ pour obtenir la courbe mélodique de synthèse désirée. Il faudra alors déterminer les emplacements des marqueurs périodiques de synthèse afin que l'équation (2.8) soit respectée :

$$t_s(j+1) = t_s(j) + P_s(t_s(j)) \tag{2.8}$$

où la période de synthèse $P_s(t_s(j))$ doit être proche de la période originale correspondante $P_o(t_s(j))$ divisée par $\phi(t_s(j))$, selon l'équation (2.9) :

$$P_s(t_s(j)) \approx \frac{P_o(t_s(j))}{\phi(t_s(j))} \tag{2.9}$$

La durée exacte de la période de synthèse $P_s(t_s(j))$ devra correspondre à la moyenne des durées transformées des périodes de l'axe original contenues dans une période de l'axe de synthèse. Ceci se traduit par l'équation (2.10).

$$P_s(t_s(j)) = \frac{1}{t_s(j+1) - t_s(j)} \int_{t_s(j)}^{t_s(j+1)} \frac{P_o(t)}{\phi(t)} dt \tag{2.10}$$

### 2.2.5 Déformation simultanée des échelles temporelle et mélodique

En connaissant les fonctions de transformation temporelle et mélodique $D(t)$ et $\phi(t)$ définies dans les sections précédentes, et en utilisant l'axe virtuel au lieu de l'axe de synthèse dans l'équation (2.10) (donc en combinant (2.7) et (2.10)), nous obtenons l'équation (2.11), qui permet de déterminer la durée d'une période de synthèse lors de transformations temporelle et mélodique simultanées :

$$P_s(t_s(j)) = \frac{1}{t_v(j+1) - t_v(j)} \int_{t_v(j)}^{t_v(j+1)} \frac{P_o(t)}{\phi(t)} dt \tag{2.11}$$

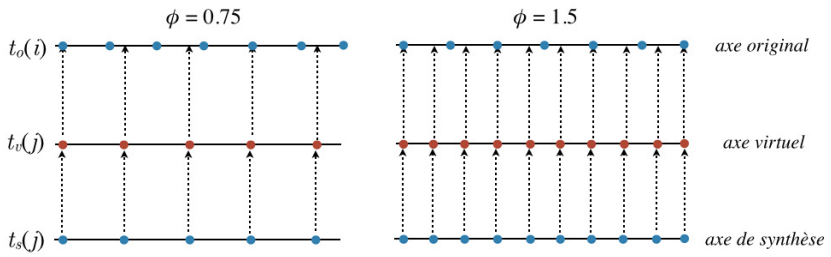


FIGURE 2.9 – *Exemple de déformation mélodique par un facteur* $\phi = 0.75$ *(à gauche) et* $\phi = 1.5$ *(à droite). Ici, la durée n'est pas modifiée, nous avons donc* $\delta = 1$.

### 2.2.6 Association des marqueurs périodiques et calcul d'une période de synthèse

L'étape suivante consiste à associer les marqueurs périodiques de synthèse $t_s(j)$ aux marqueurs périodiques originaux $t_o(i)$, comme le montre les FIGURES 2.8 et 2.9 : on calcule d'abord les positions des marqueurs périodiques virtuels avec $t_v(j) = D^{-1}(t_s(j))$, puis on vérifie leur emplacement sur l'axe original. Si $t_o(i) < t_v(j) < t_o(i+1)$, alors $y(j,n)$ correspondra à un mélange entre $x_w(i,n)$ et $x_w(i+1,n)$. Si $t_v(j) = t_o(i)$, alors nous aurons $y(j,n) = x_w(i,n)$. Ceci se traduit par l'équation (2.12) :

$$y(j,n) = (1-\alpha) \times x_w(i,n) + \alpha \times x_w(i+1,n) \tag{2.12}$$

où $\alpha(j)$ est le facteur d'interpolation permettant de mélanger deux périodes originales consécutives dans une période de synthèse, et est défini par l'équation (2.13) :

$$\alpha(j) = \frac{t_v(j) - t_o(i)}{t_o(i+1) - t_o(i)} \tag{2.13}$$

### 2.2.7 Déformation temporelle de signaux non voisés

Nous avons vu dans les parties précédentes que, lors d'un étirement temporel, l'algorithme TD-PSOLA pouvait dupliquer certaines périodes. Or, les parties non-voisées d'un signal de parole sont apériodiques. Le fait de répéter une petite section d'un signal apériodique dans des intervalles réguliers entraine un bruit tonal indésirable dans le signal de synthèse. La solution proposée dans [Moulines & Laroche 1995] permet d'étirer un signal non voisé jusqu'à 2 fois sa taille originale : $\delta_{max} = 2$. Cette méthode consiste à inverser temporellement chaque répétition successive d'une portion d'un signal non-voisé. Ainsi, si $y(j,n) = x(i,n)$ alors $y(j+1,n) = x(i,-n)$. Nous verrons dans le section 2.3.2 une méthode que nous avons développée pour le temps-réel (mais qui peut être utilisée en temps différé), qui consiste à utiliser des durées aléatoires des portions d'un signal non-voisé afin de se débarrasser du bruit tonal même lors d'un étirement temporel infini.

## 2.3 Modification en temps-réel de la hauteur, de la durée, et de la longueur du conduit vocal : VRT-PSOLA

Pour le temps-réel, les paramètres de déformation temporelle et mélodique ne sont connus qu'à l'instant présent, c'est à dire à l'instant où l'utilisateur les choisit. Donc, contrairement au temps différé, les fonctions de déformation temporelle et mélodique $D(t)$ et $\phi(t)$ sont complètement inconnues. L'utilisateur doit alors être en mesure, à tout moment, de décider quelle partie du signal original $x(n)$ il souhaite synthétiser, mais également la fréquence à laquelle il veut qu'elle soit synthétisée. Celui-ci contrôlera donc deux paramètres :

— l'instant cible dans le signal original, noté $\tau(j)$
— la fréquence de synthèse notée $f_s(j)$

Comme dans la partie précédente, les indices $i$ correspondent aux périodes du signal original et les indices $j$ aux périodes du signal de synthèse.

### 2.3.1 Modification temps-réel de signaux voisés

La Figure 2.10 décrit la façon dont un signal voisé peut être modifié en temps-réel, grâce à l'algorithme RT-PSOLA (Real-Time Pitch Synchronous Overlap-Add) [Le Beux *et al.* 2010]. Comme nous l'avons vu dans le section 2.2 pour l'algorithme TD-PSOLA original, les coefficients de modification en temps différé sont définis sur l'axe temporel original. Pour le temps-réel, ces paramètres sont définis sur l'axe de synthèse, car ils doivent être calculés à l'instant présent pour chaque nouvelle période de synthèse : aucune prédiction ne peut être faite sur les paramètres de contrôle futurs.

Les deux paramètres contrôlés par l'utilisateur sont $\tau(j)$ et $P_s(j)$. À chaque nouvelle période, l'instant $\tau(j)$ du signal original ciblé par l'utilisateur est mis à jour. La trame de synthèse $y(j,n)$ est alors calculée selon l'équation (2.14), en choisissant $i$ comme étant l'indice du marqueur de période original qui précède l'instant pointé par $\tau(j)$ :

$$y(j,n) = w(i,n) \times ((1-\alpha) \times x(i,n) + \alpha \times x(i+1,n)) \tag{2.14}$$

où $\alpha(j)$ est un facteur d'interpolation défini par l'équation (2.15) :

$$\alpha(j) = \frac{\tau(j) - t_o(i)}{t_o(i+1) - t_o(i)} \tag{2.15}$$

Ce facteur permet à l'utilisateur de faire évoluer $\tau(j)$ très lentement entre deux marqueurs de période $t_o(i)$ et $t_o(i+1)$ en évitant une transition trop brusque lors du passage d'une période à la suivante.

Chaque trame de synthèse est ensuite ajoutée au signal de sortie $y_p(n)$ avec un espacement temporel défini par $P_s(j)$. Les marqueurs périodiques de synthèse sont donc mis à jour à chaque nouvelle période de synthèse selon l'équation (2.16) :

$$t_s(j) = \sum_{k=0}^{j} P_s(k) \tag{2.16}$$

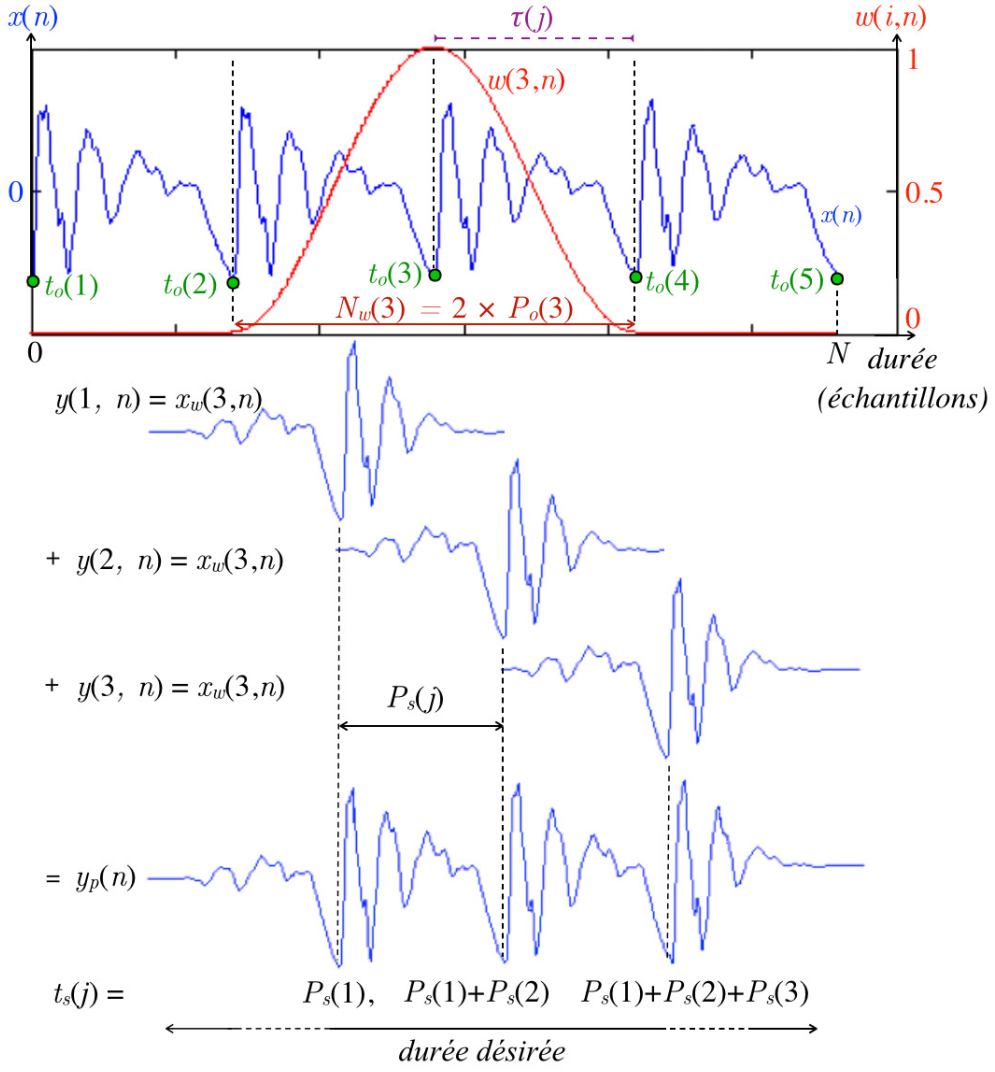


FIGURE 2.10 – *Modification temps-réel de la durée et de la hauteur d'un signal voisé par l'algorithme RT-PSOLA. La double période autour du marqueur de période $t_o(3)$ d'un signal original $x(n)$ est fenêtrée par la fenêtre de Hann $w(3,n)$ et ajoutée au signal de synthèse $y(n)$ pour une durée indéfinie à chaque période de synthèse $t_s(j)$ dont la durée est définie par $P_s(j)$.*

Dans la FIGURE 2.10, l'utilisateur maintient la double période $t_o(3)$ pendant la *durée désirée*, avec une période de synthèse $P_s(j)$ proche de la période originale $P_o(3)$.

### 2.3.2 Modification temps-réel de signaux non-voisés

La méthode de traitement temps-réel des signaux voisés que nous venons de présenter constituait l'algorithme RT-PSOLA [Le Beux *et al.* 2010]. Toutes les méthodes que nous allons présenter ci-dessous ont été développées dans le cadre de cette thèse, et constituent l'algorithme VRT-PSOLA (Vokinesis RT-PSOLA).

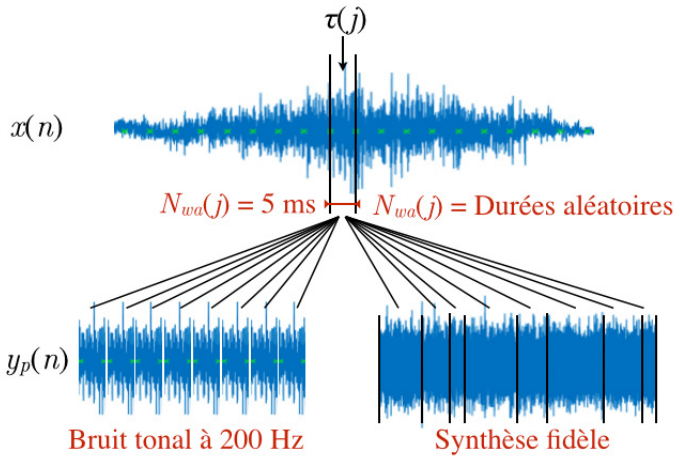


FIGURE 2.11 – *Principe de l'algorithme VRT-PSOLA pour les sons non-voisés. Des durées aléatoires sont utilisées afin d'éviter les bruits tonaux lors de l'étirement indéfini de sons non-voisés.*

Comme le montre la FIGURE 2.11, et comme expliqué dans la section 2.2.7, la répétition régulière d'une partie non-voisée d'un signal vocal crée un bruit tonal indésirable, dont la fréquence est définie par l'inverse de la durée de la partie répétée. Pour une transformation en temps différé, le fait d'inverser temporellement chaque répétition consécutive d'une même partie permet d'atteindre un facteur d'étirement maximal $\delta_{max} = 2$ [Moulines & Laroche 1995]. Cependant, dans le cas du temps réel, un facteur fini d'étirement temporel n'est plus suffisant : l'utilisateur doit être en mesure de maintenir une seule et même partie du signal pour une durée indéfinie. Ce problème a été résolu par l'utilisation de durées aléatoires pour chaque répétition d'une partie non-voisée : à chaque itération, une trame de durée aléatoire est calculée selon l'équation (2.17) :

$$y(j,n) = x\left[\tau(j) - \frac{N_{wa}(j-1)}{2}, \tau(j) + \frac{N_{wa}(j)}{2}\right] \times w_a(j,n) \times \rho_a \tag{2.17}$$

où $\rho_a$ est un facteur de puissance défini par l'équation (2.21) (nous y reviendrons plus tard), $N_{wa}$ est tirée de façon aléatoire (distribution uniforme) entre $max(N_{wa})$

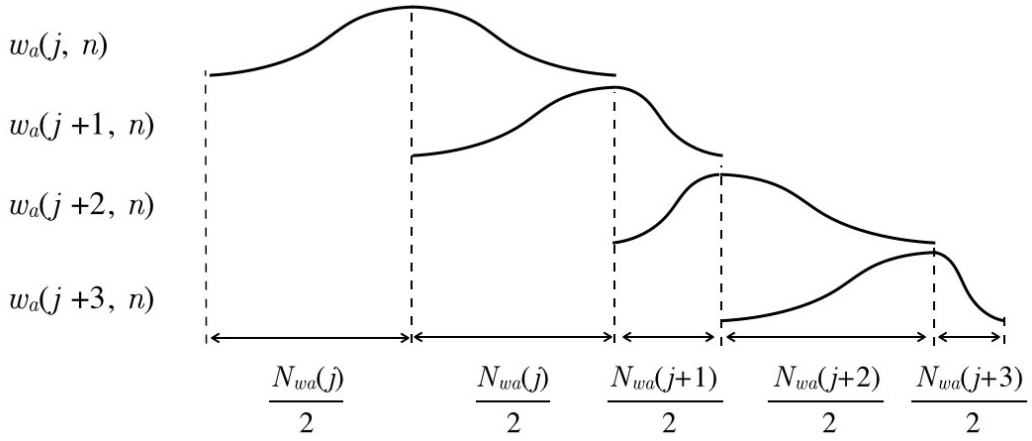


FIGURE 2.12 – *Création d'un assemblage de demi-fenêtres de Hann de tailles aléatoires à chaque nouvelle copie d'une portion d'un signal apériodique.*

et $min(N_{wa})$, $w_a(j,n)$ est une fenêtre spéciale, constituée d'un assemblage de deux moitiés de fenêtres de Hann (voir la FIGURE 2.12).

La position du marqueur de synthèse $t_s(j)$ est alors mise à jour de la même manière que sur la FIGURE 2.10 et selon l'équation (2.16) en prenant $P_s(j) = T_a(j)$, la durée définie par l'équation (2.18) :

$$T_a(j) = \frac{N_{wa}(j)}{\mu_a} \tag{2.18}$$

où $\mu_a$ est le facteur de recouvrement apériodique. Le respect de l'inégalité (2.19) permet d'éviter que le début d'une fenêtre $w_a(j)$ ne se trouve avant le début d'une fenêtre $w_a(j-1)$ :

$$\mu_a \leqslant 2 \times \frac{max(N_{wa})}{min(N_{wa})} \tag{2.19}$$

Il est préférable de garder $\mu_a$ aussi bas que possible et $min, max(N_{wa})$ aussi grand que possible afin d'économiser du temps de calcul : plus une période est grande, et plus le taux de recouvrement est faible, moins la quantité de calculs nécessaire sera importante. Après avoir testé plusieurs combinaisons, nous avons déterminé un compromis offrant une qualité convenable, défini dans l'équation (2.20) :

$$\begin{array}{rcll} \mu_a & = & 10 & \\ min(N_{wa}) & = & 3 & \text{ms} \\ max(N_{wa}) & = & 15 & \text{ms} \end{array} \tag{2.20}$$

L'utilisation d'une valeur inférieure pour $\mu_a$ crée un signal comportant un grain indésirable. Il est possible de choisir une valeur supérieure pour $max(N_{wa})$, mais plus sa valeur sera grande, plus la transition entre les parties non-voisées et voisées sera audible : une partie du signal voisé peut être incluse dans une fenêtre d'analyse.

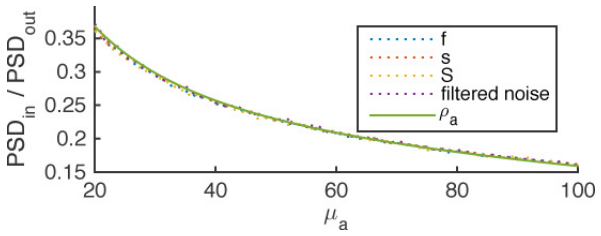


FIGURE 2.13 – *Rapport de puissance pour différentes valeurs de $\mu_a$, pour quatre signaux audio originaux, et $max(N_{wa}) = 15ms$. Ces courbes sont approximées par l'équation* (2.21).

Puisque les trames de synthèse consécutives sont recouvertes et ajoutées les unes aux autres, la puissance du signal de synthèse qui en résulte est augmentée par rapport à celle du signal original. Chaque trame de synthèse doit donc être multipliée par le facteur de puissance $\rho_a$, comme le montre l'équation (2.17).

L'équation (2.21) montre la relation qui lie $\rho_a$ à $\mu_a$ :

$$\rho_a = \frac{1}{3.9 \times e^{0.005\mu_a} - 2.9 \times e^{-0.03\mu_a}} \tag{2.21}$$

Pour la déterminer, nous avons comparé les Densités Spectrales de Puissance (PSD pour *Power Spectrum Density*) de quatre signaux non-voisés originaux ([f], [s], [ʃ] et un bruit blanc filtré passe-bande) avec celles de leurs signaux re-synthétisés. Pour chaque nouvelle période, les instants cibles $\tau(j)$ étaient tirés aléatoirement. La FIGURE 2.13 montre le rapport entre la somme des PSD originales et la somme des PSD de synthèse pour les différents signaux source, en faisant varier $\mu_a$ de 20 à 100 et en prenant $max(N_{wa}) = 15ms$. Presque aucune différence n'est visible entre les courbes des différents signaux, nous avons donc pu déterminer l'équation (2.21) grâce à la boîte à outil d'ajustement de courbes (*curve fitting toolkit*) de Matlab. Cette équation est représentée par la courbe continue sur la FIGURE 2.13.

Diminuer $min(N_{wa})$ n'a pas d'influence sur les constantes de l'équation, mais le fait d'augmenter $max(N_{wa})$ augmente le rapport de puissance. L'équation (2.21) n'est donc valide que pour $max(N_{wa}) = 15ms$. La détermination d'une équation qui prenne en compte toutes ces variables serait nécessaire pour finaliser cette méthode d'allongement infini des parties apériodiques.

### 2.3.3 Interpolation pour la concaténation

Lorsque le mode *Loop* est activé (voir section 3.3.8), une interpolation entre le dernier et le premier phonème de la boucle est effectuée, afin d'éviter toute discontinuité dans le signal de synthèse. Cette interpolation s'effectue selon l'équation (2.22) :

$$y(j,n) = x_w(\tau(j))(1 - \alpha_{loop}) + x_w(FCP_{start})\alpha_{loop} \tag{2.22}$$

avec $\alpha_{loop}$ le facteur d'interpolation de fin de boucle, défini par l'équation (2.23)

$$\alpha_{loop}(j) = \frac{\tau_{loop}(j) - p}{FCP_{end} - p} \quad pour \quad p \leq \tau_{loop}(j) \leq FCP_{end} \tag{2.23}$$

où $p$ représente le début du phonème final.

Une bonne pratique consistera alors à choisir des phonèmes final et initial identiques, ou peu éloignés. De plus, s'ils sont voisés, les marqueurs périodiques de ces deux phonèmes doivent être en phase pour assurer une qualité optimale. Cette méthode pourrait être utilisé pour des applications de synthèse par concaténation en temps-réel.

### 2.3.4 Mémoire tampon (buffer) circulaire

Puisque le signal de synthèse est créé en temps-réel, sa durée ne dépendra que du bon vouloir de l'utilisateur. Il est donc nécessaire d'allouer un espace de mémoire vive infini à ce futur signal de synthèse. Une solution, très fréquente en traitement temps-réel du signal, consiste à utiliser un espace mémoire (ou buffer) de façon circulaire, comme le montre la FIGURE 2.14. Ce buffer a une taille $N_{cb}$ finie, mais, étant circulaire, peut être considéré comme un buffer de taille infinie.

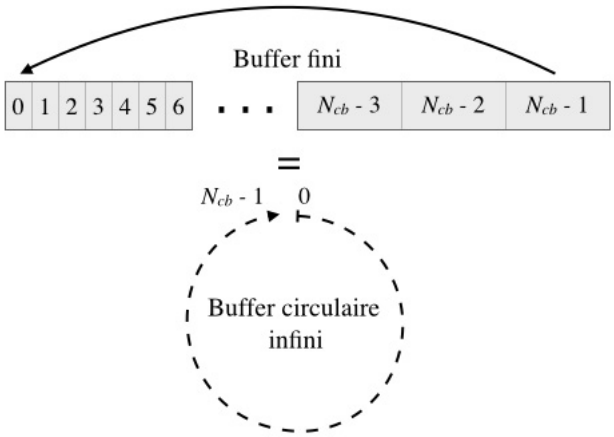


FIGURE 2.14 – *Buffer circulaire : espace mémoire fini de taille $N_{cb}$ transformé en espace mémoire infini.*

La FIGURE 2.15 détaille le fonctionnement d'un buffer circulaire $B$. Dans cette figure, nous considérons le cas simple où la période originale et la période de synthèse sont égales à la moitié de la taille de $B$ : $P_o(i) = P_s(j) = N_{cb}/2$. À chaque nouvelle période de synthèse, le contenu de $B$ est défini par $B(j, n)$, et ses positions de lecture et d'écriture sont définies par $r(j)$ et $wr(j)$.

L'initialisation du buffer circulaire consiste d'une part à le vider, et d'autre part à placer les positions d'écriture et de lecture, selon l'équation (2.24) :

$$\begin{aligned} wr(0) &= N_{cb}/2 \\ r(0) &= 0 \end{aligned} \tag{2.24}$$

Le fonctionnement du buffer circulaire s'effectue en deux phases : une phase d'écriture et une phase de lecture. Lors de la phase d'écriture, la trame de synthèse $y(j, n)$, définie par l'équation 2.12, est centrée autour de $wr(j)$, et ajoutée au contenu

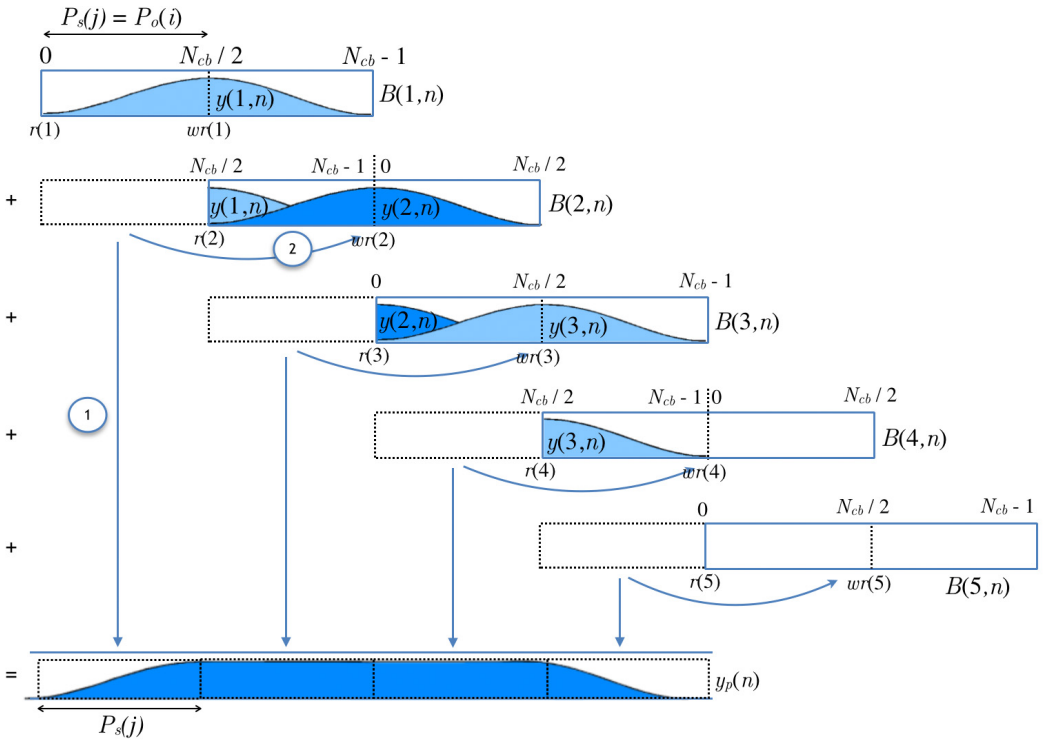


FIGURE 2.15 – *Exemple d'utilisation d'un buffer circulaire pour un signal de synthèse dont la période fondamentale de synthèse $P_s(j)$ serait égale à la période fondamentale originale $P_o(i)$ et à la moitié de la taille du buffer circulaire $N_{cb}$.*

de $B$ pour obtenir $B(j,n)$. La phase de lecture consiste ensuite à transférer une partie du contenu de $B(j,n)$ vers le vecteur de sortie $y(n)$. Dans la FIGURE 2.15, cette étape est représentée par les flèches descendantes, dont une est numérotée ①. Il s'agit d'abord d'écrire la première période de $B(j,n)$ dans le vecteur de sortie selon l'équation (2.25) :

$$y(n) = B(j,n) \quad pour \quad r(j) \leq n < r(j) + P_s(j) \tag{2.25}$$

Ensuite, le contenu qui vient d'être écrit dans le vecteur de sortie doit être effacé du buffer circulaire, afin de pouvoir réutiliser l'espace correspondant pour les écritures futures. En d'autres termes, le début du buffer est effacé pour en devenir la fin. Sur la FIGURE 2.15, cette opération est représentée par les flèches en arc de cercle dont une est numérotée ②. Les positions d'écriture et de lecture sont alors mises à jour selon l'équation (2.26) :

$$\begin{aligned} r(j) &= \sum_{k=0}^{j} P_s(k) \\ wr(j) &= r(j) + N_{cb}/2 \end{aligned} \tag{2.26}$$

Les limites qu'impose l'utilisation d'un buffer circulaire sont liées à sa taille.

D'abord, il ne faut pas choisir une taille trop grande, car, d'après l'équation (2.27), la durée de la latence $\lambda$ est proportionnelle à $N_{cb}$ :

$$\lambda = \frac{N_{cb}}{2} - P_o(i) \tag{2.27}$$

Cependant, en sachant que $N_w(i) \simeq 2P_o(i)$, si $P_o(i) > N_{cb}/2$ et $y(j,n) = x_w(i,n)$, alors $y(j,n)$ sera tronquée lors de la phase d'écriture du buffer circulaire, pour $n < r(j)$ et $n > r(j) + N_{cb}$. En d'autres termes, il ne faut pas que le buffer circulaire ait une taille inférieure à la plus grande période du signal d'entrée. En prenant en compte la modification de la longueur du conduit vocal (voir la section suivante), le buffer circulaire doit avoir une taille $N_{cb}$ qui satisfasse l'équation (2.28) :

$$N_{cb} \geq \frac{max(P_o)}{min(V_c)} \tag{2.28}$$

où $V_c$ correspond au facteur d'allongement du conduit vocal.

## 2.4 Longueur du conduit vocal

La fréquence centrale des formants d'un tube droit uniforme peut être déterminée par l'équation (2.29) :

$$F_\Phi = (2\Phi - 1) \times \frac{c}{4L} \tag{2.29}$$

avec $F_\Phi$ la fréquence centrale du $\Phi^e$ formant, $c$ la vitesse de déplacement du son en cm/s, et $L$ la longueur du tube en cm. Cette équation montre l'effet de la modification de la longueur du conduit vocal sur la fréquence des formants : un allongement du conduit vocal diminuera les fréquences des formants, et vice versa. Comme nous l'avons vu dans la section 2.2, FIGURE 2.5, le fait d'accélérer (compresser) ou de ralentir (étirer) un signal original aura pour effet d'étirer ou de compresser son enveloppe spectrale, respectivement. Afin de conserver cet effet sans affecter la fréquence fondamentale ou la durée d'un signal de synthèse, l'étirement ou la compression temporels peuvent être effectués directement sur les trames de synthèse, comme le montre la FIGURE 2.16. Ainsi, l'organisation temporelle des trames sur l'axe de synthèse est conservée, et seule l'enveloppe spectrale est modifiée. Un étirement d'une trame de synthèse permettra d'obtenir une compression de l'enveloppe spectrale, correspondant à un allongement du conduit vocal, et vice versa.

Le facteur d'allongement du conduit vocal à la $j^e$ période de synthèse est noté $V_c(j)$. Si $V_c(j) = 1$, la taille du conduit vocal n'est pas modifiée. Si $V_c(j) > 1$, le conduit vocal est allongé, et si $V_c(j) < 1$ il est raccourci. Un sous-échantillonnage permettra de compresser une trame, et un sur-échantillonnage de l'étirer. Cela consiste à recalculer la trame de synthèse $y(j,n)$ avec un nouvel indice temporel noté $k = mV_c(j)$, où $m$ est défini par l'équation (2.30) :

$$m = \left[ \sum_{-M(j)/2}^{k} \frac{1}{V_c(j)} \right] - M(j)/2 \tag{2.30}$$

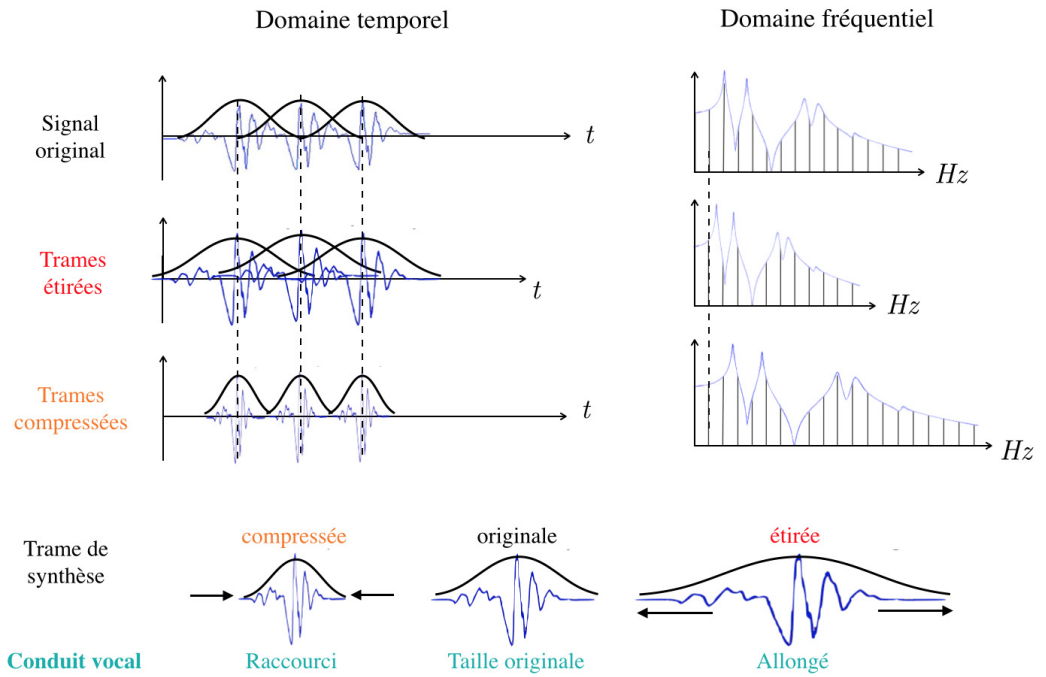


FIGURE 2.16 – *Modification de la taille du conduit vocal par compression / étirement des trames de synthèse.*

avec $M(j) = N_w(j) \times V_c(j)$ la taille de la trame modifiée. Cette trame modifiée, notée $y_c(j,k)$, est alors calculée selon l'équation (2.31), en prenant $n = \lfloor m \rfloor$ :

$$y_c(j,k) = y(j,n) + \big[y(j,n+1) - y(j,n)\big](m-n) \tag{2.31}$$

Cette équation indique que l'échantillon du signal modifié $y_c(j,k)$ sera le résultat de l'interpolation entre l'échantillon de la trame non modifiée $y(j,n)$ et la suivante $y(j,n+1)$. Nous obtenons alors une trame modifiée de taille $M$, dont le centre correspond à celui de $y(j,n)$ : $\frac{M}{2V_c} = \frac{N_w}{2}$.

## 2.5 Modification des paramètres de source : tension et effort

Nous venons de présenter l'algorithme VRT-PSOLA, permettant d'effectuer des modifications de durée, de hauteur et de taille du conduit vocal d'un signal original. Dans cette section, nous allons présenter les méthodes de modification de force et de tension vocales que nous avons mises en place.

### 2.5.1 Tension vocale

Une variation de tension vocale a pour effet d'agir sur la fréquence centrale du formant glottique. Cela influence la valeur de la différence entre l'amplitude

de l'harmonique fondamental et celles des harmoniques suivants dans le spectre de l'onde de débit glottique [Doval *et al.* 2006]. Une tension vocale élevée crée un timbre plus aigu, et une tension faible un timbre plus grave. Pour simuler une modification de la tension vocale du signal de synthèse, nous utilisons un modèle simplifié qui consiste à agir sur l'amplitude du son harmonique fondamental. Pour ce faire, nous procédons en deux étapes que nous présentons ici, et que nous détaillerons ensuite :

1. Le signal de synthèse issu de VRT-PSOLA $y_p(n)$ est filtré par un filtre passe-bande dont la fréquence de coupure est égale à la fréquence de synthèse et dont la largeur de bande est très étroite. Le signal filtré ainsi obtenu, noté $y_f(n)$, ne contient presque que l'harmonique fondamental.
2. Nous ajoutons à $y_p(n)$ le signal filtré $y_f(n)$ multiplié par un facteur $V_t$ compris entre -1 et 1 afin de modifier l'amplitude du fondamental de $y_p(n)$, et d'obtenir ainsi le signal modifié $y_t(n)$. Si $V_t = 1$, l'amplitude de l'harmonique fondamental sera augmentée (tension diminuée). Si $V_t = 0$, elle sera inchangée. Si $V_t = -1$, elle sera diminuée (tension augmentée).

**Étape 1 : filtrage du signal de synthèse**

La fonction de transfert $H_t(z)$ du filtre passe-bande que nous utilisons est définie par l'équation (2.32) :

$$H_t(z) = \frac{[1 - e^{(-2\pi T_e \Delta_f)}][1 - e^{(-2\pi T_e \Delta_f)} z^{-2}]}{1 - 2e^{(-2\pi T_e \Delta_f)} cos(f_s 2\pi T_e) z^{-1} + e^{(-4\pi T_e \Delta_f)} z^{-2}} \tag{2.32}$$

avec $T_e$ la période d'échantillonnage, $f_s$ la fréquence de synthèse, et $\Delta_f$ la largeur de bande, définie par l'équation (2.33) :

$$\Delta_f = \frac{f_s}{Q} \tag{2.33}$$

où $Q$ représente le facteur de qualité. Nous utilisons $Q = 10$ afin d'obtenir un gain de -17.6 dB au second harmonique. Le signal $y_f(n)$ est défini par l'équation (2.34) :

$$y_f(n) = \mathcal{Z}^{-1}[Y_p(z) H_t(z)] \tag{2.34}$$

**Étape 2 : calcul du signal modifié**

Le signal modifié $y_t(n)$ est alors calculé selon l'équation (2.35) :

$$y_t(n) = \Big[ y_p(n) - V_t \times y_f(n) \Big] \times \sigma \tag{2.35}$$

avec $\sigma$ le facteur de compensation énergétique défini par l'équation (2.36) :

$$\sigma = e^{ln(1.2) \times V_t} \tag{2.36}$$

où la valeur de 1.2 a été déterminée à l'oreille : les trois signaux (tension normale, maximale et minimale) ont ainsi une intensité sonore comparable.

Selon l'équation (2.35), si $V_t = 0$, le signal de synthèse ne sera pas modifié : la tension vocale restera identique à l'originale. Si $V_t = -1$, la tension vocale sera minimale, et si $V_t = 1$, elle sera maximale. Notez que Vokinesis permet d'effectuer une modification préalable de la tension vocale. Par exemple, si un signal original a été enregistré avec une tension vocale déjà élevée, ce pré-réglage permettra de la diminuer à l'avance pour permettre un contrôle qui soit adapté à cette voix particulière.

### 2.5.2 Effort vocal

Une variation de l'effort vocal a pour effet principal d'agir sur la pente spectrale de la source glottique. Pour simuler cet effet, nous utilisons le filtre proposé par [Doval *et al.* 2003, Feugère *et al.* 2017]. C'est un filtre passe-bas 2-pôles 2-zéros défini par l'équation (2.37), pour $V_e \leq 1$. L'effet d'augmentation de l'effort vocal peut alors être obtenu par l'inverse de ce filtre, dont la fonction de transfert est définie par la même équation, pour $V_e > 1$. Les équations (2.38) à (2.41) définissent les éléments de l'équation (2.37) :

$$H_e(z) = \begin{cases} H_{e1}(z) \times H_{e2}(z) & pour \quad V_e \leq 1 \\ \dfrac{1}{H_{e1}(z) \times H_{e2}(z)} & pour \quad V_e > 1 \end{cases} \tag{2.37}$$

$$H_{ek}(z) = \frac{1 - (\nu_k - \sqrt{\nu_k^2 - 1})}{1 - (\nu_k - \sqrt{\nu_k^2 - 1})z^{-1}} \tag{2.38}$$

$$\nu_k = 1 - \frac{cos(2\pi 3000Te) - 1}{10^{A_k/10} - 1} \tag{2.39}$$

$$A_1 = \begin{cases} 45(1 - V_e) \text{ dB} & pour \quad 0 \leq V_e \leq 1 \\ 45(V_e - 1) \text{ dB} & pour \quad 1 < V_e \leq 2 \end{cases} \tag{2.40}$$

$$A_2 = \begin{cases} 10(1 - V_e) \text{ dB} & pour \quad 0 \leq V_e \leq 1 \\ 10(V_e - 1) \text{ dB} & pour \quad 1 < V_e \leq 2 \end{cases} \tag{2.41}$$

avec $A_1 + A_2$ l'atténuation ($V_e \leq 1$) ou l'amplification ($V_e > 1$) à 3000Hz en dB. Le signal modifié en effort vocal $y_e(n)$ est obtenu selon l'équation (2.42) :

$$y(n) = y_e(n) = I_e \times \mathcal{Z}^{-1}[H_e(z)Y_t(z)] \tag{2.42}$$

avec $I_e = min(1, V_e^2)$. Ainsi, le volume du signal de synthèse variera de façon quadratique pour $0 \leq V_e \leq 1$ , et restera inchangé pour $V_e > 1$.

## 2.6 Conclusion

Ce chapitre nous a permis de présenter VoPTiQ, une méthode combinant l'algorithme VRT-PSOLA pour la modification temps-réel de la hauteur, de la du-

rée et de la taille du conduit vocal, et deux étapes de filtrage pour la modification de la tension et de l'effort vocal. VRT-PSOLA possède deux nouveautés par rapport à ses ancêtres RT-PSOLA [Le Beux *et al.* 2010] et TD-PSOLA [Moulines & Charpentier 1990, Moulines & Laroche 1995] : les zones non-voisées peuvent être maintenues pour une durée indéfinie sans bruit tonal indésirable grâce à l'utilisation de fenêtres d'analyse consécutives dont les durées sont tirées de façon aléatoire, et la longueur du conduit vocal peut être modifiée par étirement / compression des trames de synthèse.

Afin de compléter VoPTiQ, une étape de modification du niveau de souffle serait nécessaire. Ceci nécessiterait un traitement indépendant des composantes périodiques et apériodiques du signal de parole. Le récent vocoder WORLD [Morise *et al.* 2016], qui peut être utilisé en temps-réel, semble offrir une qualité encore meilleure que STRAIGHT. Il est basé sur une méthode précise d'estimation de la fréquence fondamentale et de l'enveloppe spectrale, avec en plus une extraction des paramètres apériodiques. Les futurs travaux devraient envisager de remplacer VRT-PSOLA par WORLD dans VoPTiQ. Il offrirait la possibilité d'ajouter un contrôle sur la quantité de souffle dans le signal de synthèse, et permettrait ainsi d'augmenter le pouvoir expressif de VoPTiQ (voir l'Annexe A). Par ailleurs, cet algorithme permettrait d'améliorer la qualité des fricatives voisées, grâce à un traitement indépendant de leurs composantes périodiques et apériodiques.

Chapitre 3

# Contrôle rythmique de la voix

## Sommaire



Nous avons vu dans le chapitre précédent que le contrôle performatif de l'articulation d'une voix de synthèse était une tâche au moins aussi complexe que pour

la voix réelle, qui demande donc un très long temps d'apprentissage : les mains et les pieds reproduisent difficilement les mouvements coordonnés des différents articulateurs actifs lors de production vocale. Si le contrôle complet de l'articulation est trop complexe à mettre en œuvre, il est nécessaire de définir de nouvelles méthodes de contrôle temporel de plus haut niveau.

Nos travaux étaient donc principalement centrés sur la recherche des méthodes les mieux adaptées au contrôle temps-réel du rythme d'un texte (parlé ou chanté), qui permette un contrôle précis et expressif, ainsi qu'une synthèse de bonne qualité. Les objectifs étaient les suivants :

- Définir l'unité temporelle suprasegmentale à contrôler pour offrir une liberté d'improvisation du rythme de textes parlés et chantés.
- Trouver les gestes les mieux adaptés au séquencement de cette unité temporelle. Ces gestes doivent permettre de conserver les paradigmes de contrôle chironomique de la hauteur avec une tablette graphique présentés dans le Chapitre 1.
- Ces gestes doivent permettre à des utilisateurs et utilisatrices de maîtriser le rythme de phrases parlées et chantées. Ils doivent donc permettre à plusieurs interprètes de produire des rythmes vocaux de façon synchronisée.

Nous présenterons dans la première section de ce chapitre les méthodes de contrôle temporel (non-rythmiques) qui étaient déjà existantes dans le système Calliphony, l'ancêtre de Vokinesis.

Le simple contrôle de la durée n'équivaut pas à un contrôle rythmique. Or, le contrôle du rythme peut permettre l'expression d'émotions (par exemple, la colère peut être exprimée par un rythme staccato [Kehrein 2002]), mais également la synchronisation dans un cadre musical. Dans les précédents systèmes, seule la pédale Vocaloid le permet (section 1.2.5), mais elle n'est destinée qu'au chant japonais (langue dont la structure syllabique est bien plus simple que celle du français, par exemple), et la guitare n'est sans doute pas l'interface la mieux adaptée au contrôle de la voix. La question qui se pose alors est la suivante : quels gestes faut-il utiliser pour le contrôle du rythme vocal (parole et chant), quel que soit le langage désiré ?

Contrairement à d'autres instruments acoustiques, pour lesquels les gestes de contrôle peuvent être facilement observables, et utilisés comme source d'inspiration pour la conception [Wanderley & Depalle 2004], les gestes de contrôle de la voix sont majoritairement internes (les gestes externes tels que ceux des mains ou du visage, par exemple, ne jouent pas un rôle sémantique dans la communication vocale, mais plutôt expressif). Une réflexion sur la nature du rythme vocal et sur la façon dont il est produit apparaît alors nécessaire, et nous verrons dans une seconde section pourquoi la syllabe est l'élément rythmique de base à contrôler. Nous tenterons donc de décrire son organisation temporelle en proposant une structure rythmique inter-linguistique du séquencement syllabique. Nous présenterons dans une troisième section les stratégies de contrôle gestuel que nous avons mises en place pour le séquencement du rythme vocal. La quatrième section fournit une explication de la façon dont les signaux originaux doivent être enregistrés et étiquetés pour assurer

une qualité de synthèse et de contrôle optimale. La dernière section présentera les évaluations objective et subjective de nos méthodes de contrôle du rythme vocal.

## 3.1 Calliphony : contrôle de la durée

Avant de parler de contrôle rythmique, nous souhaiterions présenter les méthodes de contrôle de la durée de Calliphony [Le Beux *et al.* 2007, Le Beux 2009] dont Vokinesis a hérité. Son fonctionnement est représenté FIGURE 3.1. Il permet, grâce à l'algorithme RT-PSOLA, de modifier en temps-réel la durée et la hauteur d'un signal de parole pré-enregistré à partir des coordonnées $(x, y)$ d'un stylet sur une tablette graphique. Les données de contrôle et les données audio ainsi créées peuvent être sauvexgardées sur le disque dur. Il possède deux modes de contrôle temporel non rythmiques, l'un permettant de contrôler la vitesse de lecture (mode *Speed*), et l'autre permettant de contrôler directement la position de l'instant cible (mode *Scrub*). Nous allons les présenter ci-dessous.

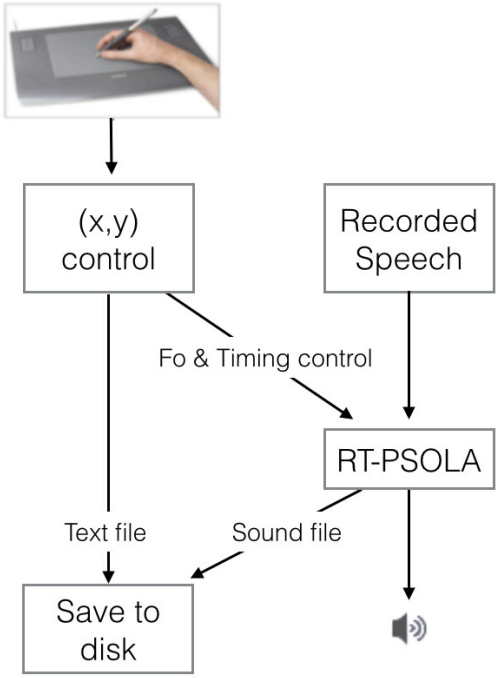


FIGURE 3.1 – *Fonctionnement du système Calliphony. Image adaptée de* [Le Beux *et al.* 2007].

### 3.1.1 Contrôle direct de l'instant cible : mode *Scrub*

Le mode *Scrub* permet de cibler directement l'instant du signal original à resynthétiser (nommé *instant cible*, et noté $\tau$), grâce à un contrôleur continu tel que l'axe vertical de la tablette (contrôle lié : la mélodie et la durée sont contrôlées par

le même membre) ou un potentiomètre (contrôle libre : les contrôles mélodique et temporel sont indépendants). La valeur du contrôleur, comprise entre 0 et 1, est multipliée par la durée du signal original. Le synthétiseur reçoit donc une valeur temporelle correspondant à l'instant qui doit être re-synthétisé dans le signal original, comme l'indique l'équation (3.1) :

$$\tau = p_c \times N \tag{3.1}$$

avec $N$ la taille du signal original (en échantillons) et $0 \leq p_c \leq 1$ la position du contrôleur. Cette méthode permet d'effectuer des variations très fines et de décomposer le signal original avec précision. Cependant, elle ne permet pas un contrôle précis du rythme syllabique, car l'amplitude des mouvements à effectuer varie selon la durée de la syllabe, et selon la durée du signal original. Un nouvel apprentissage serait donc nécessaire pour chaque syllabe de chaque signal original.

### 3.1.2 Contrôle de la vitesse de lecture : mode *Speed*

le mode *Speed* (ou vitesse) permet de contrôler la vitesse de lecture du signal original avec un contrôleur continu. Nous verrons dans le CHAPITRE 5, section 5.5.2.2, que la plage de contrôle de la vitesse peut être réglée avec deux paramètres nommés $v_{min}$ (vitesse minimale) et $v_{max}$ (vitesse maximale). L'équation (3.2) définit les valeurs extrêmes qui peuvent être assignées à ces paramètres :

$$\begin{array}{rcl} v_{min} & \leq & min(v_{max}, 1) \\ v_{max} & \geq & max(v_{min}, 0.1) \end{array} \tag{3.2}$$

La vitesse de lecture $v_s$ est calculée selon la position $p_c$ du contrôleur continu, et différemment selon le signe de $v_{min}$, comme le montre l'équation (3.3) (Calliphony ne permettait pas d'assigner une valeur négative à $v_{min}$, ceci constitue donc une évolution propre à Vokinesis) :

$$v_s = \begin{cases} (2p_c - 1)v_{max} & pour \quad 0.5 < p_c \leq 1 \quad et \quad v_{min} \leq 0 \\ -(2p_c - 1)v_{min} & pour \quad 0 \leq p_c \leq 0.5 \quad et \quad v_{min} \leq 0 \\ e^{ln(v_{max}) \times (2p_c - 1)} & pour \quad 0.5 < p_c \leq 1 \quad et \quad v_{min} > 0 \\ e^{ln(\frac{1}{v_{min}}) \times (2p_c - 1)} & pour \quad 0 \leq p_c \leq 0.5 \quad et \quad v_{min} > 0 \end{cases} \tag{3.3}$$

La FIGURE 3.2 schématise cette équation dans le cas où $p_c$ correspond à la position du stylet sur l'axe vertical de la tablette. Quand $v_{min} \leq 0$ (à gauche de la figure), la vitesse de lecture $v_s$ est positive dans la partie supérieure de la tablette (la lecture se fait en avant), et négative dans la partie inférieure (lecture arrière). Au centre de la tablette, la vitesse est nulle, et le signal de synthèse reste bloqué au dernier instant cible calculé. Si $v_{min} = 0$, alors la partie inférieure de la tablette est inactive : $v_s$ restera à 0. Dans le cas où $v_{min} > 0$ (à droite de la figure), la vitesse de lecture $v_s > 1$ dans la partie supérieure de la tablette (la lecture est accélérée), et $v_s < 1$ dans la partie inférieure (la lecture est ralentie). Au centre de la tablette,

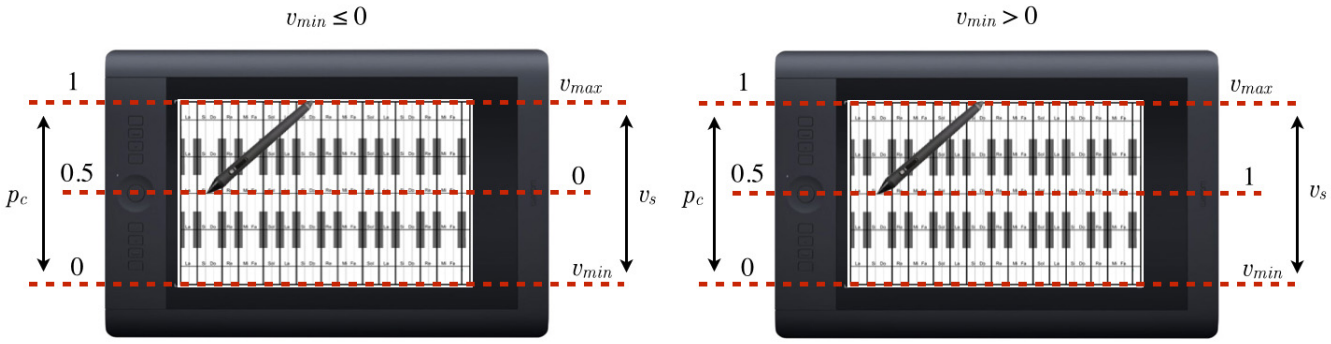


FIGURE 3.2 – *Contrôle de la vitesse de lecture sur l'axe y de la tablette graphique.*

la vitesse est égale à la vitesse original. Si $v_{min} = 1$, alors la partie inférieure de la tablette est inactive : $v_s$ restera à 1. L'instant cible est ainsi calculé selon l'équation (3.4) :

$$\tau(n) = \sum_{m=0}^{n} v_s(m) \tag{3.4}$$

avec $n$ l'échantillon présent du signal de synthèse. Un appui sur le bouton supérieur du stylet permet de réinitialiser $n$ à zéro.

## 3.2 Le rythme vocal

Dans un contexte musical ou dans le cas de la parole, nous pouvons dire que le rythme correspond à des séquences structurées dans le temps d'événements perceptivement saillants. Dans cette section, nous allons tenter de définir une structure générale du rythme vocal qui puisse être appliquée à n'importe quelle langue.

### 3.2.1 Hiérarchie temporelle de la production et de la perception de la voix

Pour déterminer comment contrôler le rythme de la production vocale, il est tout d'abord nécessaire d'en comprendre l'organisation temporelle. La FIGURE 3.3 est une représentation schématique de la structure du mot anglais « *tomato* ». Au niveau segmental sont produits les phonèmes (voyelles et consonnes), qui constituent les plus petites unités acoustiques de la parole. Au niveau suprasegmental se trouvent l'intonation et les accents toniques. La syllabe est à l'interface entre les deux niveaux : elle régit l'organisation rythmique suprasegmentale (répartition des accents toniques et des pics d'intonation), et l'organisation des phonèmes au niveau segmental. Nous détaillons ci-dessous chaque élément de la figure, en commençant par ceux qui varient le plus lentement.

L'intonation représente la forme générale que suivra la fréquence fondamentale ($f_0$ ou hauteur), qui correspond à la fréquence de vibration des plis vocaux. Elle

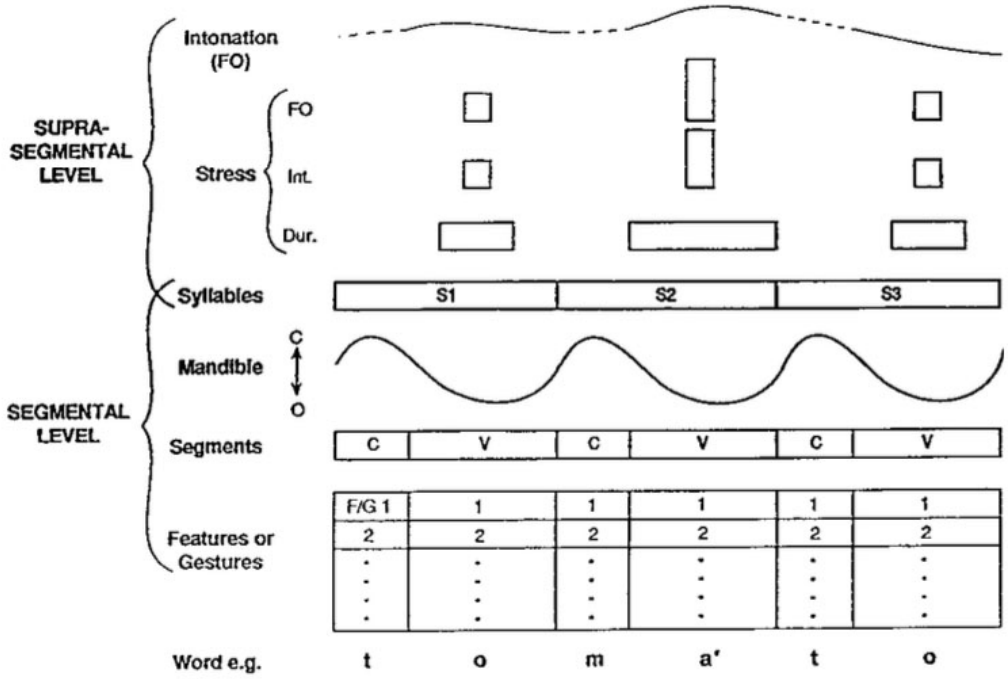


FIGURE 3.3 – *Représentation schématique de l'organisation temporelle de la parole pour le mot anglais « tomato ». Image issue de* [MacNeilage 1998].

permet à un locuteur de signifier l'intention d'une phrase. Par exemple, en français, un courbe de hauteur montante permet d'exprimer une question, une courbe descendante une affirmation, etc...

L'accent tonique (ou *stress*) permet d'accentuer une syllabe, qui sera alors perçue comme plus fortes que celles qui l'entourent. La façon dont une syllabe est accentuée dépend de la langue. En anglais, les syllabes sont accentuées par un pic dans la courbe de $f_0$ et d'intensité, et un allongement de leur durée. En français, l'allongement des syllabes accentuées est moins important, mais le pic de hauteur est bien existant [Wagner 2008]. Dans les règles poétiques des langues dont l'organisation rythmique est dite *accentuelle* (*stress-timed languages*) telles que l'anglais, un vers est composé d'un certain nombre (fixé) de pieds, et un pied est composé d'un certain nombre (variable) de syllabes, dont une est accentuée. C'est donc le nombre d'accents toniques qui régit la composition d'un vers. Dans les langues dont l'organisation rythmique est dite *syllabiques* (*syllable-timed languages*) telles que le français, c'est le nombre de syllabes qui est fixé, et donc qui régit la composition d'un vers. Dans les langues dites *moriques* (*mora-timed languages*), la structure d'un vers est définie par le nombre de mores, une unité temporelle qui définit la longueur d'une syllabe (plus de détails dans la section 3.2.2).

La syllabe est la plus petite unité d'organisation rythmique suprasegmentale. Sur la figure, chaque syllabe est produite par un cycle d'ouverture/fermeture de la

mandibule, les ouvertures correspondant à des phases vocaliques, et les fermetures à des phases consonantiques. Au niveau segmental, chaque phonème est produit par des gestes articulatoires (mouvements de la langue, des lèvres, de la mandibule...) et par des états actifs ou non de certaines caractéristiques (état de voisement, nasalisation). De plus amples détails seront fournis dans les sections 3.2.2 et 3.2.4.

Dans le cadre de la perception du temps, les recherches en neurosciences et en psychoacoustique ont permis de mettre en évidence l'existence de fenêtres d'intégration temporelle de différentes tailles, qui traiteraient les informations temporelles de différentes manières. [Wagner 2008] propose de les assimiler à différents niveaux temporels de la parole, ce que nous résumons ci-dessous :

| | |
|---|---|
| *Fenêtre syllabique* : | Les signaux audio reçus par le cerveau seraient segmentés en groupes de 150 - 200ms. Cette durée est tout à fait comparable à la durée inter-linguistique moyenne de la syllabe. |
| *Présent psychologique* ou *Fenêtre du pied* : | Correspond approximativement à la durée de deux syllabes / d'un pied (400-600ms). Ce serait la plus petite fenêtre permettant de détecter un motif rythmique. |
| *Fenêtre de la phrase* : | Environ 3 secondes d'après [Pöppel 1994]. Elle permettrait d'intégrer différents motifs temporels (pieds) formant ainsi des motifs rythmiques plus longs (vers, phrases). |

La fenêtre du présent psychologique a la plus courte durée d'intégration permettant de détecter des motifs rythmiques. La syllabe est donc la plus petite unité permettant de créer des motifs rythmiques. En effet, d'après [Wagner 2008], la syllabe est le niveau d'organisation prosodique qui crée l'impression de battements.

### 3.2.2 Composition de la syllabe

Nous venons de décrire la hiérarchie temporelle de la production et de la perception de la voix, et avons vu que la syllabe en constituait l'élément rythmique porteur. C'est donc le séquencement des syllabes qui permet de contrôler la base rythmique de la voix. Pour comprendre comment séquencer le rythme syllabique, nous aurons besoin de connaître les règles phonologiques de la composition d'une syllabe.

Pour les phonologues, la syllabe est définie par trois phases : l'attaque (consonnes), le noyau (voyelle), et la coda (consonnes), le noyau et la coda constituant la rime (voir la FIGURE 3.4). L'attaque peut être composée d'une ou plusieurs consonnes (comme pour les mots *ta* [**t**a], *toit* [**tw**a], *trois* [**tʁw**a]), ou alors être inexistante (comme pour le mot *a*). Le noyau est toujours présent. Il est la plupart du temps composé d'une voyelle. Il existe cependant des exceptions. En anglais, par exemple, certaines consonnes (nasales ou latérales) [Wells 1965] peuvent prendre la place du

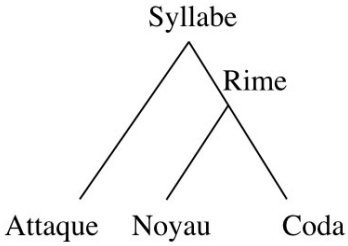


FIGURE 3.4 – *Structure phonologique d'une syllabe.*

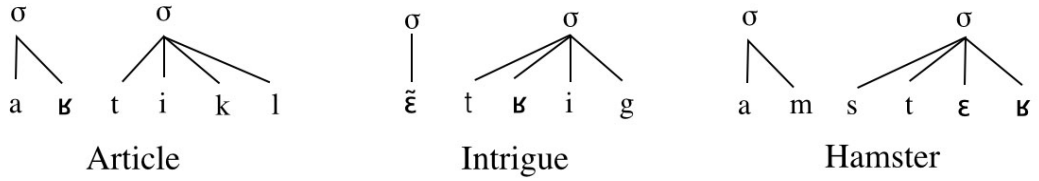


FIGURE 3.5 – *Syllabification des mots « article », « intrigue » et « hamster ».*

noyau syllabique. Elles seront alors nommées *consonnes syllabiques*. Le noyau de la deuxième syllabe du mot « *little* » [lit̩l̩] s'écrit alors [l̩], celui du mot « *seven* » [sɛvn̩] s'écrit [n̩]. La coda est également constituée d'une ou plusieurs consonnes (comme pour les mots *Yves* [iv], *ivre* [ivʁ], *fichtre* [fiʃtʁ]), et peut aussi être inexistante (comme pour le mot *vie*).

Dans une production vocale, les syllabes s'enchaînent. L'appartenance d'une consonne à une syllabe ou à une autre dépend des règles phonotactiques de la langue. Le principe est le suivant : si les règles phonotactiques d'une langue interdisent l'association de deux consonnes consécutives en début de mot, alors ces consonnes doivent appartenir à des syllabes différentes. Sinon, toutes les consonnes d'un groupe consonantique qui respecte les règles phonotactiques sont regroupées avec le noyau suivant [Katamba 1989]. Considérons par exemple les mots « article », « intrigue » et « hamster », dont la syllabification est présentée FIGURE 3.5. En français, le mot imaginaire « rticle » est inconcevable : les règles phonotactiques interdisent l'association [ʁt] en début de mot. Le [ʁ] et le [t] appartiennent donc à des syllabes différentes pour le mot « article ». Par contre, le mot « trigue » serait tout à fait concevable, l'ensemble [tʁ] est donc associé à la même syllabe pour « intrigue ». Enfin, le mot « mster » est inconcevable, mais le mot « ster » l'est. L'ensemble [st] constitue donc l'attaque de la seconde syllabe du mot « hamster », et le [m] constitue la coda de la première. Ces règles ne s'appliqueraient pas de la même manière pour des langues donc les règles phonotactiques diffèrent. [Katamba 1989] donne comme exemple le Swahili, l'une des nombreuses langues africaines qui autorisent des séquences initiales telles que [nd] ou [ŋg]. Les mots [ndugu] ou [ŋguruwe] qui signifient « frère » et « cochon » possèdent des séquences initiales inconcevables en français. Les règles de syllabification des deux langues sont donc différentes.

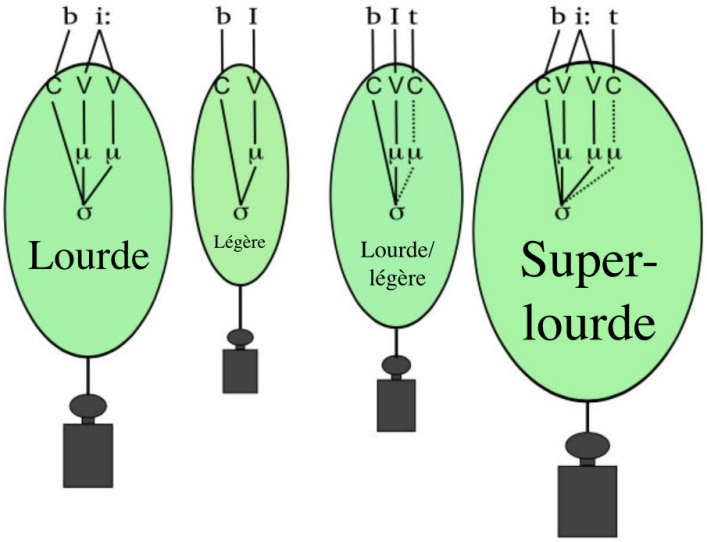


FIGURE 3.6 – *Poids d'une syllabe en fonction du nombre de mores qu'elle contient. Figure issue de* [Wagner 2008].

La more est une unité temporelle qui permet de définir la longueur d'une syllabe : une syllabe peut être composée d'une ou plusieurs mores. le nombre de mores comprises dans la syllabe définit son *poids*. Une syllabe dite *lourde* comportera plus de mores qu'une syllabe dite *légère* (voir la FIGURE 3.6). La répartition du poids des syllabes régit l'organisation syllabique (répartition des syllabes accentuées) de manière différente selon le langage. Cependant, la détermination du poids est universelle : une more donne une syllabe légère, deux mores une syllabe lourde, et trois mores une syllabe très lourde. le nombre de mores ne correspond pas nécessairement au nombre de phonèmes : certaines voyelles peuvent être allongées et comporter alors deux mores.

### 3.2.3 Centre perceptif (*p-center*) et rythme syllabique

Nous savons que la syllabe est l'unité de base pour la perception du rythme vocal, ou encore pour la production de motifs rythmiques. En musique, chaque syllabe est associée à un événement rythmique sur une partition. Considérons par exemple les partitions de la FIGURE 3.7, qui décrivent des chansons populaires française et anglaise (les langues syllabique et accentuelle prototypiques). Chaque syllabe est associée à un événement rythmique, ou battement, dont la longueur peut varier. Au niveau de l'avant dernière mesure de la chanson française, une seule syllabe est prononcée alors que deux événement rythmiques sont indiqués : il s'agit ici d'une variation rythmique mélodique. La syllabe est alors maintenue (la durée de la voyelle est allongée, telle une more supplémentaire) tandis qu'un changement de note est produit.

L'événement rythmique perçu d'une syllabe se nomme le *centre perceptif*,

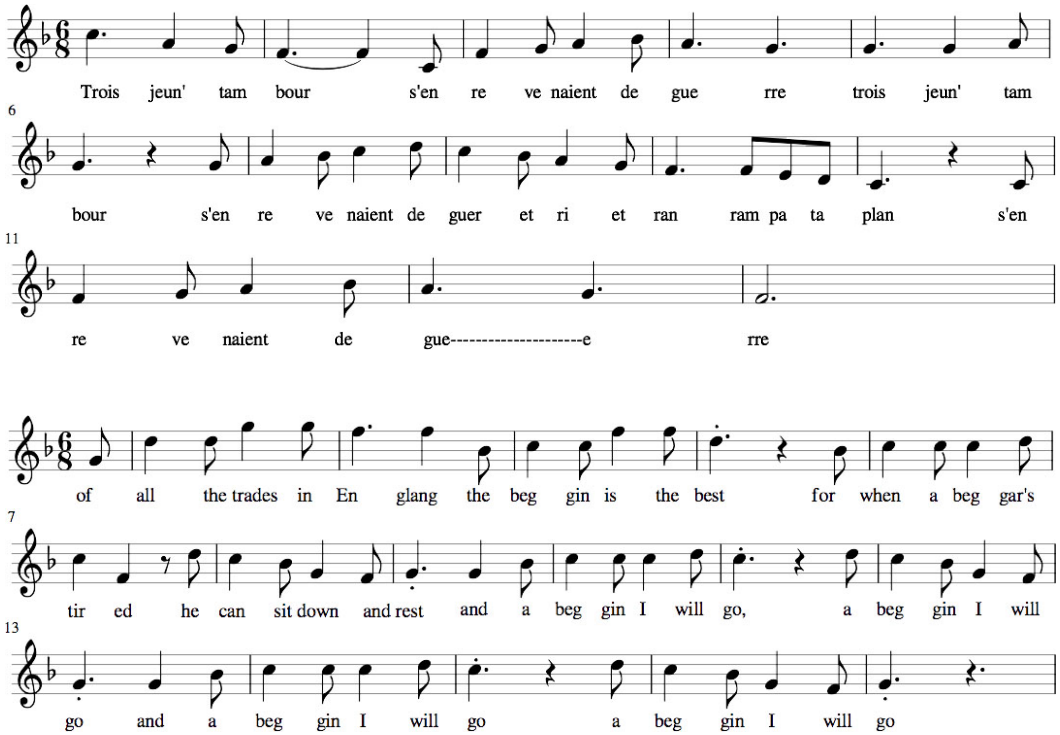


FIGURE 3.7 – *Partitions du chant populaire français « Trois jeunes tambours » (en haut) et du chant populaire anglais « A begging I will go » (en bas). Chaque syllabe est assignée à un événement rythmique.*

ou encore le *p-center* (pour *perceptual center*). Il y a un large consensus pour définir son emplacement comme étant proche du début du noyau syllabique [Pompino-Marschall 1989, Scott 1993, Barbosa *et al.* 2005, Mairano 2011], qu'il soit composé d'une voyelle ou d'une consonne syllabique [Wagner 2008]. Sur une partition, le début de chaque nouvel élément rythmique indique l'emplacement d'un p-center ; cela implique qu'un nouveau noyau syllabique doit démarrer au même instant qu'une nouvelle note. La durée qui sépare deux noyaux syllabiques est donc définie par la durée qui sépare deux événements rythmiques. L'allongement ou le raccourcissement d'une syllabe consiste principalement à modifier la durée de son noyau, les durées des consonnes étant très peu modifiées [Bartkova & Sorin 1987, Kuwabara 1996]. Nous pouvons alors supposer que la durée d'un événement rythmique d'une partition indique la durée du noyau vocalique, dont la fin serait déterminée par anticipation de la durée nécessaire pour produire les consonnes qui le séparent du noyau suivant. Cependant, de nombreuses productions chantées contredisent cette supposition. La chanson « Amsterdam » de Jacques Brel en est un parfait exemple. Nous invitons lectrices et lecteurs à écouter le début de cette chanson (par exemple avec cet enregistrement [1]), et à porter une attention particulière sur les durées des différentes consonnes. Ces allongements temporels sont d'une impor-

1. https://youtu.be/2U06PicY2C4

tance capitale pour la perception des intentions émotionnelles émises par le chanteur (parmi d'autres éléments tels que le vibrato, les variations d'effort vocal, pour ne parler que du chant). En parole, le contrôle du rythme syllabique permet également de véhiculer des intentions expressives. Par exemple, la colère peut être exprimée par des rythmes staccato (raccourcissement des noyaux) [Kehrein 2002]. Cependant, à notre connaissance, aucune règle d'écriture musicale ne permet de réguler les durées des phonèmes. Un événement rythmique d'une partition définit donc le démarrage d'un noyau et la durée qui le sépare du noyau suivant, mais ne donne aucune indication sur la durée propre du noyau : l'interprète régulera les durées segmentales selon ses intentions expressives.

### 3.2.4 Phonologie articulatoire

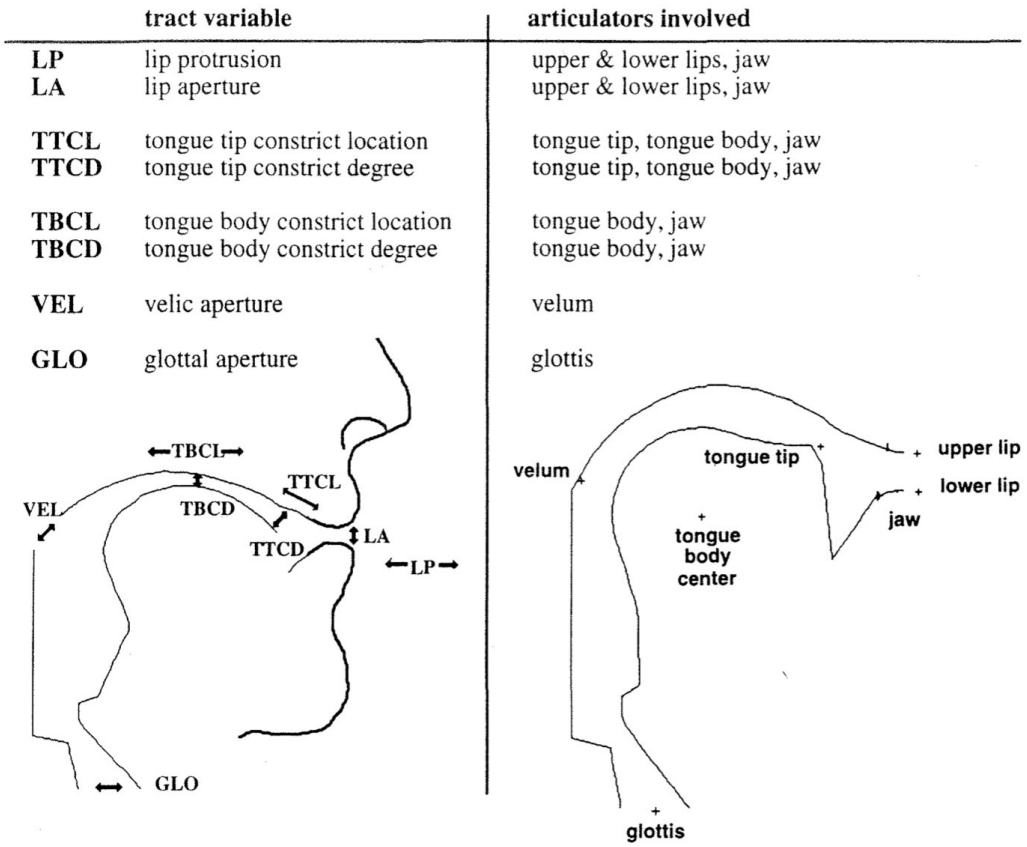


FIGURE 3.8 – *Les articulateurs (à droite) et leurs variables de conduit correspondantes (à gauche)* [Browman & Goldstein 1990a].

TABLEAU 3.1 – *Cibles gestuelles et symboles correspondants* [Browman & Goldstein 1990b]. *Les cibles gestuelles de la glotte et du vélum ne sont pas représentés ici.*

| Symbole | Signification | Variable de conduit |
|---|---|---|
| i | geste palatal (étroit) | TBCD, TBCL |
| a | geste pharyngé (étroit) | TBCD, TBCL |
| β | geste de fermeture bilabiale | LA, LP |
| τ | geste de fermeture alvéolaire | TTCD, TTCL |
| σ | geste de quasi-fermeture alvéolaire (permet la friction) | TTCD, TTCL |
| λ | geste de fermeture alvéolaire latérale | TTCD, TTCL |
| κ | geste de fermeture vélaire | TBCD, TBCL |

Nous avons vu qu'une syllabe comportait un événement rythmique perceptivement saillant (le *p-center*) ainsi que des éléments sémantiques (voyelles et consonnes) et expressifs (durée des segments). Nous allons à présent nous pencher du côté de la phonologie articulatoire, et plus particulièrement sur le modèle de [Browman & Goldstein 1992], qui définit l'organisation des gestes articulatoires effectués pour la production de la parole.

Le conduit vocal est composé de plusieurs articulateurs plus ou moins dépendants les uns des autres. Par exemple, la lèvre inférieure et la langue sont attachées à la mandibule et leur position dépend donc du degré d'ouverture de la mâchoire. Dans leur modèle, [Browman & Goldstein 1990a] proposent de simplifier l'étude complexe des mouvements respectifs des articulateurs et de leurs interactions, grâce à l'utilisation de *variables de conduit* (ou *tract variable*). La FIGURE 3.8 présente les articulateurs du conduit vocal (à droite) et leurs variables de conduit correspondantes (à gauche). Un ensemble d'articulateurs est associé à un couple de variables de conduit. Par exemple, les lèvres et la mâchoire sont associées au couple L (Lèvre), qui indique le degré de protrusion (LP) et d'ouverture (LA) des lèvres. L'apex de la langue, son corps et la mâchoire sont associés au couple TT (*Tongue Tip*, Apex de la langue), qui représentent le lieu et le degré de constriction (CL et CD) de l'apex. Le couple TB (*Tongue Body*, corps de la langue) est associé à la mâchoire et au corps de la langue. Enfin, le vélum (VEL) et la glotte (GLO) étant indépendants des autres articulateurs considérés, seule leur ouverture est prise en compte.

À chaque phonème peut être apparenté une ou plusieurs cibles gestuelles des variables de conduit. Par exemple, chaque phonème de la phrase anglaise « *piece plots* » peut être apparenté à une cible gestuelle du TABLEAU 3.1. La transcription phonétique s'écrira [pisplats], et la transcription gestuelle correspondante {βiσβλaτσ}. Chaque cible gestuelle implique un mouvement des variables de conduit auxquelles elle est associée. Bien que les phonèmes soient considérés comme des entités distinctes, les gestes articulatoires de phonèmes consécutifs se recouvrent. Plus l'énoncé sera rapide, plus les gestes consécutifs seront recouverts.

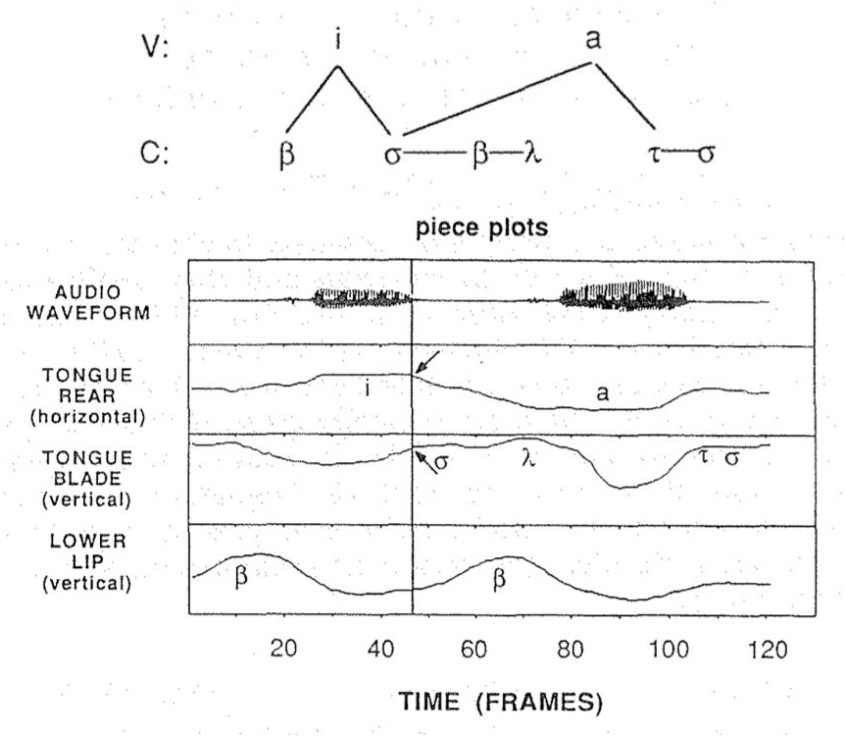


FIGURE 3.9 – *Cibles et mouvement articulatoires pour la phrase « piece plots ». En haut : cibles vocaliques* (V) *et cibles consonantiques* (C). *En bas : mouvements articulatoires correspondants mesurés par rayon-x pour le corps de la langue* (Tongue Rear)*, son apex* (Tongue Blade) *et la lèvre inférieure* (Lower Lip). [Browman & Goldstein 1990b].

[Browman & Goldstein 1990b] ont alors pu définir des règles de suppression de certains phonèmes en parole spontanée lorsque les gestes articulatoires sont trop recouverts (par exemple la suppression du [t] dans l'énoncé « *must be* », prononcé [mʌsbi] au lieu de [mʌstbi]). Lorsque l'on ordonne à une personne d'articuler, on lui demande en réalité de prendre le temps d'atteindre toutes les cibles articulatoires, et donc de faire attention à produire entièrement tous les gestes articulatoires de la phrase qu'elle souhaite prononcer.

Deux grands types de cibles articulatoires peuvent être distingués : les cibles vocaliques (V), indiquées par des lettres latines, et les cibles consonantiques (C), indiquées par des lettres grecques (FIGURE 3.9, en haut). Le séquencement des syllabes consiste à lier des cibles vocaliques par des mouvements vocaliques, qui peuvent être couplés à des mouvements consonantiques. La durée prise pour passer d'une cible vocalique à l'autre dépendra du nombre de cibles consonantiques qui les séparent. Par exemple, si aucune consonne ne sépare deux voyelles, alors le geste vocalique qui les liera sera très court. En haut de la FIGURE 3.9, chaque ligne représente un geste articulatoire. Un geste vocalique démarre en même temps que le premier geste consonantique d'un groupe de consonnes. Ceci est représenté par la

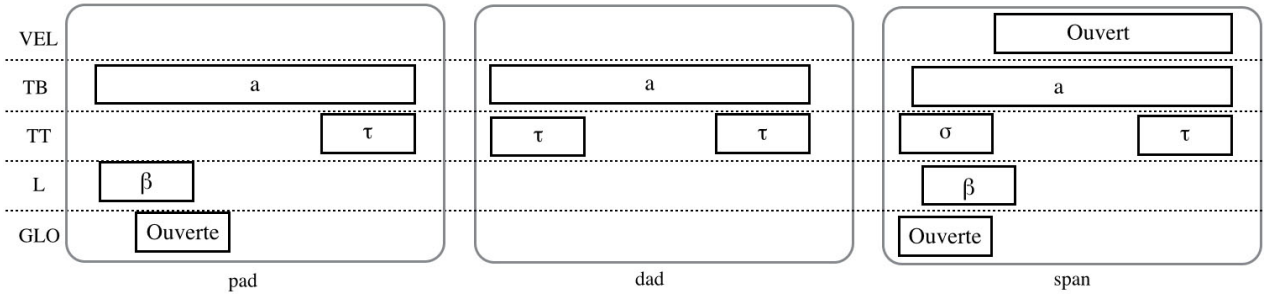


FIGURE 3.10 – *Partitions gestuelles. Le vélum et la glotte prennent des états binaires. Si le vélum est ouvert, un son nasal est produit. Si la glotte est ouverte, un son non-voisé (sans vibration des plis vocaux) est produit* [Browman & Goldstein 1992].

ligne reliant {i} à {β} et par celle reliant {a} à {σ} : à chaque fois qu'un mouvement vocalique démarre, les mouvements consonantiques qui y seront couplés démarrent également, dans l'ordre d'apparition des cibles consonantiques correspondantes. Sur la partie inférieure de la figure, le démarrage du mouvement vocalique reliant les cibles {i} et {a} (déplacement du corps de la langue) est indiqué par une flèche. Ce mouvement démarre donc au même instant que le mouvement consonantique {σ}, dont le démarrage est également indiqué par une flèche. Ce mouvement vocalique est ensuite couplé avec les mouvements consonantiques suivants {β} et {λ}. En haut de la figure, les traits qui relient les cibles consonantiques représentent le recouvrement des mouvements consonantiques correspondants. Ceux-ci peuvent être observés dans le bas de la figure. Par exemple, le second mouvement {β} effectué par la lèvre inférieure recouvre la fin du mouvement {σ} et le début du mouvement {λ} effectués par l'apex de la langue. Notez que les règles de synchronisation des mouvements consonantiques aux mouvements vocaliques ne tiennent pas compte des règles d'appartenance d'une consonne à une syllabe ou à une autre : même si les règles phonotactiques impliquent qu'une certaine consonne appartienne à la syllabe précédente, son mouvement consonantique démarrera au même instant que le mouvement vocalique visant le noyau syllabique suivant.

Des exemples de partitions gestuelles sont présentés dans la FIGURE 3.10, pour les mots monosyllabiques « *pad* », « *dad* » et « *span* ». Dans ces exemples, chaque rectangle des lignes TB, TT et L représente un mouvement du couple de variables de conduit correspondant, visant la cible articulatoire indiquée. Les rectangles des lignes VEL et GLO représentent quant à eux des états binaires de l'ouverture du vélum et de la glotte : lorsqu'un son nasal est produit, le vélum est ouvert ; lorsqu'un son non-voisé est produit, la glotte est ouverte.

En résumé, les mouvements vocaliques sont des gestes articulatoires qui permettent de passer d'une cible vocalique à la suivante. Ces liaisons vocaliques sont souvent couplées avec des mouvements plus fins et plus rapides qui visent des cibles consonantiques. Le rythme des gestes articulatoires est donc régi par le séquencement de **cibles vocaliques**, liées par des mouvements coarticulatoires plus ou moins

complexes que nous nommerons **liaisons vocaliques**.

### 3.2.5 Cadre syllabique : La théorie Frame/Content

Les sections précédentes nous ont permis de définir l'organisation rythmique de la parole et l'organisation segmentale de la syllabe d'un point de vue phonologique. Dans cette section, nous allons présenter la théorie *Frame / Content* (F/C, ou *cadre / contenu*) de l'évolution de la parole [MacNeilage 1998], qui permet de décrire l'organisation cognitive du séquencement phonémique et syllabique.

D'après cette théorie, l'apparition de la parole chez l'humain serait le résultat d'une réutilisation des mouvements cycliques primitifs de l'ingestion tels que la mastication. L'étude du cerveau humain et sa comparaison avec celui d'autres mammifères ont permis de montrer que les développements du cerveau pour la production de la parole ont principalement pris place aux alentours et à l'intérieur de la zone de Broca, zone cérébrale connue pour jouer un rôle primaire dans le contrôle des mouvements d'ingestion chez les mammifères. Selon l'auteur, les étapes d'apprentissage chez l'enfant (l'ontogenèse) offriraient une bonne représentation des étapes évolutives à l'échelle de l'humanité (phylogenèse). Or, il semblerait que le babillage (productions syllabiques des nouveaux nés effectuées par des mouvements d'ouverture/fermeture de la mâchoire), soit une étape universelle de l'apprentissage de la parole [Canault 2007]. L'apprentissage de l'articulation commencerait donc par l'apprentissage du séquencement du *cadre syllabique* (*frame*), créé par cycles d'ouverture / fermeture de la mandibule (le babillage), auquel le *contenu sémantique* (*content*), c'est à dire les variations fines des autres articulateurs qui permettront un jour de produire de la parole qui ait du sens, est ajouté petit à petit, au fur et à mesure que l'enfant apprend à préciser les mouvements des muscles correspondants.

Chez l'adulte, l'organisation cognitive de la production syllabique conserverait cette dichotomie. En effet, l'observation d'erreurs de prononciation ont permis de montrer qu'elles concernent la plupart du temps des échanges d'un même segment syllabique : l'attaque, le noyau ou la coda d'une syllabe seront respectivement échangés avec l'attaque, le noyau ou la coda d'une autre syllabe. Cognitivement, ce type d'erreurs prendrait place à l'interface du système lexical et du système moteur, responsable de l'organisation rythmique générale des syllabes (alternance de positions ouvert/fermé du conduit vocal), mais également de la modulation de ces cycles par la production de consonnes et de voyelles durant les phases de fermeture et d'ouverture. Après avoir observé les effets des lésions de certaines zones du cerveau, [MacNeilage 1998] a pu supposer que la planification du cadre syllabique et du contenu sémantique prendraient place dans deux zones distinctes : le système pré-moteur médian pour la cadre, et le système pré-moteur latéral pour le contenu.

### 3.2.6 Détermination d'une structure rythmique inter-linguistique du séquencement syllabique

La FIGURE 3.11 réunit les concepts que nous avons présentés dans les sections précédentes, tout en considérant des exceptions de structuration syllabique. Ainsi, la phrase « *little man* » [lit̩mæn] comporte un [l̩] syllabique, et le mot « *menstrual* » [mɛnstrʊəl] comporte un groupe consonantique multiple [nstr] qui appartient à deux syllabes différentes selon les règles phonotactiques de l'anglais, et une syllabe sans coda suivie d'une syllabe sans attaque : [strʊ] et [əl]. Nous allons nous servir de ces deux exemples pour tenter de définir une structure rythmique générale du séquencement syllabique.

D'un point de vue phonologique, nous savons qu'une syllabe est toujours composée d'un noyau syllabique, comportant généralement une voyelle, et parfois une consonne. Cependant, les règles de structuration syllabique ne permettent pas de définir une structure rythmique universelle, car elles dépendent des règles phonotactiques de la langue, et car leur structure à trois phases est très variable : seul le noyau est toujours présent. D'un point de vue articulatoire, nous avons vu dans la section 3.2.4 que le séquencement syllabique était effectué par l'enchaînement de cibles vocaliques (noyaux syllabiques) et de liaisons vocaliques (transitions articulatoires entre deux cibles vocaliques). Cependant, ces appellations ne sont pas généralisables au cas des consonnes syllabiques. Nous proposons donc de renommer le terme « cible vocalique » par « **noyau rythmique** », et le terme « liaison vocalique » par « **liaison rythmique** ». Nous préférons « rythmique » à « syllabique » car, comme nous le

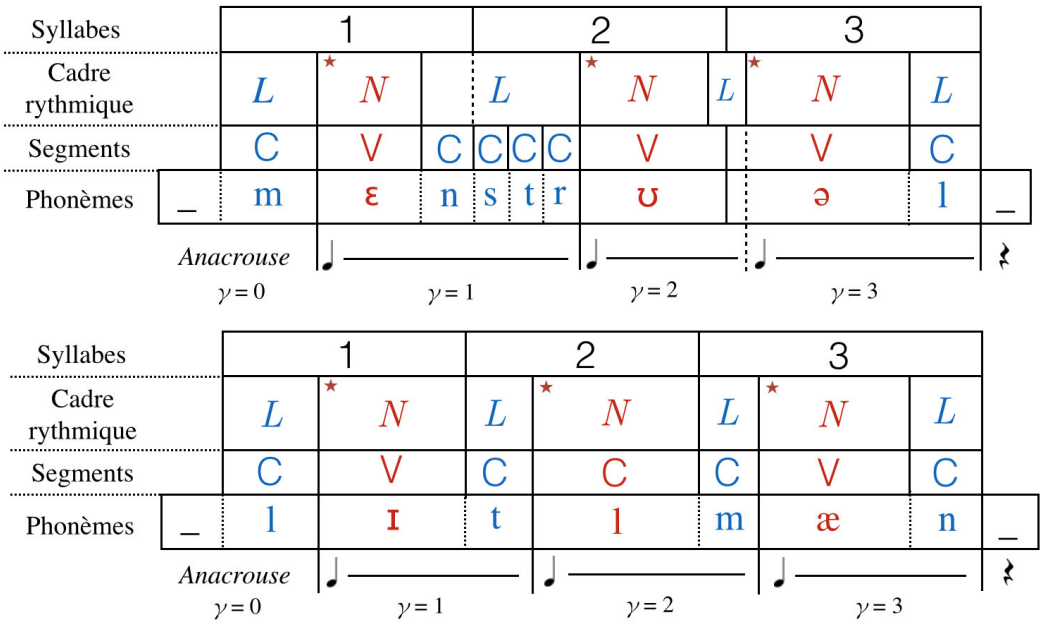


FIGURE 3.11 – *Organisation structurelle du séquencement syllabique. L : liaison rythmique ; N : Noyau rythmique ; Étoile : p-center ; C : consonne ; V : voyelles*

verrons plus bas, les règles d'appartenance de certaines consonnes à un groupe rythmique ne respectent pas forcément les règles de syllabification. Ainsi, sans remettre en cause la théorie F/C, nous pensons que l'appellation *cadre syllabique* n'est plus appropriée. Nous utiliserons donc dorénavant le terme « **cadre rythmique** », que nous considérerons composé d'une alternance de noyaux et de liaisons rythmiques.

Nous pouvons à présent définir une structure générale du cadre rythmique. Sur la FIGURE 3.11, nous avons représenté le cadre rythmique par des alternances de liaisons et de noyaux rythmiques. Les p-centers, représentés par des étoiles, se trouvent au début des noyaux rythmiques. Nous pouvons donc déjà observer un motif *Liaison / Noyau* dont chaque répétition contient un p-center. [Barbosa & Bailly 1994] proposent une unité alternative à la syllabe : le ***groupe inter-p-center*** (ou IPCG pour *inter-perceptual-center group*). Cette unité est mieux adaptée aux règles d'écriture musicale du rythme que la syllabe : un événement rythmique d'une partition définit la durée qui sépare un p-center du suivant. Dorénavant, nous utiliserons le terme ***groupe rythmique*** pour désigner l'ensemble des phonèmes dont la durée correspond à un événement rythmique d'une partition. Dans la figure, les groupes rythmiques auxquels appartiennent les phonèmes sont mis en évidence et numérotés par les indices $\gamma$. Pour les deux exemples, le groupe $\gamma = 0$ (silence - liaison) se trouve avant le premier p-center. Il peut donc être considéré comme une anacrouse : en musique, une anacrouse correspond à une note ou un groupe de notes précédant le premier temps fort de la première mesure ; un exemple d'anacrouse peut être observé dans la partition du chant anglais présentée FIGURE 3.7. Tous les groupes rythmiques suivants sont donc composés du motif *Noyau / Liaison*. Ainsi, l'analogie avec l'écriture musicale est respectée : un événement rythmique d'une partition définit la durée d'un groupe rythmique.

## 3.3 Séquencement du cadre rythmique

Dans la section précédente, nous nous sommes appuyés sur des notions phonologiques et perceptives de la voix pour montrer que son cadre rythmique était constitué d'une alternance de noyaux et de liaisons rythmiques, portant le contenu sémantique (voyelles et consonnes). Dans cette section, nous allons présenter les différentes méthodes de contrôle du cadre rythmique que nous avons pu explorer.

Replaçons donc dans le contexte du contrôle performatif de la synthèse vocale, et plus particulièrement du logiciel Vokinesis. Les interprètes utilisant Vokinesis modifient des signaux de voix pré-enregistrés. Ils peuvent contrôler des paramètres de hauteur et de qualité vocale (voir le CHAPITRE 4), mais également des paramètres temporels. Lors du contrôle temporel, le paramètre qui est sous contrôle se nomme *l'instant cible*, et se note $\tau(t)$. Il représente l'instant $\tau$ du signal original que les interprètes souhaitent synthétiser à un instant $t$ de la vie réelle. La façon dont l'instant cible est contrôlé dépend du mode de contrôle temporel sélectionné (pour plus de détails sur la configuration du logiciel, se référer au CHAPITRE 5 ). En plus des modes de contrôle temporel *Speed* et *Scrub* que comportait Calliphony,

et que nous avons présentés dans la section 3.1, Vokinesis possède deux modes de contrôle rythmique : le mode *Tap* (contrôle binaire du cadre rythmique par des mouvements percussifs) et le mode *Fader* (contrôle continu des liaisons rythmiques par des interfaces continues de type potentiomètre).

Dans la procédure de conception d'un système interactif, il est préférable de concevoir d'abord des méthodes d'interactions, puis de tester ensuite différentes interfaces qui pourront permettre de les mettre en œuvre, plutôt que de se restreindre à une interface et de tenter de trouver une méthode d'interaction qui convienne [Beaudouin-Lafon 2004]. C'est effectivement la procédure que nous avons suivie. Ainsi, nous présenterons le fonctionnement de chaque méthode d'interaction que nous avons mises en place avant de présenter les différentes interfaces de contrôle testées.

De récentes recherches ont permis de montrer que l'audition dominait largement la vision pour la perception des durées [Ortega *et al.* 2014]. Nous privilégierons donc les méthodes de contrôle rythmique ne nécessitant pas l'usage de la vision, qui joue un rôle très important pour le contrôlé mélodique dans le cas de l'utilisation d'une tablette graphique [Perrotin & D'alessandro 2016a].

### 3.3.1 Frame Control Points (FCP)

Le contrôle du cadre rythmique nécessite la mise en place de points d'ancrages temporels sur le signal original : un point d'ancrage par noyau rythmique, et un point d'ancrage par liaison rythmique (donc deux points d'ancrage par groupe rythmique). Ces points d'ancrage permettront au logiciel de déterminer la position de l'instant cible selon les mouvements de contrôle rythmique des interprètes. Nous les nommerons *Points de Contrôle du Cadre*, ou FCP pour *Frame Control Points*. Un groupe rythmique contiendra toujours deux FCP : le point nucléique, nommé $P_n(\gamma)$, et le point de liaison, nommé $P_l(\gamma)$, avec $\gamma$ le numéro de groupe rythmique. En haut de la Figure 3.12, nous pouvons voir le spectrogramme de la phrase « *my name is* » [majnejmɪz] étiquetée phonétiquement. Les lignes bleues verticales représentent ses FCP. Les $P_n$ sont toujours situés au sein des noyaux rythmiques, et les $P_l$ au sein des liaisons. Le premier $P_l$ correspondant à l'anacrouse, il sera toujours numéroté $P_l(0)$. Leur emplacement précis est très important pour assurer la précision du contrôle rythmique. Pour le moment, nous pouvons simplement dire qu'un $P_n$ se trouve au centre d'un noyau rythmique, et qu'un $P_l$ se trouve au centre de la dernière consonne d'une liaison rythmique. Cependant, les règles précises de placement des FCP peuvent différer légèrement selon le mode de contrôle rythmique. Nous les définirons donc plus en détails dans la section 3.4, après avoir défini le fonctionnement de chacun des modes de contrôle rythmique.

### 3.3.2 Contrôle binaire du cadre rythmique : mode *Tap*

En mode *Tap*, le séquencement du cadre rythmique s'effectue par des mouvements percussifs : l'appui et le relâchement d'une touche de contrôle permet de

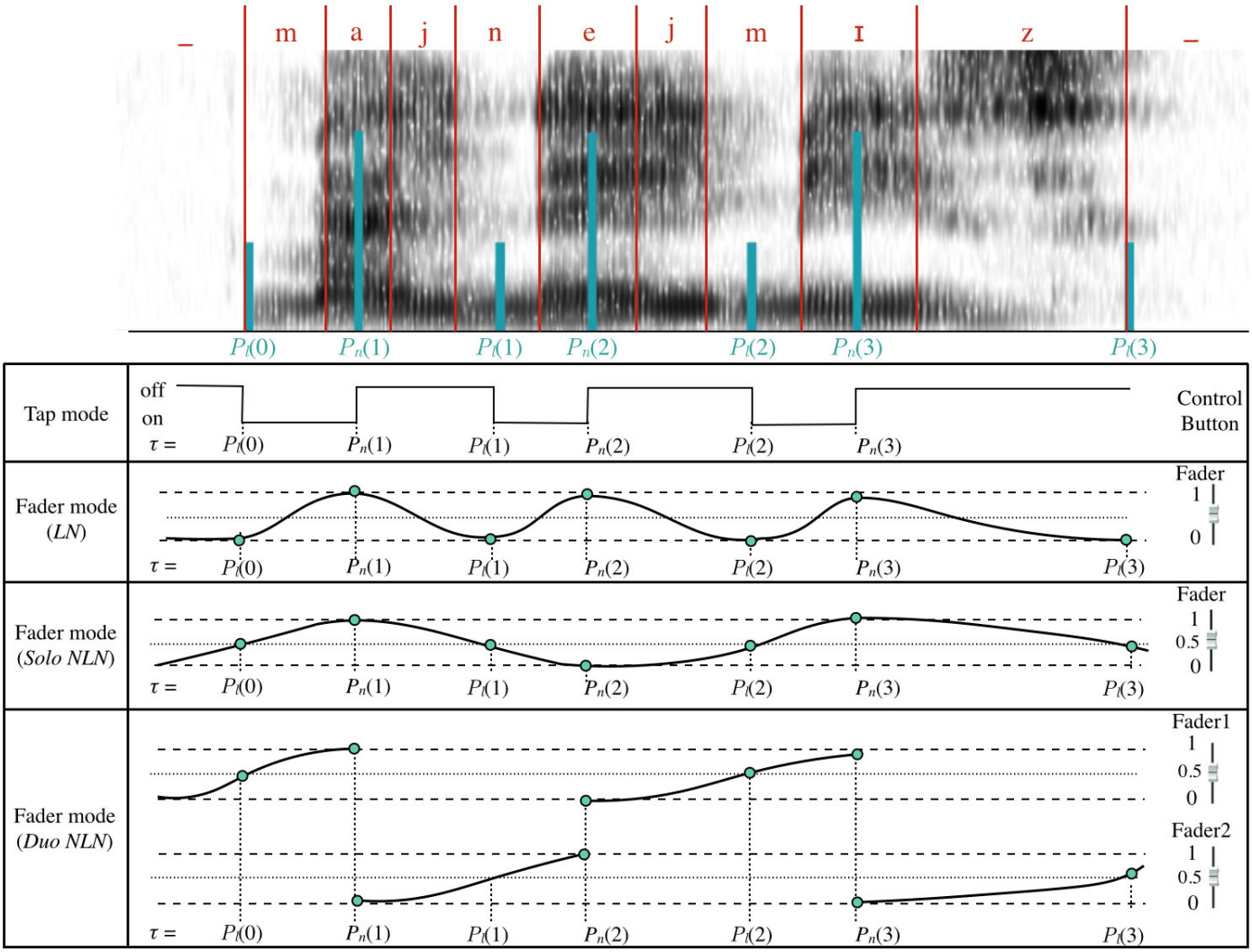


FIGURE 3.12 – *Contrôle du cadre rythmique de la phrase* « my name is ». *En haut : étiquetage des phonèmes, spectrogramme et points de contrôle du cadre. En bas : modes de contrôle rythmique — de haut en bas : mode Tap, mode Fader Liaison-Noyau (LN), modes Fader Noyau-Liaison-Noyau (NLN) solo et duo.*

déclencher les noyaux et les liaisons rythmiques, et la durée de maintien d'un état de la touche permet de définir la durée de chacune de ces phases. La ligne *Tap mode* de la FIGURE 3.12 représente l'état appuyé (*on*) ou relâché (*off*) de la touche de contrôle. Un appui déclenche la transition entre un point de liaison et un point nucléique. La durée du maintien de la touche définit la durée du noyau. Le relâchement déclenche une transition entre ce point nucléique et le point de liaison suivant. Au début, le premier $P_l$ est sélectionné ($\tau = P_l(0)$). Lors d'un appui sur la touche de contrôle, l'instant cible évolue du premier $P_l$ au premier $P_n$. Une fois ce $P_n$ atteint, ($\tau = P_n(1)$), le système synthétisera la période originale correspondante pour une durée indéfinie. Cette période originale sera donc dupliquée en plusieurs périodes de synthèse, aussi longtemps que la touche de contrôle sera maintenue enfoncée. Une fois la touche relâchée, l'instant cible évolue du $P_n$ actuel ($P_n(1)$) jusqu'au $P_l$ suivant ($P_l(1)$). Un nouvel appui sur la touche fera évoluer l'instant cible du $P_l$ actuel ($P_l(1)$) jusqu'au $P_n$ suivant ($P_n(2)$), et ainsi de suite jusqu'à ce que la fin du signal original soit atteinte. Une syllabe est donc entièrement prononcée par une séquence *relâchement-appui-relâchement*, et un groupe rythmique par une séquence *appui-relâchement*. La vitesse de lecture des transitions est prédéfinie, et peut être réglée différemment pour les noyaux, les liaisons et les silences (voir le CHAPITRE 5, section 5.5.2.2), mais ne peut pas être contrôlée en temps-réel. Nous utilisons généralement une vitesse de 3 fois l'originale pour les noyaux et 1.5 fois pour les liaisons. Ces vitesses de lecture permettent la production d'un rythme au tempo plus rapide que celui de l'enregistrement original. Puisque le noyau constitue la partie la plus stable d'un groupe rythmique, il peut être lu rapidement sans causer de problème d'intelligibilité. Cependant, les liaisons étant constituées d'éléments courts et rapidement variables, il est important de ne pas trop les accélérer pour conserver l'intelligibilité et le naturel du signal original [Lindblom & Studdert-Kennedy 1967].

Ce mode de contrôle binaire est une bonne analogie de la représentation biphasique du contrôle rythmique de la théorie F/C : les états *appuyé* et *relâché* de la touche de contrôle peuvent être apparentés aux états *ouvert* et *fermé* du conduit vocal.

### 3.3.3 Interfaces pour le contrôle binaire du cadre rythmique

L'interface qui nous semble la plus évidente pour le contrôle binaire du cadre rythmique est la barre espace de l'ordinateur : les interprètes contrôlent la mélodie avec le stylet en utilisant leur main préférée, et le cadre rythmique avec la main restante sur la barre espace (un exemple de cette configuration peut être visualisé dans la vidéo `Ex07` à partir de 1m31s). Cependant, le contrôle du cadre rythmique peut être effectué par tout contrôleur pouvant fournir un état binaire. Par exemple, s'il est assigné à l'état de toucher du stylet sur la tablette graphique, alors le rythme et la hauteur pourront être contrôlés par un seul membre. Cette configuration peut être très utile pour le contrôle de langues tonales, comme nous le verrons dans le CHAPITRE 7. Une autre méthode consiste à contrôler le rythme et la mélodie avec un clavier MIDI (un exemple de cette configuration peut être visualisé dans la vidéo

`Ex08` ; pour le contrôle mélodique, se référer au CHAPITRE 4) : un appui sur une touche déclenche un noyau, qui sera maintenu tant que toutes les touches jouées ne seront pas relâchées. Une autre configuration consiste à assigner ce contrôle binaire à l'intensité d'un signal audio d'entrée, capté par exemple par un microphone. Un seuil d'intensité réglable (voir la section 5.5.3.4) définira alors une frontière binaire permettant de simuler l'état d'une touche de contrôle : si l'intensité du signal d'entrée dépasse ce seuil, cela sera considéré comme un état *appuyé* de la touche de contrôle. Si elle est inférieure à ce seuil, cela correspondra à un état relâché de la touche. Une transition faible - forte intensité (une attaque) déclenchera un noyau, et une transition forte - faible intensité (un relâchement) déclenchera une liaison.

### 3.3.4 Contrôle continu des liaisons rythmiques : mode *Fader*

Nous avons vu que le mode *Tap* était une bonne analogie de la représentation biphasique du contrôle rythmique de la théorie F/C. Cependant, la nature binaire de ce mode de contrôle ne permet pas de maîtriser la vitesse de lecture des transitions, et ne permet donc pas un contrôle fin des durées segmentales des liaisons rythmiques. Cela constitue donc une limite du point de vue de l'expressivité (voir la section 3.2.3). Le mode *Fader* fait l'usage d'interfaces continues de type potentiomètres, qui permettent un contrôle précis des durées des liaisons rythmiques.

La FIGURE 3.12 montre les trois modes de contrôle continu disponibles. Les positions extrêmes des potentiomètres sont représentées par les valeurs 0 et 1. Déplacer un potentiomètre d'une position extrême à l'autre fait avancer l'instant cible $\tau$ dans le signal original.

Dans le premier mode, *liaison-noyau* (*LN*), déplacer le potentiomètre de 0 à 1 effectue une transition LN et le déplacer de 1 à 0 effectue une transition NL[2]. Ce mode de contrôle continu, tout comme le mode *Tap*, peut être apparenté à la représentation biphasique du contrôle rythmique de la théorie F/C de [MacNeilage 1998] (section 3.2.5) : la position 0 correspond à la phase fermée du conduit vocal, et la position 1 à sa phase ouverte. Bien que cette analogie soit intéressante, le mode *LN* sera ignoré par la suite, car les mouvements impliqués ne permettent pas un contrôle assez rapide du cadre rythmique.

Dans le second mode, *Solo noyau-liaison-noyau* (*NLN*), un déplacement du potentiomètre de la position 0 à 1 ou de la position 1 à 0 effectue une transition NLN. Ainsi, les noyaux rythmiques correspondent directement aux cibles du potentiomètre (ses positions extrêmes), et les liaisons rythmique sont directement assignées aux mouvements de transition entre les deux cibles du potentiomètre. Ce mode de contrôle continu peut être apparenté à la théorie F/C, mais également aux notions de Phonologie Articulatoire (PA) de [Browman & Goldstein 1992] (section 3.2.4) :

2. Si le potentiomètre utilisé est une pédale d'expression, le contrôle ressemble alors exactement à celui de la fameuse pédale *wah-wah*. Cette pédale pour guitare applique un filtre formantique variable au signal d'entrée. La position « talon » de la pédale permet de prononcer un [w] (la consonne) et la position « pointe » un [a] (la voyelle). La transition entre les deux positions extrêmes de la pédale permet d'effectuer la liaison entre ces deux phonèmes.

les phases ouvertes du conduit vocal (F/C) ou les cibles vocaliques (PA) sont visées par les positions extrêmes du potentiomètre, et les phases de fermeture/ouverture (F/C) ou les liaisons vocaliques (PA) sont effectuées par les transitions d'une position extrême à l'autre.

Dans le troisième mode, *Duo NLN*, deux potentiomètres sont utilisés au lieu d'un seul (par exemple deux pédales d'expression, comme nous le verrons dans la section 3.3.7). Dans ce mode, les pédales sont actives uniquement lors de leur déplacement positif ($0 \rightarrow 1$ : talon $\rightarrow$ pointe), et doivent être utilisés de façon successive : sur la FIGURE 3.12, un déplacement du *Fader 1* de 0 à 1 effectue la première transition NLN, puis un déplacement du *Fader 2* de 0 à 1 effectué la seconde transition NLN, et ainsi de suite.

### 3.3.5 Traitement du geste de contrôle continu

Tout comme un signal audio, les signaux émis par des potentiomètres peuvent être traités et transformés. On parle alors de *traitement* ou d'*édition du geste* [Ramstein 1991]. Nous avons mis en place deux méthodes importantes permettant d'optimiser le contrôle continu du rythme vocal. La première consiste à permettre le retour du potentiomètre avant qu'il ait atteint la fin de sa course, afin d'éviter aux interprètes d'avoir impérativement à atteindre une position extrême pour passer à la syllabe suivante. La seconde consiste à ralentir le signal de contrôle dans les phases importantes pour l'intelligibilité (liaisons rythmiques), et à l'accélérer dans les phases plus stables (noyaux rythmiques).

#### 3.3.5.1 Retour

Afin de permettre une liberté de mouvement optimale, nous avons fait en sorte que les interprètes n'aient en réalité pas besoin d'atteindre une position extrême du potentiomètre pour déclencher la transition suivante, comme l'illustre la FIGURE 3.13.

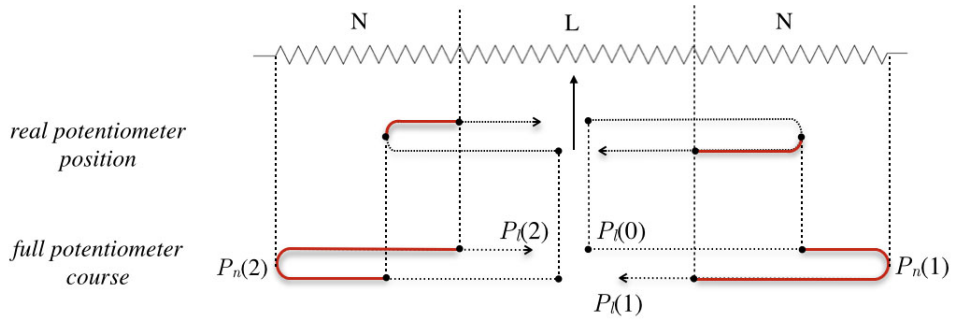


FIGURE 3.13 – *Un changement de direction d'un potentiomètre avant l'atteinte de sa position extrême simule la course complète qui aurait dû être effectuée.*

Dès qu'un mouvement de retour est détecté dans le deuxième moitié de la course du potentiomètre, la transition suivante est déclenchée. Afin d'empêcher un saut

brutal de la position de l'instant cible dans le signal original, toute la course qu'aurait parcouru le potentiomètre en atteignant sa position extrême (tracé rouge en bas de la FIGURE 3.13) est simulée à partir du moment où le mouvement de retour est détecté, jusqu'à la fin du noyau rythmique (courbe rouge au centre de la figure). Une course plus courte du potentiomètre balayera donc la même portion de signal qu'une course complète. Ceci fonctionne de la même manière pour le mode *Duo NLN*, à la seule différence qu'un retour dans les autres modes correspond à un démarrage de déplacement positif du prochain potentiomètre actif dans celui-ci.

### 3.3.5.2 Correction temporelle des liaisons rythmiques

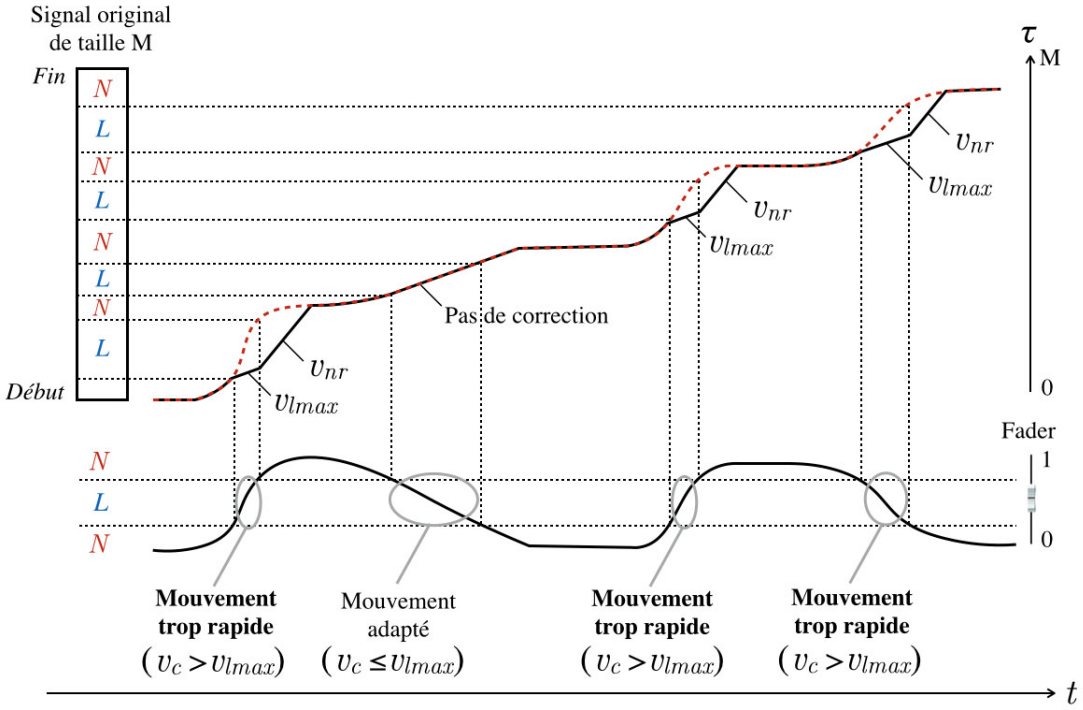


FIGURE 3.14 – *Correction temporelle des liaisons lors de déplacements trop rapides du potentiomètre en mode Fader NLN. (En haut) ligne rouge pointillée : évolution de l'instant cible $\tau$ sans correction temporelle ; ligne noire continue : évolution de $\tau$ avec correction temporelle. (En bas) évolution du potentiomètre de contrôle.*

Les liaisons sont très importantes à préserver pour une bonne intelligibilité [Lindblom & Studdert-Kennedy 1967]. Si elles sont lues de façon trop rapide, les informations acoustiques (et donc sémantiques) qu'elles comportent seront perdues. Or, lors de la production de rythmes rapides, la transition d'une position extrême à l'autre d'un potentiomètre peut être quasiment instantanée. Puisque la position de l'instant cible est directement liée à celle du potentiomètre, une correction temporelle doit y être appliquée : la vitesse d'évolution de l'instant cible doit être limitée pendant les liaisons rythmiques par une *vitesse maximale des liaisons* (notée

$v_{lmax}$), et une *vitesse de rattrapage des noyaux* (notée $v_{nr}$) doit accélérer l'évolution de l'instant cible pour rattraper la position du potentiomètre pendant les phases nucléiques. Ceci est illustré sur la FIGURE 3.14. Si la vitesse de contrôle $v_c$ est supérieure à $v_{lmax}$ durant une liaison, alors la correction temporelle est activée et l'évolution de l'instant cible $\tau$ est ralentie à la vitesse $v_{lmax}$. Une fois le début du noyau suivant atteint, l'évolution de $\tau$ est accélérée à la vitesse $v_{nr}$, jusqu'à ce qu'il rattrape l'instant qu'il aurait dû atteindre sans correction temporelle, ciblé par le potentiomètre. Les vitesses $v_{lmax}$ et $v_{nr}$ peuvent être définies dans les réglages de Vokinesis (voir le CHAPITRE 5, section 5.5.2.2).

### 3.3.6 Potentiomètres manuels

Nous avons testé plusieurs stratégies manuelles de contrôle continu du cadre rythmique, que nous présenterons dans cette section.

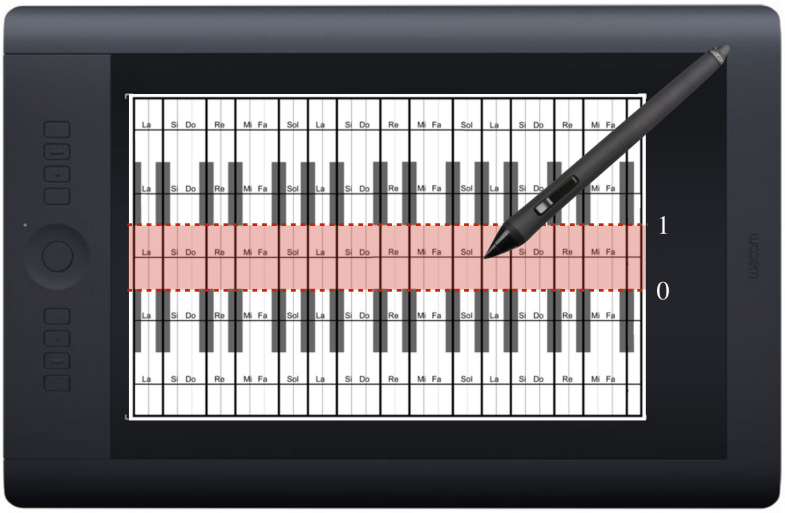


FIGURE 3.15 – *Le potentiomètre peut être simulé sur l'axe vertical de la tablette graphique.*

La première méthode que nous avons explorée consiste à utiliser la zone rouge de la tablette graphique présentée FIGURE 3.15 comme potentiomètre. Cette méthode fonctionne, mais il est difficile de se concentrer à la fois sur le contrôle de la mélodie et sur celui des motifs rythmiques qui peut être complexe. Sur la même surface, il est possible d'utiliser un doigt de la seconde main à la place du stylet. Un des problèmes dans ce cas de figure est le manque de retour tactiles : le visuel étant concentré sur le contrôle mélodique, et il est difficile de se repérer sans regarder la main gauche pour le contrôle rythmique. Ceci nous a menés à essayer des potentiomètre physiques, tel qu'un *crossfader* ou encore un joystick de manette de jeux vidéo. Dans tous les cas, l'influence des mouvements d'une main sur ceux de l'autre est trop grande, ce qui conduit à des variations de hauteur ou de rythme non contrôlées. Nous invitons les lecteurs et lectrices à déplacer leur seconde main rapidement de gauche à

droite, tout en essayant d'écrire le mot « *bonjour* » avec la première. Ils verront alors que la tâche est difficile : chaque main influence les mouvements de l'autre. Le principe de *synergie* dit que « l'évolution a sélectionné des [...] "mouvements naturels" qui impliquent des groupes de muscles et de membres qui travaillent (en grec *ergos*) ensemble (*syn*) » [Berthoz 1997]. En effet, [Rispal-Padel *et al.* 1982] ont montré sur un groupe de babouins que les projections neuronales impliquées dans un acte précis sont regroupées. Ils ont pu observer que l'activation d'une zone cérébrale impliquée dans les déplacements latéraux d'un membre déclenchait également une co-contraction des muscles des zones voisines. Nous pouvons donc penser qu'effectuer des mouvements latéraux avec chaque main qui soient indépendants (rythme à gauche et mélodie à droite) demanderait un entraînement assez long avant de pouvoir se défaire de ces contraintes. Le problème est plus faible lors de l'utilisation d'un mouvement avant-arrière de la seconde main, mais il persiste tout de même.

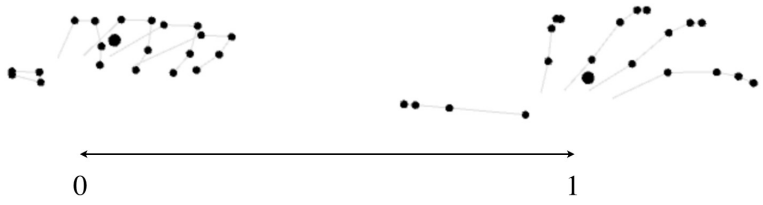


FIGURE 3.16 – *Représentation d'une main en position fermée (à gauche) et ouverte (à droite) à partir des données émises par la Leap Motion. La main fermée émet la valeur* 0*, la main ouverte la valeur* 1*, et toute position entre fermé - ouvert émet une valeur interpolée comprise entre* 0 *et* 1*.*

Pour nous affranchir des problèmes cités ci-dessus, nous avons créé un potentiomètre fictif dont la valeur varie selon l'état d'ouverture de la seconde main : sa valeur prend 1 lorsque la main est ouverte, et 0 lorsqu'elle est fermée (voir la FIGURE 3.16). Pour ce faire, nous utilisons une *Leap Motion*, un appareil qui capte et transmet les coordonnées $(x, y, z)$ des différents éléments des mains. Dans la FIGURE 3.16, les représentations d'une main fermée (à gauche) et ouverte (à droite) ont été obtenues à partir de ces données. Pour convertir les transitions de fermeture - ouverture de la main en une valeur de potentiomètre variant entre 0 et 1, nous avons utilisé le logiciel *Wekinator* [Fiebrink & Cook 2010], un logiciel d'apprentissage machine (ou *machine learning*) spécialement conçu pour des applications multimédias. La configuration consiste à envoyer à l'entrée de Wekinator des données de la Leap Motion correspondant à une main fermée tout en lui indiquant qu'elles doivent correspondre à une valeur de sortie de 0. La même opération doit être effectuée avec la main ouverte, en lui indiquant que le signal de sortie doit valoir 1. Une fois le modèle entraîné, Wekinator interpole les données complexes et multiples de la Leap Motion entre main fermée - main ouverte pour obtenir un signal de sortie simple et unique compris entre 0 et 1, qui représentera notre potentiomètre fictif. Cette mé-

thode de contrôle permet de réduire le problème d'influence de la seconde main sur la première. Si les lecteurs et lectrices essaient d'écrire avec leur première main tout en effectuant des cycles d'ouverture - fermeture avec la seconde, ils rencontreront sans doute moins de difficultés que lors de l'expérience précédente. En effet, une même zone musculaire serait liée à plusieurs zones neuronales, qui correspondraient elles à des tâches spécifiques [Berthoz 1997]. Ainsi, la zone dédiée à une action de déplacement latéral du bras et du poignet de la première main (écriture) serait plus éloignée de celle dédiée a une action d'ouverture - fermeture de la seconde main que de celle dédiée à ses déplacements latéraux. De plus, nul besoin de regarder sa seconde main pour savoir dans quelle position d'ouverture - fermeture elle se trouve. Enfin, ce geste peut s'effectuer très rapidement. Cependant, la Leap Motion n'est pas assez réactive pour le contrôle du cadre rythmique : le retard induit devient très vite gênant. Nous supposons alors qu'un jeu de capteurs dont la réponse serait plus rapide, tels qu'un gant de contrôle similaire à ceux de [Fels & Hinton 1998], offrirait une stratégie manuelle de contrôle continu du rythme syllabique de bonne qualité.

### 3.3.7 Potentiomètres pédestres

Il nous fût difficile de trouver une modalité de contrôle manuel et continu du cadre rythmique qui soit tout-à-fait satisfaisante. Nous nous sommes donc penchés vers l'utilisation de pédales d'expressions (Figure 3.17), qui sont des potentiomètres commandés par des mouvements pédestres. Cette solution semble mieux adaptée au contrôle simultané de la hauteur et du cadre rythmique, car l'influence des mouvements de contrôle pédestre du rythme sur ceux du contrôle chironomique de la hauteur est faible. Nous invitons les lecteurs et lectrices à une troisième expérience, qui consistera à écrire tout en tapant des pieds, talons cloués au sol, pour mimer le contrôle de la pédale. Cette tâche est sans doute la plus simple des trois, les mains et les pieds étant très éloignés, et les mouvements effectués assez différents. Le contrôle pédestre du rythme permet en plus de libérer la seconde main, qui pourra alors contrôler la hauteur d'une seconde voix, ou encore un autre paramètre vocal qui lui sera assigné.


Figure 3.17 – *Pédale d'expression Dunlop.*

L'utilisation du mode de fader *Solo NLN* présenté plus haut (et donc d'une pédale unique) n'est pas la meilleure solution pour le contrôle pédestre : la géométrie musculaire implique que les mouvements avant et arrière du pied ne fassent

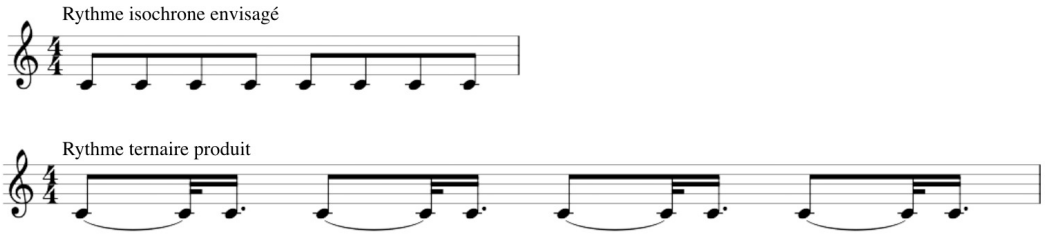


FIGURE 3.18 – *Erreur de contrôle rythmique lorsqu'une seule pédale est utilisée et que l'attention n'est pas concentrée sur le contrôle rythmique. Le rythme ternaire a été écrit avec une signature binaire afin de conserver le tempo.*

pas intervenir les mêmes muscles [Berthoz 1997], alors qu'ils sont tous deux sensés contrôler une transition NLN. Nous avons ainsi remarqué une tendance à produire le rythme ternaire de la FIGURE 3.18 au lieu du rythme isochrone envisagé lorsque l'attention n'est pas concentrée sur le contrôle du rythme. C'est pour cette raison que nous avons mis en place le mode de contrôle continu *Duo NLN*, dont un exemple peut être visualisé dans le vidéo `Ex07` à partir de 2m12s. Chaque transition NLN est contrôlée par un mouvement avant des pieds (le contrôle du cadre rythmique peut alors être comparé au contrôle de la grosse caisse à la double pédale dans le cas de la batterie). Les deux mouvements se faisant en alternance, ce mode permet un séquencement plus rapide du cadre rythmique : un mouvement avant du second pied peut être effectué avant la fin d'un mouvement avant du premier.

Bien que le mode de contrôle continu *LN* soit une bonne analogie avec les mouvements d'ouverture/fermeture de la mandibule impliqués dans le séquencement des syllabes selon la théorie F/C [MacNeilage 1998] (section 3.2.5), nous ne considérerons par la suite plus que les modes *NLN*, qui permettent un contrôle au moins deux fois plus rapide. Comme nous l'avons vu plus haut, ils offrent également une bonne analogie avec les principes de phonologie articulatoire de [Browman & Goldstein 1992] (section 3.2.4) : les cibles des mouvements pédestres correspondent aux cibles vocaliques (les noyaux rythmiques), et les transitions entre deux cibles pédestres correspondent aux liaisons vocaliques.

### 3.3.8 Mode *Loop*

Le mode *Loop* (ou boucle) peut être utilisé avec n'importe-quel mode de contrôle rythmique ou temporel. Ce mode consiste à faire boucler la position de l'instant cible entre deux FCP prédéfinis. Ainsi, si l'instant cible dépasse le FCP final ($FCP_{end}$), il retournera au FCP initial ($FCP_{start}$), et vice versa. Le calcul de l'instant cible en mode *Loop* se fait selon l'équation (3.5) :

$$\tau_{loop} = (\tau - FCP_{start}) \mod (FCP_{end} - FCP_{start} + 1) + FCP_{start} \tag{3.5}$$

Afin d'obtenir une qualité optimale, nous préconisons l'utilisation d'une consonne identique pour les position $FCP_{start}$ et $FCP_{end}$. Afin d'éviter toute discontinuité

dans le signal de synthèse, nous effectuons une interpolation du signal entre le phonème final et le phonème initial, qui sera présentée en détails dans la section 2.3.3.

## 3.4 Préparation et étiquetage des signaux originaux

Dans cette section, nous allons d'abord présenter les bonnes pratiques d'enregistrement des signaux originaux pour l'obtention d'une synthèse de bonne qualité. Nous détaillerons ensuite les règles de positionnement des FCP selon l'étiquetage des phonèmes et le mode de contrôle rythmique. Cela nous permettra alors d'indiquer comment étiqueter les phonèmes pour assurer le bon fonctionnement de toutes les fonctionnalités de contrôle rythmique de Vokinesis.

### 3.4.1 Enregistrement des signaux originaux

N'importe quel signal vocal peut être utilisé et modifié dans Vokinesis. Cependant, la qualité du signal de synthèse peut être optimisée si certaines règles d'enregistrement sont respectées, outre l'absence de réverbération, d'écho, ou de bruit.

Pour le contrôle du chant, il est préférable d'enregistrer des signaux originaux à fréquence fondamentale constante ou peu variable, et sans vibrato. En effet, un changement de fréquence fondamentale modifie le timbre de la voix. Si les interprètes souhaitent garder une note constante et qu'un vibrato a été produit lors de l'enregistrement, alors les variations de timbre induites par le vibrato seront audibles dans le signal de synthèse, tel un vibrato d'enveloppe spectrale à fréquence fondamentale fixe. Il est également préférable d'exagérer légèrement l'articulation de tous les phonèmes afin d'assurer une prononciation correcte lors de leur maintien, ou bien lors de leur accélération. Les signaux originaux doivent être enregistrés de manière isochrone, sans quoi un geste de contrôle identique aura un rendu différent selon la longueur des noyaux. Un tempo d'environ 300 bpm semble judicieux, car il correspond à des groupes rythmiques de 200 ms, durée légèrement supérieure à la moyenne inter-linguistique de la durée d'une syllabe [Wagner 2008].

Pour la parole, il peut être utile d'enregistrer des signaux qui possèdent déjà un rythme naturel, dans la mesure où l'énoncé reste neutre et à un tempo normal. En effet, il est probable qu'un rythme similaire soit reproduit par les interprètes, et les différences de durées entre groupes rythmiques deviennent alors moins gênantes, surtout pour le français dont les durées inter-nucléiques varient peu.

### 3.4.2 Règles de positionnement des FCP

Comme nous l'avons vu, un groupe rythmique contient deux FCP. L'un est placé au sein du noyau ($P_n$), et l'autre au sein de la liaison ($P_l$). La façon précise dont ils doivent être positionnés au sein de chaque phase dépend du mode de contrôle rythmique. Nous allons détailler ces règles empiriques ci-dessous.

#### 3.4.2.1 Mode *Tap*

En mode *Tap*, un appui sur la touche de contrôle doit déclencher un noyau rythmique. En effet, en considérant le cas des percussions, le contrôle rythmique serait impossible si la frappe ne déclenchait pas instantanément un événement rythmique. Les $P_l$ doivent donc être placés au niveau de la dernière consonne d'une liaison articulatoire. Si c'est une plosive non-voisée, le $P_l$ doit être placé durant le silence qui précède l'explosion. Ainsi, un maintien du bouton relâché permettra de maintenir ce silence, et un appui déclenchera l'explosion. Sinon, placer le $P_l$ au centre de la consonne semble être une bonne solution pour obtenir la meilleure prononciation possible lors de son maintien. Si la liaison ne contient pas de consonne, alors le $P_l$ correspondant sera placé au niveau de la délimitation des deux voyelles consécutives.

Un relâchement du bouton de contrôle doit déclencher le début d'une liaison rythmique. Les $P_n$ doivent donc être placés à la fin de la partie stable du noyau rythmique. Ceci assure sa prononciation correcte lors du maintien du bouton d'un part, un déclenchement immédiat de la transition NL lors de son relâchement d'autre part, et permet également d'atteindre le $P_l$ suivant le plus rapidement possible. Si le noyau original est assez court (locution normale), son centre sera suffisamment proche de la liaison suivante pour y placer le $P_n$, et permettra la meilleure prononciation possible lors de son maintien.

#### 3.4.2.2 Mode *Fader*

Le positionnement des FCP en mode *Fader* s'effectue selon deux conditions essentielles :

1. La correspondance mouvement / événement rythmique doit être conservée pour chaque syllabe.
2. La prononciation des voyelles doit être de bonne qualité lorsque le potentiomètre est en position extrême.

Une solution logique pour satisfaire la première condition consisterait à faire en sorte que les positions extrêmes du potentiomètre correspondent aux p-centers. Cependant, la seconde condition ne serait alors pas respectée, les p-centers ne se trouvant pas dans les parties stables des noyaux rythmiques. Si le p-center ne peut pas correspondre aux positions extrêmes du potentiomètre, il doit alors correspondre à son centre, afin qu'il ait la même position à chaque transition NLN effectuée. Les $P_l$ doivent donc se trouver, comme pour le mode *Tap*, dans la dernière consonne d'une liaison rythmique. La meilleure solution pour respecter la seconde condition consiste alors à placer les $P_n$ au centre des noyaux. En effet, si nous les placions au début de la partie stable du noyau, alors la voyelle prendrait une trop grande place dans le mouvement de contrôle de la prochaine transition NL. À l'inverse, si nous le placions à la fin, alors la voyelle aurait pris trop de place dans la transition LN précédente. Placer les $P_n$ au centre des noyaux permet donc un contrôle plus précis des transitions LN et NL. Les règles de placement automatique sont donc *a priori*

identiques pour les modes *Tap* et *Fader*. Il existe cependant quelques exceptions, que nous détaillerons dans la section 3.4.3.

### 3.4.3 Cas particuliers

Le FCP de l'anacrouse est toujours noté $P_l(0)$. Si l'anacrouse contient des consonnes, elles devront être prononcées avant le premier noyau rythmique. Nos règles de placement automatique se fient à l'étiquetage des phonèmes pour placer $P_l(0)$ au début de la première consonne. Lorsqu'un silence est indiqué dans l'étiquetage des phonèmes, le FCP qui le suit sera toujours placé tel que nous venons de le décrire pour $P_l(0)$.

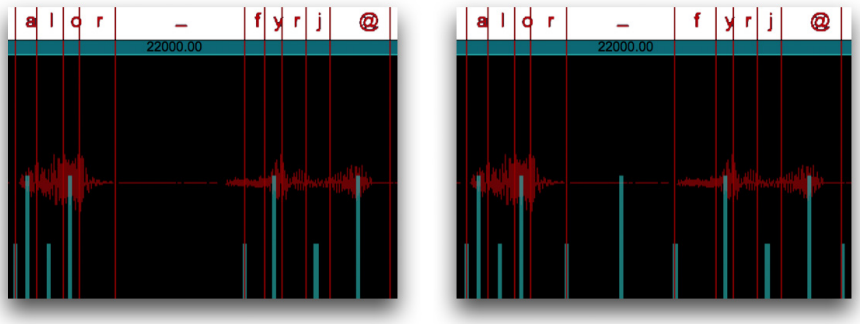


Figure 3.19 – *Cas particuliers de positionnement des FCP pour la phrase « alors, furieux » (forme d'onde en rouge). En mode Tap (à gauche), le silence est considéré comme appartenant à une liaison rythmique, et le premier $P_l$ qui suit le silence est au centre de la consonne anacrouse. En mode Fader (à droite), le silence central est considéré comme un noyau rythmique, deux FCP sont donc rajoutés.*

Sur la Figure 3.19, la forme d'onde de la phrase « alors, furieux » étiquetée phonétiquement et marquée par des FCP (lignes verticales bleues, les grandes sont des $P_n$ et les petites des $P_l$) est représentée pour le mode *Tap* à gauche et pour le mode *Fader* à droite. Ignorons pour l'instant la première partie de la phrase, et imaginons que le FCP qui suit le silence soit $P_l(0)$. En mode *Fader*, $P_l(0)$ doit être placé juste avant la première consonne. Cela permet d'assurer sa prononciation complète lors du premier déplacement du potentiomètre. Cependant, en mode *Tap*, un tel placement risque de décaler l'instant d'occurrence du p-center par rapport à l'instant de frappe de la touche de contrôle. Dans le cas où l'anacrouse ne contient qu'une consonne, il est plus judicieux de placer $P_l(0)$ en son centre, comme sur la Figure 3.19, à gauche, sur le [f] qui suit le silence (la délimitation du démarrage du phonème a été avancée pour déplacer le FCP correspondant). Le contrôle de l'anacrouse consistera alors à anticiper le posé du stylet avant la première frappe de la touche de contrôle, pour prononcer le centre de la première consonne, avant le déclenchement du premier noyau rythmique. Le cas où l'anacrouse contient plusieurs consonnes constitue une limitation du mode *Tap* : les interprètes devront anticiper la frappe pour prononcer toutes le consonnes et faire retentir le début du noyau

rythmique au bon moment.

Sur la FIGURE 3.19, nous pouvons voir que l'organisation des FCP dans les silences diffère selon le mode de contrôle rythmique. En mode *Tap*, les silences sont contrôlés comme des liaisons rythmiques : un relâchement de la touche prononce le silence, et un appui prononce le noyau rythmique suivant. En mode *Fader* ils sont contrôlés comme des noyaux rythmiques : une position extrême du potentiomètre correspond au centre de la respiration, permettant ainsi un contrôle précis de sa temporalité.

### 3.4.4 Étiquetage des phonèmes

TABLEAU 3.2 – Règles d'étiquetage des phonèmes pour une prononciation optimale.

| **Phase rythmique** | **Composition phonétique** | **Étiquetage** |
|---|---|---|
| Noyaux | Voyelle | Voyelle (partie stable uniquement) |
| | Consonne syllabique | |
| Liaisons | Consonne | Consonne |
| | transition VV | Consonne voisée |
| Silences | Respiration | Silence |
| | Occlusive glottique | Plosive |

Les règles de placement des FCP que nous avons présentées section 3.4.2 ont été développées selon les règles phonotactiques du français : le noyau rythmique correspond toujours à une voyelle. Ainsi, une consonne syllabique devra être étiquetée comme une voyelle afin d'indiquer au système qu'il s'agit d'un noyau rythmique. Par ailleurs, les liaisons sont indiquées au système par des consonnes. Ainsi, dans le cas de deux voyelles consécutives, la transition qui les sépare peut être étiquetée comme une consonne afin d'indiquer au système d'y effectuer une correction temporelle des liaisons rythmiques en mode *Fader* (section 3.3.5.2) pour conserver une intelligibilité optimale. De même, seule la partie stable d'un noyau doit être étiquetée par une voyelle afin de conserver les durées des parties transitoires en mode *Fader*. Pour le cas des silences, ceux correspondant à une respiration devront être étiquetés par des silences, mais ceux qui correspondent à des occlusives glottiques devront être étiquetés comme des plosives. Cela permettra en mode *Fader* d'y appliquer la correction temporelle d'une part, et d'éviter d'avoir à contrôler ce silence consonantique comme un noyau rythmique d'autre part. Les règles d'étiquetage que nous venons d'énoncer sont résumées dans le TABLEAU 3.2.

## 3.5 Évaluation des méthodes de contrôle du rythme articulatoire

Afin d'évaluer nos méthodes de contrôle rythmique, nous avons mis en place des expériences de contrôle temporel et mélodique de la parole et du chant. La

première d'entre elles fournit une évaluation objective des capacités d'un groupe de sujets à imiter le rythme et l'intonation de phrases parlées en mode *Tap*. Elle sera présentée ci-dessous. Nous avons également mis en place deux autres expériences. L'une est une réplique de celle que nous venons d'évoquer, avec plus de sujets et avec un paradigme d'imitation naturelle supplémentaire. L'autre consistait à évaluer nos méthodes de contrôle rythmique (*Tap* et *Fader*) dans le cadre du chant. Le temps imparti pour cette thèse ne nous a pas permis d'effectuer l'analyse objective des résultats de ces deux dernières expériences. Cependant, certains sujets nous ont confié leurs impressions concernant nos méthodes de contrôle, et nous en fournirons une synthèse.

### 3.5.1 Première expérience de contrôle du rythme de la parole

La capacité d'un groupe de sujet à reproduire le rythme et l'intonation d'un jeu de phrases naturelles en mode *Tap* a été évaluée par un test d'imitation prosodique, dont le protocole, les méthodes de mesure et les résultats seront présentés dans ci-dessous.

#### 3.5.1.1 Protocole

Un jeu de 8 phrases dont la taille variait de 2 à 9 syllabes, enregistrées par un homme et une femme, étaient présentées dans un ordre aléatoire à 8 sujets. Les phrases que nous avons utilisées sont présentées dans le Tableau 3.3. Ce sont les mêmes que celles utilisées par [d'Alessandro *et al.* 2011]. Les sujets pouvaient effectuer autant d'essais qu'ils le souhaitaient pour chaque phrase. Une fois convaincus par leur performance, ils pouvaient passer à la phrase suivante. Cette procédure devait être effectuée deux fois : la première fois sans contrôle de la hauteur (la hauteur de synthèse était laissée identique à l'originale), et une autre avec le contrôle simultané de la hauteur et du rythme. Le test durait une cinquantaine de minutes, et les sujets n'avaient reçu qu'un entraînement minimal.

Tableau 3.3 – *Phrases utilisées pour nos tests d'imitation prosodique*

| **Syllabes** | **Phrase** | **Transcription phonétique** |
|---|---|---|
| 2 | Salut | [saly] |
| 3 | Répétons | [ʁepetõ] |
| 4 | Marie chantait | [maʁiʃãtɛ] |
| 5 | Marie s'ennuyait | [maʁisãnɥijɛ] |
| 6 | Marie chantait souvent | [maʁiʃãtɛsuvã] |
| 7 | Nous voulons manger le soir | [nuvulõmãʒeləswaʁ] |
| 8 | Sophie mangait des fruits confits | [sofimãʒɛdefʁɥikõfi] |
| 9 | Sophie mangeait du melon confit | [sofimãʒɛdyməlõkõfi] |

#### 3.5.1.2 Mesures

Les phonèmes originaux ont été étiquetés avec Vokinesis, en utilisant les représentations visuelles des signaux et des spectrogrammes des phrases originales. Tout comme le faisait [Levitt 1991], chaque incertitude à propos d'une certaine délimitation était résolue à l'écoute, et la fin d'une phrase était déterminée par la fin de la périodicité.

La position de l'instant cible était enregistrée pour chaque performance. Elle a ensuite été utilisée pour déterminer l'étiquetage des phonèmes de synthèse, selon l'étiquetage original. Les durées des groupes rythmiques étaient mesurées à partir de l'étiquetage des phonèmes. Les durées des groupes rythmiques originaux sont notées $\Delta_{orig}(\gamma)$ et celles des groupes rythmiques des imitations de synthèse $\Delta_{imit}(\gamma)$, avec $\gamma$ le numéro de groupe rythmique. Pour évaluer la précision de reproduction du rythme, nous avons utilisé les différences entre $\Delta_{orig}$ et $\Delta_{imit}$, selon l'équation (3.6) :

$$\Delta_{diff}(\gamma) = \Delta_{orig}(\gamma) - \Delta_{imit}(\gamma) \tag{3.6}$$

#### 3.5.1.3 Résultats

Cette étude se concentrant principalement sur la qualité du contrôle rythmique, nous ne parlerons que brièvement du contrôle de la hauteur qui sera traité dans le Chapitre 4.

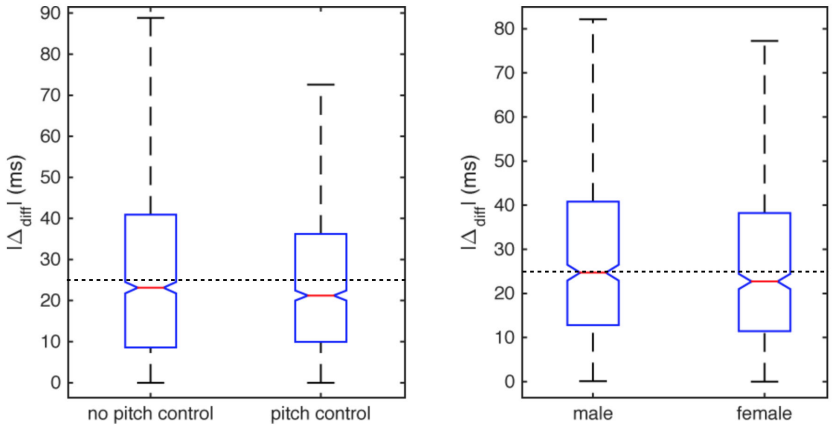


Figure 3.20 – $|\Delta_{diff}|$ *avec et sans contrôle de la hauteur (à gauche), pour la voix d'homme et la voix de femme (à droite).*

Pour tous les sujets et pour toute les phrases, $\overline{|\Delta_{diff}|}$ avoisine les 20ms. Or, d'après [Wagner 2008], le plus courte différence de durée perceptible (ou JND pour *Just Noticeable Difference*) pour une unité temporelle de la taille d'une syllabe est d'environ 25ms. Nous pouvons donc conclure que le cadre rythmique de la parole peut être reproduit avec une très bonne précision. La Figure 3.20 montre que ni le contrôle de la hauteur ni le locuteur original n'ont d'effet significatif sur la précision rythmique ; $\overline{|\Delta_{diff}|}$ est toujours inférieur à 25ms. La Figure 3.21 montre deux effets :

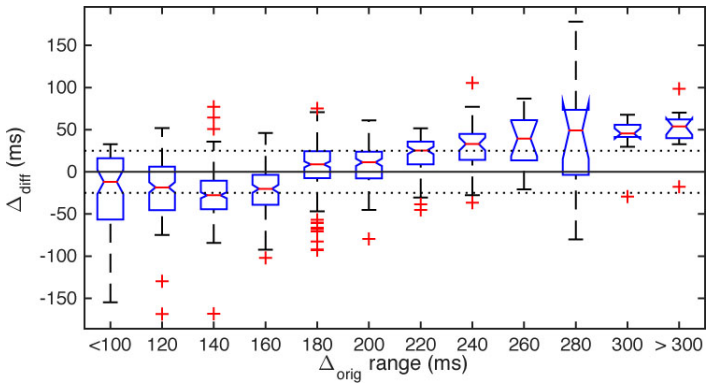


FIGURE 3.21 – $\Delta_{diff}$ *pour différentes valeurs de* $\Delta_{orig}$. *La JND de la longueur d'une syllabe est indiquée par les lignes pointillées.*

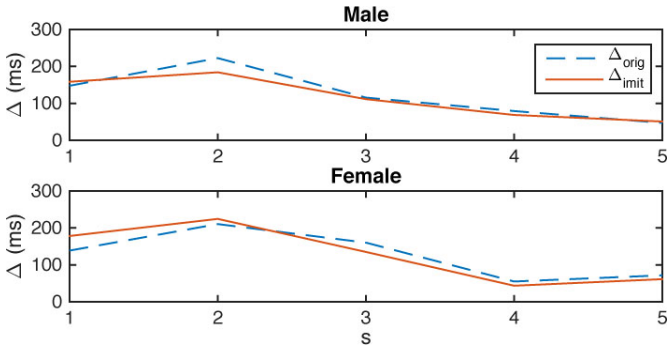


FIGURE 3.22 – $\Delta_{orig}$ *(lignes pointillées) et moyenne de* $\Delta_{imit}$ *(lignes continues) pour la phrase composée de 5 syllabes, enregistrées par l'homme (en haut) et la femme (en bas), pour tous les sujets, avec et sans contrôle de la hauteur.*

1) Les syllabes longues (> 220ms) sont plus difficiles à reproduire, puisque $\overline{|\Delta_{diff}|}$ peut atteindre 50ms, alors que pour les syllabes plus courtes, $\overline{|\Delta_{diff}|}$ a une valeur maximale de 25ms.

2) Il y a une tendance à raccourcir les longues syllabes et à rallonger les courtes. Cependant, la FIGURE 3.22 montre que les sujets étaient en mesure de suivre les variations de durée des groupes rythmiques originaux au sein d'une phrase.

#### 3.5.1.4 Contrôle simultané de la hauteur et du cadre rythmique

La FIGURE 3.23 montre l'évolution de la hauteur de la phrase à 8 syllabes enregistrée par l'homme, ainsi que celle des 2 signaux reproduits. Bien que les courbes de hauteur évoluent avec des variations moins abruptes, la forme générale est conservée lors de la synthèse, comme nous pouvions nous y attendre selon [d'Alessandro *et al.* 2011]. D'après les auteurs, la différence perceptive entre les contours naturels et les meilleures imitations chironomiques serait nulle (voir la section 4.1.1).

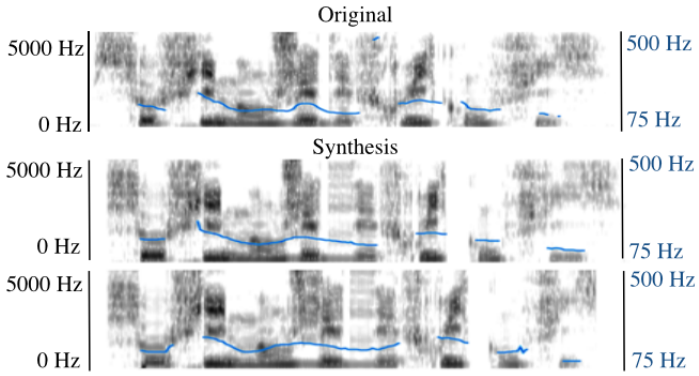


FIGURE 3.23 – *Spectrogrammes et courbes d'intonation de la phrase à 8 syllabes original (en haut) et re-synthétisée (au milieu et en bas). L'axe de gauche correspond au spectrogramme, et celui de droite à la fréquence fondamentale.*

### 3.5.2 Évaluation subjective des modalités de contrôle rythmique de la parole et du chant

Nous avons récemment reproduit le test de la section précédente avec un plus grand nombre de sujets (15 sujets), et avec en plus une modalité d'imitation naturelle des phrases originales. Nous avons également souhaité évaluer la précision du contrôle du rythme musical, en demandant à un groupe de 16 sujets musiciens de produire les phrases musicales présentées FIGURE 3.24 en suivant un métronome, avec différentes modalités de contrôle. Nous différencierons ces deux tests par les appellations *test parole* et *test chant*.

Les séries de test ont été menées par notre collègue psychologue Gabriela Patino-Lakatos, qui a pu relever les impressions des sujets. L'analyse objective des performances des sujets est encore en cours. Nous fournissons donc ici une courte synthèse des impressions des sujets par rapport à nos méthodes de contrôle performatif de la voix.

#### 3.5.2.1 Test parole

Pour le test parole, les 2/3 des sujets ont fait part de la difficulté éprouvée lorsque le rythme et la mélodie sont contrôlés de manière simultanée. Certains ont jugé la tâche stressante. Les autres sujets semblent avoir développé des stratégies liées à la concentration ciblée sur le contrôle mélodique : les variations mélodiques à la main droite dirigeraient les variations rythmiques de la main gauche. L'un des sujets a d'ailleurs éprouvé des problèmes de synchronisation lors du contrôle du rythme uniquement, qui auraient disparu avec le contrôle simultané de la mélodie. Un autre sujet a évoqué la similarité avec le contrôle de l'onde Martenot : mélodie à la main droite, et rythme à la main gauche. Enfin, les phrases qui semblent avoir causé des difficultés dans le contrôle rythmique sont celles qui contiennent la plus grande variabilité de durée des groupes rythmiques. Dans l'ensemble, les sujets ont

trouvé ce test assez difficile, et plusieurs d'entre eux l'ont trouvé stressant.

#### 3.5.2.2 Test chant

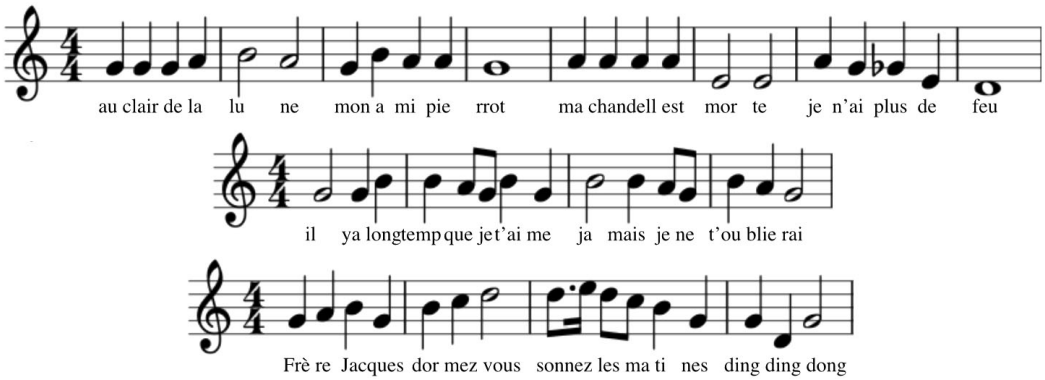


Figure 3.24 – *Partitions des chants proposés aux sujets. De haut en bas : Au clair de la lune, À la claire fontaine, frère Jacques.*

Lors du test chant, nous avons demandé à un groupe de 16 sujets de reproduire le plus précisément possible les phrases musicales présentées Figure 3.24 en suivant un métronome à un tempo de 120 bpm, avec différentes modalités de contrôle : avec un barre espace (mode *Tap*), avec une pédale d'expression (mode *Fader Solo NLN*), puis deux (*Duo NLN*). Pour chacune de ces modalités, les sujets devaient d'abord reproduire le rythme uniquement de chacune des phrases musicales, puis le rythme et la mélodie de façon simultanée. Pour chaque phrase musicale, les sujets pouvaient écouter un exemple audio et la partition leur était présentée. Ils disposaient d'autant d'essais qu'ils le souhaitaient avant de valider celui qui leur semblait convenir.

Sur les 16 sujets, 7 ont préféré le contrôle rythmique avec la barre espace et 7 autres ont préféré la pédale. Les 2 sujets restants n'ont pas éprouvé de préférence particulière. En ne considérant que le contrôle continu, 12 ont préféré le contrôle rythmique bi-pédestre, souvent jugé plus naturel et plus intuitif, et plus efficace pour la production de rythmes rapides. Seuls 2 sujets ont préféré le contrôle mono-pédestre.

Dans l'ensemble, les sujets ont éprouvé du plaisir à effectuer ce test, qui n'a pas apporté le stress ressenti lors du test parole. Les sujets qui ont passé les deux tests ont d'ailleurs souvent exprimé leur préférence du contrôle du chant par rapport à celui de la parole. Nous pensons que la modalité de contrôle ouverture/fermeture de la main (section 3.3.6) en mode *Fader Solo NLN* permettrait de réduire considérablement le stress induit par le test parole, car ce sont des mouvements très rapides (au moins deux fois plus rapide que le paradigme de frappe) et peu fatigants. Cependant, nous pensons également que le contrôle binaire du rythme serait moins stressant pour des langues au tempo syllabique plus lent, telles que l'anglais, l'allemand ou le polonais [Wagner 2008].

Comme nous le verrons dans le CHAPITRE 6, le contrôle du chant en mode *Tap* et *Fader* a été testé par plusieurs musiciens dans la cadre de représentations publiques. Ils étaient en mesure de jouer de façon synchronisée avec d'autres instruments et entre voix de synthèse, avec un temps d'entraînement relativement court.

## 3.6 Conclusion

Ce chapitre a permis de mettre en lumière l'existence d'une unité suprasegmentale comparable à la syllabe du point de vue de sa durée, mais dont la composition segmentale diffère légèrement. Cette unité, que nous avons nommée *groupe rythmique*, offre une meilleure représentation inter-linguistique du séquencement rythmique de la voix que la syllabe. En effet, quelle que soit la composition phonétique d'un groupe rythmique, il comportera toujours deux phases : le *noyau* et la *liaison* rythmiques. Au contraire, le nombre de phases que comporte une syllabe varie selon son contenu phonétique (l'attaque et la coda peuvent exister ou non). Par ailleurs, les groupes rythmiques peuvent être directement associés à des notations musicales (un événement rythmique d'une partition définit la durée d'un groupe rythmique), alors que le lien entre la syllabe et les notations musicales est plus ambigu.

Nous avons exploré différentes méthodes de contrôle rythmique basées sur deux modes de contrôle principaux : les modes *Tap* et *Fader*. Le mode *Tap* permet de contrôler le séquencement du cadre rythmique de façon binaire. Ce mode de contrôle représente une bonne analogie avec la représentation biphasique de la théorie F/C : les états *appuyé* et *relâché* d'une touche de contrôle peuvent être apparentés aux états *ouvert* et *fermé* du conduit vocal. Les analyses objectives des performances d'un groupe de sujets à imiter le rythme de phrases parlées ont pu montrer que ce mode de contrôle offrait une remarquable précision. Cependant, les analyses subjectives ont laissé entrevoir des difficultés, voire du stress lors de la production de rythmes parlés, qui n'ont pas été ressentis de la même manière dans le cas de la voix chantée. Il serait donc intéressant de vérifier si cette méthode de contrôle est mieux adaptée à des langues dont le tempo est plus lent, telles que l'anglais, l'allemand ou le polonais [Wagner 2008]. Le principal défaut du mode *Tap* vient de sa nature binaire, qui ne permet pas le contrôle de la vitesse de lecture des liaisons rythmiques. Ce problème peut être contourné grâce à l'utilisation du mode *Fader*, qui, grâce à des potentiomètres, permet un contrôle continu des liaisons rythmiques. Parmi les différentes interfaces continues que nous avons testées, nous avons surtout retenu l'utilisation de pédales d'expression. En effet, par synergie, l'utilisation de potentiomètres avec la main gauche influençait trop le contrôle de la hauteur effectué avec la main droite. De plus, le contrôle pédestre du rythme permet de libérer la main gauche qui, comme nous le verrons plus tard, pourra alors contrôler la hauteur d'une seconde voix de synthèse. Malgré le fait que nous n'ayons pas eu le temps d'évaluer la précision rythmique qu'offre ce mode de contrôle de façon objective, nous verrons dans le CHAPITRE 6 qu'il permet à des musiciens de jouer de façon synchronisée sans difficulté particulière. Les analyses subjectives n'ont pas permis

de déceler une préférence pour le mode *Tap* ou *Fader*. Par contre, en mode *Fader*, le contrôle bi-pédestre semble être préféré au contrôle mono-pédestre.

À l'heure actuelle, nous pensons que les modes *Tap* et *Fader* sont tous deux tout à fait adaptés au contrôle rythmique du chant. En effet, ils ont été pratiqués par des musiciens à plusieurs reprises, sans difficulté particulière pour la synchronisation. Le mode *Fader* offre cependant une plus grande liberté expressive que le mode *Tap*, car il permet un contrôle fin des durées segmentales.

Pour le cas de la parole, nous pensons que le contrôle continu des liaisons rythmiques est rarement nécessaire, car le tempo d'une phrase parlée est généralement assez rapide. Nous pensons d'ailleurs qu'il est souvent trop rapide pour un contrôle aisé en mode *Tap* : même si les sujets étaient capables d'imiter le rythme de phrases parlées avec précision, beaucoup ont jugé la tâche difficile, et parfois même stressante, ce qui n'a pas été le cas pour la tâche musicale. Nous faisons l'hypothèse que la meilleure solution consisterait à utiliser le mode *Fader* avec des gestes d'ouverture/fermeture de la main gauche (pour les droitiers). En effet, ces gestes semblent avoir peu d'influence synergique sur le contrôle de la hauteur à la main droite, ils sont extrêmement rapides, et peu fatigants.

Quoi qu'il en soit, nos méthodes de contrôle rythmique de la voix nécessitent un étiquetage des phonèmes préalable. De longs signaux originaux demandent donc un temps de préparation non-négligeable. Cependant, nous pouvons sans difficulté imaginer l'utilisation de signaux originaux calculés par des synthétiseurs vocaux à partir du texte. En effet, de tels systèmes sont capables de fournir des signaux vocaux qui respectent nos préconisations d'enregistrement (hauteur constante et rythme isochrone), et qui soient déjà étiquetés phonétiquement, et donc prêtes à être contrôlées rythmiquement.

Chapitre 4

# Contrôle expressif de la hauteur et de la qualité vocale

## Sommaire



Dans le chapitre précédent, nous avons présenté les différentes stratégies de contrôle du rythme de la parole que nous avons explorées lors de nos travaux. Ce chapitre se concentre sur le contrôle de la hauteur pour la parole (intonation) et pour le chant (mélodie), avec une attention particulière sur le contrôle des modulations expressives associées (variations de hauteur et de qualité vocale). Dans la première section, nous présenterons d'abord les travaux d'évaluation du contrôle de la hauteur par des gestes d'écriture (ou chironomiques, du grec *cheir* : la main, et *nomos* :

la règle) dans le cas de la parole et du chant. Nous y verrons ensuite les nouvelles méthodes de contrôle de la hauteur et de la qualité vocale qui ont été mises en place lors de nos travaux. La seconde section présente les stratégies existantes pour le contrôle expressif d'un synthétiseur avec un clavier MIDI, et nous montrerons comment les adapter au contrôle de la synthèse vocale. La section suivante fournira une présentation des interfaces de type PMC (*Polyphonic Multidimensionnal Controllers*), et de leur adaptation au contrôle de la synthèse vocale. Les capacités de ces trois types d'interfaces pour un contrôle expressif du chant seront comparées dans une quatrième section. Nous y verrons que les claviers MIDI offrent des possibilités inférieures aux tablettes graphiques et PMC. Enfin, les capacités d'amélioration de l'expressivité de signaux de parole issus d'un système de synthèse HMM-TTS expressive par modification chironomique de la hauteur seront démontrées dans une cinquième section.

## 4.1 Tablettes graphiques

La puissance expressive du contrôle de la hauteur par des gestes d'écriture et de dessin sur une tablette graphique semble avoir convaincu de nombreux chercheurs [Kessous 2004b, D'Alessandro & Dutoit 2007, Le Beux *et al.* 2007, Astrinaki *et al.* 2012, Feugère *et al.* 2017]. En effet, l'apprentissage du maniement du stylo dès le plus jeune âge implique une maîtrise déjà experte de cette modalité. Nos travaux poursuivent ceux que nous allons présenter dans les sections 4.1.1 à 4.1.4. En effet, nous utilisons les mêmes méthodes de contrôle de hauteur intonative et mélodique de la voix. Nous présenterons dans les sections 4.1.5 à 4.1.6 les nouvelles stratégies de contrôle chironomique de la hauteur et de la qualité vocale que nous avons eu l'occasion d'explorer. Nous y verrons comment utiliser la tablette graphique pour un contrôle polyphonique, pour la production de mélodies de type yodel, ainsi que pour le contrôle de la tension vocale et de la taille du conduit vocal.

### 4.1.1 Contrôle intonatif dans le cas de la parole

L'utilisation d'une tablette graphique pour contrôler la hauteur vocale a soulevé la questions suivante : dans quelle mesure les gestes chironomiques sont-ils capables de reproduire les gestes vocaux ? Les travaux effectués par [d'Alessandro *et al.* 2011] ont tenté d'y répondre en demandant à un groupe de sujets non entraînés d'imiter au mieux les contours intonatifs de phrases originales, à l'aide du système Calliphony d'une part (présenté section 3.1), et avec leurs voix naturelles d'autre part. Les analyses ont été effectuées de façon objective et subjective : les distances entre les contours intonatifs imités et originaux ont été calculées par ordinateur, et un test perceptif sur un groupe de 15 sujets a été mené pour évaluer la qualité des imitations chironomiques. L'analyse objective a montré que les imitations vocales étaient légèrement meilleures que les imitations chironomiques, mais que les deux étaient tout de même comparables. Les contours chironomiques possédaient des variations moins abruptes que les contours intonatifs naturels. Cependant, du point de vue

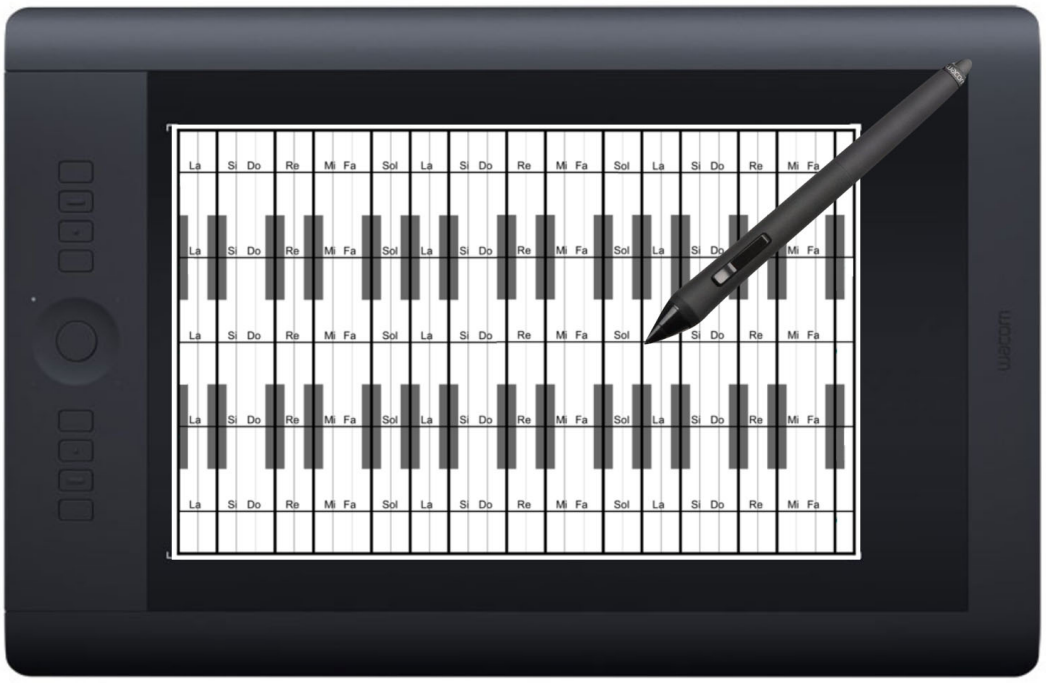


FIGURE 4.1 – *Masque de contrôle mélodique ajouté à la tablette graphique Wacom Intuos 5 Touch.*

perceptif, les meilleures imitations chironomiques étaient quasiment indiscernables des productions naturelles originales. Ce travail a donc offert un premier argument en faveur de la suppléance du geste vocal par le geste manuel.

### 4.1.2 Contrôle mélodique dans le cas du chant

D'un point de vue musical cette fois, les travaux de [d'Alessandro *et al.* 2014] ont cherché à évaluer la justesse et la précision offerte par cette interface en utilisant le synthétiseur Cantor Digitalis, présenté section 1.3. La justesse est mesurée par la moyenne de la différence entre les hauteurs d'une production mélodique et les hauteurs cibles, alors que la précision est mesurée par l'écart type entre les hauteurs cibles et les hauteurs produites. Trois tâches musicales ont été proposées à des groupes de 20 et 28 sujets, pour la plupart musiciens. Dans chacune de ces tâches, les sujets entendaient un exemple chanté accompagné de la partition correspondante, et il leur était demandé de reproduire cet exemple de façon chironomique d'une part, et avec leur voix naturelle d'autre part, en suivant un tempo imposé par un métronome. La première tâche consistait à produire différents intervalles montants et descendants (donc deux notes) à un tempo de 120 bpm (battement par minutes). La seconde consistait à produire différentes mélodies isochrones de 7 notes à 120 bpm également. La troisième consistait à produire des séries de doubles intervalles (3 notes) montants/descendants, ou descendants/montants, à trois tempi différents : 120, 179 et 240 bpm. Le contrôle de la hauteur s'effectuait grâce à une tablette

graphique sur laquelle était apposé le masque de la Figure 4.1. Les résultats ont permis de montrer que le contrôle chironomique offrait pour la plupart des sujets une meilleure précision et une meilleure justesse que la voix naturelle, quel que soit l'écart entre les deux notes d'un intervalle, la durée de la mélodie ou le tempo. Cependant, ces travaux ont également montré que le contrôle chironomique de la mélodie était largement dégradé en l'absence de repères visuels. De plus, [Perrotin 2015] a comparé les performances d'un groupe de sujets à viser un point selon trois modalités : avec un retour visuel seulement, avec un retour auditif seulement, puis avec les deux en même temps. Les résultats ont montré que « *le retour auditif perd toute influence sur le mouvement moteur en présence d'une modalité visuelle.* [...] *L'influence plus forte du retour visuel permet un contrôle plus simple de l'instrument.* » Pour un débutant, le contrôle mélodique consiste donc principalement à viser les lignes verticales du masque de la Figure 4.1 pour atteindre la note désirée.

### 4.1.3 Justesse, correction dynamique de la hauteur et modulations expressives

La nature continue de la tablette graphique apporte une difficulté dans la tâche de visée d'une note. Cet interface musicale étant nouvelle, il a fallu mettre en place des techniques de jeu particulières qui y soient adaptées. Le Chorus Digitalis est un ensemble de musiciens jouant principalement avec le Cantor Digitalis depuis plusieurs années (nous présenterons certaines de leurs représentations dans le Chapitre 6). Leur expérience a permis de mettre en place des techniques de jeu pour le contrôle de la hauteur avec un tablette graphique. Comme le montre la Figure 4.1, les notes sont disposées sur l'axe horizontal de la tablette graphique. La visée d'une note consiste donc à viser la ligne verticale correspondante. Pour passer d'une note à l'autre, un musicien non-entraîné aura tendance à tirer un trait droit avec le stylet entre ces deux notes. Avec cette solution, plus l'écart sera grand et le mouvement rapide, plus les erreurs de précisions dans la visée seront amplifiées. La Figure 4.2 montre les déplacements du stylet sur une tablette graphique lors de la production d'une mélodie simple par un musicien entraîné par sa participation au Chorus Digitalis. Les transitions entre deux notes sont effectuées par des tracés en arc de cercle. Ainsi, la visée d'une ligne verticale s'effectue avec un déplacement vertical du stylet à l'arrivée, ce qui permet un jeu rapide dépourvu de dépassement de la note visée.

Bien que le repère visuel sur la tablette graphique soit une aide importante à la précision du contrôle mélodique, il peut arriver au musicien d'avoir à détourner son regard (lecture de partition, communication avec d'autres musiciens...) Dans ce cas, les erreurs de pointage peuvent être accentuées. Pour palier ce problème, [Perrotin & d'Alessandro 2013] ont développé une méthode de déformation de hauteur dynamique, qui permet d'améliorer significativement la justesse et la précision des attaques et des notes tenues, sans altérer les modulations expressives du musicien telles que le vibrato ou le glissando.

En effet, de telles modulations de hauteur correspondent à des paramètres de contrôle expressif du chant, et permettent d'améliorer le naturel du chant de syn-

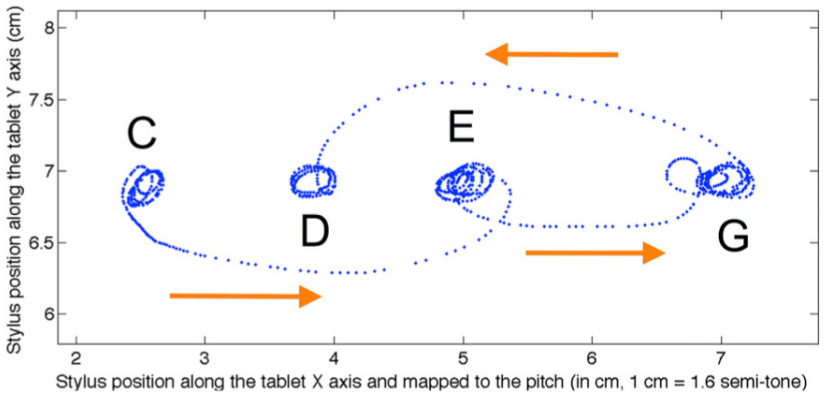


FIGURE 4.2 – *Position du stylet sur la tablette graphique lors de transitions de notes et de vibratos. Les flèches oranges représentent le temps. Figure issue de* [Feugère *et al.* 2017].

thèse. Plutôt que de contrôler des paramètres tels que l'amplitude et la fréquence du vibrato, comme dans le SPASM (section 1.2.3), ou encore d'avoir à régler à l'avance la durée des glissandi, la tablette graphique permet de contrôler ces modulations directement par le geste manuel. La FIGURE 4.2 montre le déplacement du stylet pour une mélodie simple. Le musicien produit un vibrato en dessinant des cercles autour de la note jouée. Le geste manuel permet non seulement de contrôler l'amplitude et la fréquence du vibrato, mais également l'enveloppe temporelle de ces paramètres. De même, la durée du glissando peut être simplement modulée par la durée du geste qui lie une note à la suivante. La tablette graphique permet de contrôler les variations mélodiques et expressives de la hauteur avec une seule modalité de contrôle, alors que deux mains sont nécessaires dans le cas d'un clavier MIDI (hauteur sur le clavier, amplitude du vibrato avec la roue de modulation, fréquence du vibrato préréglée). De plus, le clavier MIDI ne permet pas de contrôler la durée des glissandi, qui doit être réglée à l'avance. Les interfaces continues offrent un contrôle expressif plus intuitif que les discrètes, et sont donc bien mieux adaptées au contrôle de la synthèse vocale.

### 4.1.4 Rôle des modalités

Le stylet d'une tablette graphique est une interface qui offre de nombreuses dimensions de contrôle. En effet, il est bien sûr possible de détecter sa position $(x, y)$ sur la tablette graphique, mais également la pression qui y est appliquée, ainsi que son inclinaison sur les axes $x$ et $y$. De plus, le stylet possède deux boutons dont l'état peut être détecté. Tous les paramètres que nous venons de citer sont également détectables lorsque le stylet n'est pas en contact avec la tablette, jusqu'à une distance de quelques centimètres.

Le Cantor Digitalis étant un synthétiseur reconnu (premier prix du concours Guthman d'instruments de musique), nous avons décidé de conserver les stratégies

de contrôle de la hauteur et de l'effort vocal qui soient pertinentes pour notre application. Ainsi, le contrôle de la hauteur se fera de façon identique sur l'axe $x$ de la tablette graphique, et l'effort vocal sera contrôlé par la pression appliquée au stylet. Les autres modalités pourront alors librement être assignées à tel ou tel paramètre vocal, ce que nous verrons plus tard dans cette section. Le Cantor Digitalis faisait également usage de la fonction tactile de la tablette graphique pour le contrôle des voyelles. Notre application de modification de signaux pré-enregistrés libère alors la fonction tactile de la tablette : la prononciation est programmée à l'avance. Si la méthode de contrôle temporel ou rythmique choisie ne fait pas usage de la main libre, celle-ci peut alors être utilisée pour contrôler d'autres paramètres vocaux, tels que la hauteur d'une seconde voix de synthèse. Les paramètres disponibles sont alors la position $(x, y)$ du doigt sur la tablette, ainsi que sa surface de contact.

Dans les sections suivantes, nous présenterons nos contributions dans le domaine du contrôle chironomique de la hauteur et d'autres paramètres de qualité vocale.

### 4.1.5 Polyphonie

L'un des principaux objectifs qui ont guidé le développement de Vokinesis était de permettre à des interprètes de surpasser les possibilités de production vocale solitaire. La faculté de jouer des polyphonies permettrait à un seul musicien de produire un chœur de voix chantée. Vokinesis permet ainsi de contrôler jusqu'à deux voix de synthèse en simultané. La méthode que nous avons mise en place consiste à utiliser la position d'un doigt en contact avec la tablette graphique pour contrôler la hauteur d'une seconde voix de synthèse. Par défaut, notre système permet de contrôler l'effort vocal de la seconde voix avec la pression du stylet, auquel cas les deux voix auront le même effort vocal au même moment. Comme nous l'avons vu dans la section 4.1.2, les repères visuels sur la tablette graphique sont d'une grande aide pour un contrôle juste et précis de la hauteur mélodique. Or, le fait de contrôler deux hauteurs simultanées implique une nécessité de regarder à deux endroits en alternance, ce qui complique évidemment le contrôle mélodique.

Ce mode de contrôle a été testé lors de représentations publiques par deux musiciens. L'un d'entre eux utilisait une modalité de contrôle rythmique pédestre qui lui laissait la deuxième main libre (mode *Fader* contrôlé par une pédale, voir la section 3.3), l'autre une modalité manuelle qui ne lui offrait pas cette possibilité (mode *Tap*). Le premier contrôlait donc la note de la seconde voix avec l'index de sa seconde main (nous l'appellerons ici l'interprète *bi-manuel*), et le deuxième avec l'auriculaire de la main qui tient le stylet (l'interprète *mono-manuel*). Les deux stratégies ont leurs avantages et leurs inconvénients. Pour planifier et corriger les mouvements de hauteur de sa seconde voix, l'interprète mono-manuel disposait, en plus des indices audio-visuels, d'indices kinesthésiques importants correspondant à l'écart entre la position de son stylet et son auriculaire. Cependant, il avait moins de liberté de contrôle de sa seconde voix, car la plage de contrôle de son auriculaire dépendait évidement de la position de son stylet. Quoi qu'il en soit, avec un peu d'entraînement, les deux étaient capables de jouer des polyphonies justes.

### 4.1.6 Yodel

Le yodel est un style de chant particulier qui consiste à effectuer des transitions brusques d'un mécanisme laryngé à l'autre, avec une volonté de marquer l'instant de transition entre les deux mécanismes [Wise *et al.* 2007]. C'est une technique vocale répandue dans le monde. En voici quelques exemples : une version européenne chantée par Franzl Lang (Allemagne) [1], une autre chantée par Melanie Oesch (Suisse) [2], une version américaine chantée par Wanda Jackson (États-Unis) [3], une version africaine chantée par des femmes du peuple Baka (Cameroun et Gabon) [4], ou bien de façon plus rare lors des notes finales de certaines phrases de cette chanson des Cramberies (Irlande) [5]. Le passage d'un mécanisme laryngé à l'autre se traduit souvent par des variations de hauteur de l'ordre de l'octave. Or, la distance à parcourir avec le stylet sur la tablette graphique est trop grande pour produire l'effet brutal du changement de mécanisme du yodel.

L'une des configurations de Vokinesis permet d'assigner l'octave du signal de synthèse à la position du stylet sur l'axe vertical de la tablette graphique, comme l'illustre la FIGURE 4.3. C'est un contrôle discret : passer à l'octave supérieure double la fréquence de synthèse, mais il n'y a pas d'interpolation entre deux octaves. Cela permet une transition instantanée d'une octave à l'autre sans avoir à parcourir les 12 demi-tons correspondants sur l'axe horizontal de la tablette, et de simuler ainsi une transition brusque qui rappel le yodel. Notez que nous n'effectuons pas de modification particulière du signal pour donner une impression de changement de mécanisme laryngé, et il serait intéressant de vérifier si l'implémentation d'un tel effet améliorerait le naturel de la synthèse.

### 4.1.7 Taille du conduit vocal et tension vocale

La taille du conduit vocal et la tension vocale peuvent être ajustées à l'avance à partir de la fenêtre principale de Vokinesis (voir la section 5.5.2), mais également contrôlées en temps-réel par l'interface qui leur aura été assignée (voir la section 5.5.3).

Pour une synthèse vocale qui se rapproche du naturel, la modification de la taille du conduit vocal a plus vocation a être effectuée au préalable qu'en temps-réel. Nous vous présenterons des exemples sonores de modification temps-réel de la taille du conduit vocal dans la section 6.2, *Au delà du chant*. La modification préalable du conduit vocal permet de modifier le timbre des voix de synthèse, en transformant par exemple une voix d'homme en voix de femme, une voix d'adulte en voix d'enfant, ou encore de minuscule ou de géant dans le cas de modifications extrêmes.

Le cas de la tension vocale est légèrement différent. En effet, celle-ci pourrait être considérée comme une variation prosodique suprasegmentale plus lente encore

1. https://www.youtube.com/watch?v=vQhqikWnQCU
2. https://www.youtube.com/watch?v=AWhMLfnlYIc
3. https://www.youtube.com/watch?v=fTxVdsjOX_U
4. https://www.youtube.com/watch?v=cATZe_jlc9g
5. https://www.youtube.com/watch?v=6Ejga4kJUts

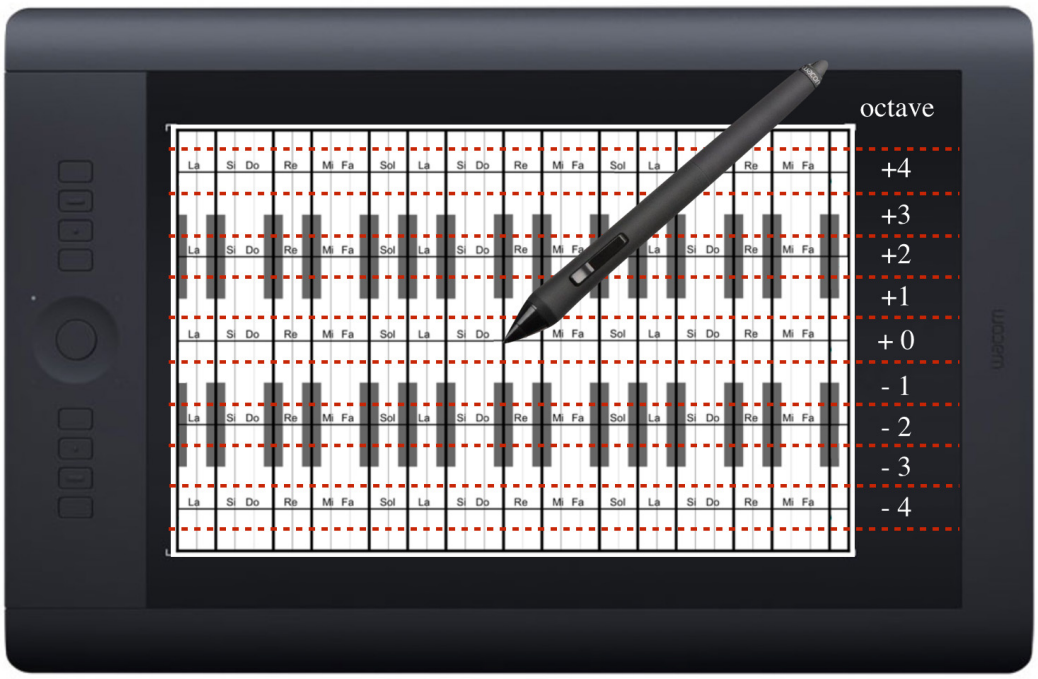


Figure 4.3 – *Contrôle de l'octave sur l'axe vertical de la tablette.*

que l'intonation ou la mélodie. Un chanteur pourra décider de serrer plus ou moins la voix pour exprimer différents sentiments. En l'assignant par exemple à la position $y$ du stylet, ou à son inclinaison, les interprètes pourront alors augmenter le pouvoir expressif de leurs gestes de contrôle.

## 4.2 Claviers et contrôleurs MIDI

Nous avons vu dans la section précédente que la tablette graphique était une interface puissante pour le contrôle mélodique et intonatif de la hauteur vocale, avec un nombre de degrés de liberté qui permet un contrôle précis, intuitif et expressif. Cependant, ces interfaces sont plutôt destinées à des applications graphiques, et ne sont pas extrêmement répandues chez les musiciens. Nous avons donc souhaité explorer les interfaces dédiées à la musique pour le contrôle de la mélodie et des modulations expressives associées. Dans cette section, nous présenterons d'abord le fonctionnement du protocole MIDI, une méthode universelle de communication entre différents appareils numériques dédiés à la musique. Nous verrons comment les synthétiseurs interprètent les données MIDI, et les stratégies qui sont mises en place pour transformer ces signaux de contrôle discrets en signaux sonores qui varient dans le temps. Nous pourrons alors expliquer les différentes stratégies que nous avons mises en place pour permettre un contrôle expressif de la voix avec les possibilités offertes par le protocole MIDI.

### 4.2.1 Interfaces et protocole MIDI

Le protocole MIDI (*Musical Instrument Digital Interface*) [IMA 1983] est une méthode de communication entre différents appareils numériques dédiés à la musique. C'est un protocole dont le langage peut être reconnu par n'importe-quel équipement musical qui en est doté. Tout d'abord, il faut différencier les claviers MIDI des contrôleurs MIDI. Les claviers MIDI sont organisés à la manière d'un piano (en haut de la FIGURE 4.4). Ils servent à émettre des notes MIDI à un synthétiseur. Ils peuvent être équipés de contrôleurs supplémentaires tels que des potentiomètres ou des boutons. Une action d'un contrôleur envoie un message MIDI *control change* (cc) [6] pour transmettre une valeur numérique, codée sur 8 ou 14 bits. Le clavier en haut de la figure est équipé de deux mollettes noires (à gauche du clavier), nommées *pitch bend* et *modulation wheel*. Ce sont deux cc standards du protocole MIDI, permettant d'effectuer des modifications continues de la hauteur (*pitch bend*), ou d'un ou plusieurs paramètres de synthèse prédéfinis (*modulation wheel*). Le synthétiseur interprète ces messages selon ses réglages, paramétrés par l'utilisateur (nous y reviendrons plus tard). L'interface en bas de la figure ne possède que des cc (potentiomètres rotatifs et linéaires, ainsi que des boutons), et pas de clavier.

Plusieurs appareils MIDI peuvent être connectés au même terminal. Par exemple, un ordinateur peut comporter plusieurs synthétiseurs, qui seront contrôlés par différentes interfaces MIDI. Pour indiquer au terminal la façon dont l'installation doit fonctionner, chaque interface MIDI possède son propre nom. Il est alors possible de sélectionner l'*interface 1* pour le *synthétiseur 1*, et l'*interface 2* pour le *synthétiseur 2*. Ensuite, chaque interface possède 16 canaux MIDI. Si l'*interface 1* émet sur le canal 5, alors le *synthétiseur 1* devra écouter les messages du canal 5. Enfin, chaque canal peut émettre les données relatives aux notes, et les données relatives à 121 cc. Lorsqu'une note est jouée, le message émis par le contrôleur est le suivant : [note on, valeur de la note, vélocité] (la vélocité correspond à la force de frappe). Lorsqu'une note est relâchée, le message émis est le suivant : [note off, valeur de la note, vélocité de relâchement]. Le synthétiseur saura alors que la note en question ne doit plus retentir. Lorsqu'un cc est activé, le message émis est le suivant : [numéro du cc, valeur du cc]. Ainsi, si le cc 10 est assigné au volume, alors à chaque fois que le potentiomètre correspondant au cc10 sera activé, le volume du synthétiseur sera modifié. Certains claviers MIDI possèdent également l'*aftertouch*. C'est une donnée MIDI correspondant à la pression appliquée à une touche du clavier lors du maintien d'une note. Toutes les notes maintenues sur un même canal MIDI sont affectées par l'aftertouch.

Outre les notes jouées, les sons émis par le synthétiseur pourront varier selon les paramètres suivants : la vélocité, l'aftertouch, les valeurs des cc, les enveloppes et les LFO (nous définirons ces deux derniers dans les sections suivantes). Chacun de ces paramètres de contrôle pourra être assigné à un paramètre particulier d'un synthétiseur. Les enveloppes et les LFO sont alors très utiles pour assigner à un son synthétique une variation temporelle qui imite celle d'un son naturel.

6. Nous utiliserons « cc » pour désigner un contrôleur MIDI


Figure 4.4 – *Clavier MIDI « M-Audio Keystation 61es » équipé de quelques contrôleurs (en haut) et interface MIDI « Akai APC40 MKII » uniquement équipée de contrôleurs (en bas). Images issues de* https://fr.audiofanzine.com/clavier-maitre-midi-61-touches/m-audio/Keystation-61es/ *et* https://www.energyson.fr/akai-apc40-mk2-controleurs-midi_p3895.htm

### 4.2.2 Enveloppes et LFO (Low Frequency Oscillators)

Lorsqu'une note est jouée, une enveloppe est déclenchée, et la façon dont elle varie dans le temps peut être prédéfinie. La Figure 4.5 montre les paramètres d'enveloppe standards qui peuvent être réglés dans un synthétiseur. Nous garderons ici les termes anglais largement utilisés de façon internationale. Imaginons que l'enveloppe ait été assignée au volume d'un synthétiseur. L'*attack* correspond au temps qu'il prendra à atteindre son volume maximal une fois la note jouée. Le *decay* indique au synthétiseur la durée que devra prendre la transition entre la valeur de volume maximale et celle indiquée par le *sustain*. Tant que la touche du clavier est maintenue, la valeur du volume restera égale à celle du *sustain*. Enfin, une fois que la touche est relâchée, la durée que prendra la volume à atteindre sa valeur nulle est indiquée par le *release*.

Un LFO est un oscillateur à basse fréquence (un signal périodique), présent dans

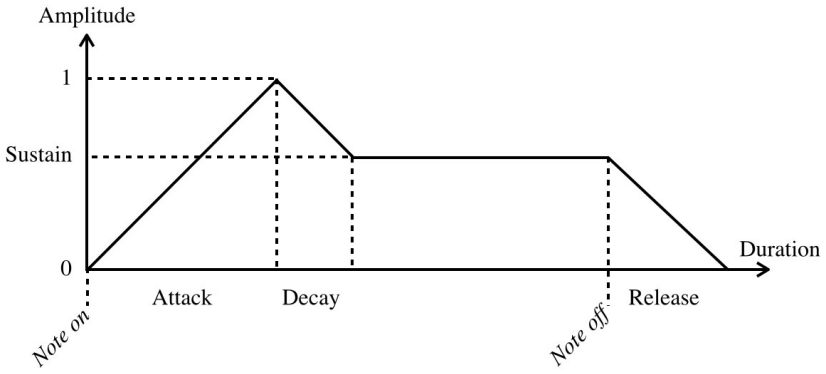


FIGURE 4.5 – *Paramètres standards des enveloppes temporelles d'un synthétiseur.*

la plupart des synthétiseurs, dont la forme d'onde et la fréquence peut être définie, et qui permet de moduler un paramètre sonore auquel il sera assigné. Par exemple, si un LFO est assigné à la fréquence fondamentale, celle-ci oscillera autour de la valeur émise par la note jouée sur le clavier selon les paramètres d'amplitude et de fréquence du LFO. C'est cette technique que nous utiliserons pour permettre aux interprètes de produire un vibrato. Nous utiliserons alors des LFO sinusoïdaux, et nous verrons plus bas comment en contrôler la fréquence et l'amplitude.

### 4.2.3 Contrôle de la hauteur vocale et modulations expressives avec un clavier MIDI

Le contrôle de la hauteur avec un clavier MIDI est par nature discret. Cependant, il existe des stratégies pour transformer les commandes discrètes du clavier en signaux de commande continus : le portamento pour les changements de notes, les LFO pour le vibrato, et les enveloppes pour l'intensité vocale.

Dans notre système, le contrôle de la hauteur avec un clavier MIDI fonctionne selon les stratégies de contrôle des synthétiseurs monophoniques. Celles-ci ne permettent pas de jouer des accords, mais permettent de produire des transitions continues lors du passage d'une note à la suivante (portamento). Un appui sur une touche du clavier émet la fréquence de la note correspondante au synthétiseur. Si une deuxième note est jouée lorsque la première est maintenue, le portamento est activé : la valeur de fréquence fondamentale émise au synthétiseur évoluera entre la fréquence de la première note et celle de la seconde à une vitesse prédéfinie. Le portamento permet donc d'éviter des transitions trop abruptes entre deux notes successives. Seulement, sa durée est réglée à l'avance (voir la section 5.5.3.2), et ne peut pas être contrôlée en temps-réel, ce qui enlève un degré de liberté en terme d'expressivité, et donc de naturel. Une autre manière d'effectuer un portamento consiste à utiliser la *Pitch Bend Wheel* (ou roue de courbure de la hauteur), présente sur la plupart des claviers MIDI. C'est un contrôleur continu de type potentiomètre linéaire équipé de ressorts et dont la position de repos est au centre. Elle permet de modifier la hauteur du signal de synthèse de manière continue sans avoir à appuyer

sur une autre touche du clavier. Les durées des glissandi sont directement contrôlées, mais cette méthode implique l'utilisation de deux mains pour contrôler les variations d'un seul paramètre.

Le vibrato relève du contrôle d'un LFO. La fréquence du LFO est généralement réglée à l'avance, et son amplitude contrôlée par la *Modulation Wheel* (ou roue de modulation), un contrôleur continu de type potentiomètre linéaire, qui peut posséder un ressort ou non. Cependant, si l'utilisateur est équipé d'une pédale d'expression, il peut lui assigner le contrôle de la fréquence du LFO. En assignant un LFO à la hauteur, celle-ci oscillera autour de la note jouée sur la clavier, avec une fréquence et une amplitude gérées par les contrôleurs continus que nous venons d'évoquer. Encore une fois ce qui pouvait être contrôlé avec un seul membre sur la tablette graphique en nécessite ici trois.

L'effort vocal est contrôlé par la vélocité d'une note jouée. Sa variation est régie par une enveloppe, dont les durées d'*attack* et de *release* peuvent être préalablement réglées. Le *sustain* est maintenu à 1, et seul l'aftertouch pourra modifier l'effort vocal lors du maintien d'une note. Lors d'un portamento, une interpolation linéaire entre les vélocités de la première et de la seconde note jouées est effectuée, avec une durée identique à celle définie pour la hauteur.

Comme nous le verrons dans la section 5.5.3.3, il est possible de définir une note de séparation sur le clavier MIDI afin de permettre la production de polyphonies. Ainsi, toutes les notes inférieures à la note de séparation contrôleront la voix de synthèse principale, et les autres la voix secondaire. Chaque partie du clavier conservera les stratégies de contrôle monophonique que nous venons de présenter. Ce mode de contrôle peut alors être considéré comme une mode *bi-monophonique*. Un exemple de contrôle bi-monophonique effectué au clavier MIDI par Christophe d'Alessandro est présenté dans la vidéo `Ex08`. Pour cet exemple, le clavier MIDI a été configuré de telle sorte que chaque main contrôle une voix de synthèse différente, aussi bien au niveau du cadre rythmique (mode *Tap*) que de la mélodie et de l'effort vocal. Le musicien a près de 50 ans de pratique d'instruments à clavier (piano, orgue, clavicorde), contre une seule minute de pratique au clavier sur Vokinesis.

## 4.3 Polyphonic Multidimensional Controllers (PMC)

### 4.3.1 Interfaces PMC et méthode MPE

Les PMC (*Polyphonic Multidimensional Controller*) sont de nouvelles interfaces de contrôle dédiées à la musique. Elles font usage de la méthode MPE (*Multidimensional Polyhonic Expression*), une nouvelle façon d'utiliser le protocole MIDI (plus de détails seront donnés plus bas). Elles permettent un contrôle polyphonique, tout comme le clavier MIDI : plusieurs notes peuvent être jouées de façon simultanée. La nouveauté vient de la possibilité de contrôler une multitude de paramètres de façon continue et indépendante pour chaque note : le son correspondant à une note jouée avec un doigt peut être modulé en déplaçant ce même doigt sur l'axe horizontal ou vertical de la surface de contrôle, ou encore en y appliquant des variations de

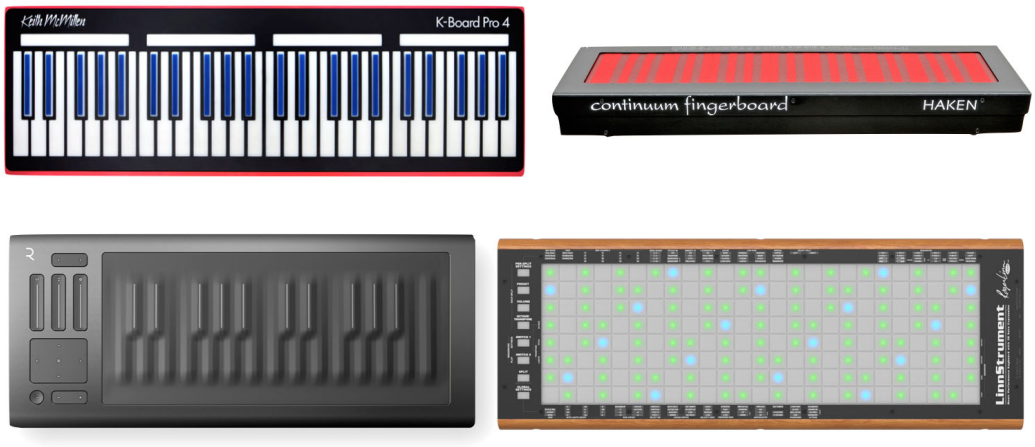


FIGURE 4.6 – *Exemples de PMC. Dans l'ordre de lecture : K-Board Pro 4 (Keith McMillen Instruments), Continuum Fingerboard (Haken), Seaborad (Roli), LinnStrument (Roger Linn Design). Tous ces contrôleurs sont sensibles à la pression et à la position* $(x, y)$ *de chaque doigt, et émettent des messages MIDI selon la méthode MPE. L'échelle n'est pas respectée.*

pression. Si plusieurs doigts jouent plusieurs notes au même instant, les modulations effectuées seront indépendantes pour chaque doigt.

La FIGURE 4.6 présente quatre PMC du marché. Nous les avons sélectionnés pour leur notoriété et leurs différences respectives, afin de présenter les larges possibilités de conception matérielle offertes par la méthode MPE. Toutes ces interfaces possèdent des surfaces de contrôle en 5 dimensions : elles sont sensibles aux vélocité d'attaque et de relâchement de chaque doigt, ainsi qu'à leur pression et à leur position $(x, y)$.

En haut à gauche de la figure se trouve le *K-Board Pro 4*. Il conserve le principe de fonctionnement d'un clavier MIDI : chaque touche est indépendante, et fournit des données de vélocité lors de l'appui et du relâchement. Cependant, les capteurs en tissus ajoutés sur chacune des touches permettent un contrôle du vibrato par des mouvements horizontaux, du volume par des variations de pression, et d'un autre paramètre assigné à la position du doigt sur l'axe vertical de la touche. Le *Seaboard* (en bas à gauche de la figure) a également conservé l'organisation d'un clavier. Cependant, sa surface est entièrement continue. Il permet donc un jeu similaire à celui du clavier grâce aux reliefs de la surface, mais également similaire à celui de la tablette graphique grâce aux espaces plats présents dans les parties inférieure et supérieure de la surface. Les reliefs de la surface apportent donc un retour tactile que ne possède pas la tablette graphique. Le *continuum fingerboard* possède quant-à lui une surface tout à fait plate, qui n'apporte donc pas de retour sensoriel supplémentaire : le visuel y jouera donc un rôle très important. Le *LinnStrument* possède des rangées de touches : chaque colonne correspond à un espacement d'un demi-ton, et chaque ligne peut être organisée à la guise d'un utilisateur. L'organisation des notes pourra alors

ressembler à l'accordage d'une guitare, d'un violon, etc... En plus de retours visuels offerts par les éclairages de chaque touche, le *LinnStrument* offre un retour tactile : chaque touche est déparée par un creux fin et peu profond. Des exemples d'utilisation de ces interfaces peuvent être visualisés aux adresses suivantes : *K-Board Pro 4*[7], *Continuum fingerboard*[8], *Seaboard*[9], *LinnStrument*[10]

L'indépendance du contrôle des paramètres sonores de chaque note est rendu possible grâce à la méthode MPE (*Multidimensional Polyphonic Expression*), qui fait usage du protocole MIDI, selon une technique particulière basée sur l'utilisation de plusieurs canaux. Un canal global est utilisé pour transmettre les messages MIDI concernant des modifications qui devront être appliquées à toutes les notes, par exemple la position d'une pédale d'expression. Chaque canal restant est utilisé pour transmettre les messages qui correspondent à une note jouée, mais également aux modifications expressives qui pourront y être appliquées (déplacements d'un doigt sur les deux axes et variations de pression). Le protocole MIDI faisant usage de 16 canaux, la méthode MPE permet de jouer jusqu'à 15 notes de façon simultanée et indépendante.

### 4.3.2 Contrôle de la hauteur vocale et modulations expressives avec un PMC

La diffusion *open source* du Cantor Digitalis a permis à des utilisateurs tiers de rendre ce synthétiseur compatible avec plusieurs PMC : le *Continuum fingerboard*[11], le *Soundplane*[12] et le *Seaboard*[13]. Le contrôle continu en trois dimension qu'offrent les PMC réduit l'importance des enveloppes et des LFO pour une synthèse naturelle et expressive. Les variations de pression permettent un contrôle manuel de l'enveloppe du volume, et de légers mouvements horizontaux des doigts créent des effets de vibrato qui peuvent remplacer les LFO. La durée des portamenti est également contrôlée directement par le déplacement d'un doigt. Les modulations expressives liées à des changements de qualité vocale peuvent quant-à elles être effectuées par des déplacements sur l'axe vertical.

De la même manière que pour les claviers MIDI, afin de permettre le jeu polyphonique, nous avons mis en place une note de séparation qui permet d'assigner une partie d'un PMC à une première voix de synthèse, et une autre partie à une seconde (voir la section 4.2.3).

7. https://youtu.be/q6F-_m9g6O8
8. https://youtu.be/OccKkhHrMOs
9. https://youtu.be/Oy4Y5xPAC7A
10. https://youtu.be/flKMz9UmqAw
11. https://youtu.be/R2XRfhu95Dc
12. https://youtu.be/oVQMHX4bQuo
13. https://youtu.be/Xtx8WcKBQbk

## 4.4 Comparaison des interfaces pour le contrôle de la mélodie

Nous allons à présent comparer les différentes interfaces de contrôle de la hauteur que nous avons évoquées jusqu'ici, et nous de soulignerons les avantages et les inconvénients de chacune dans le cadre d'un contrôle expressif du chant de synthèse. Le Tableau 4.1 fournit une comparaison des tablettes graphiques, des claviers MIDI et des PMC . Pour les PMC, nous considérerons des interfaces telles que le *Seaboard* ou le *LinnStrument*, qui offrent des surfaces continues similaires à celle de la tablette graphique, mais dont les reliefs offrent en plus un retour tactile important.

### 4.4.1 Mélodies monophoniques

Tout d'abord, nous avons déjà vu que des débutants à la tablette graphique étaient en mesure de jouer des mélodies avec une meilleure précision qu'avec leur propre voix [d'Alessandro *et al.* 2014]. Ainsi, tant que le regard des interprètes reste concentré sur la tablette, le seul retour audiovisuel qu'elle offre, et l'absence de retours tactiles et kinesthésiques par rapport aux claviers MIDI et PMC considérés ne semblent pas être un frein à la précision du contrôle monophonique de la mélodie. Par ailleurs, il serait intéressant de comparer les durées d'apprentissage d'une mélodie complexe par des débutants de chacune de ces interfaces. Nous supposons en effet que la maîtrise experte quasi-universelle du stylo donnerait un certain avantage au stylet de la tablette graphique.

Pour des variations mélodiques rapides de type yodel qui sont de l'ordre de l'octave, les claviers MIDI et les PMC semblent avantageux. En effet, l'usage de la main permet le passage d'une note à l'autre par alternance des doigts, alors que l'unique point de contrôle que fournit le stylet implique une durée de transition minimale entre deux notes, liée à la vitesse maximale du mouvement de la main. Cependant, l'utilisation de l'axe $y$ de la tablette graphique pour le contrôle de l'octave permet de corriger ce défaut. Nous avons tout de même jugé cette méthode comme désavantageuse dans le Tableau 4.1 car elle implique un contrôle de la hauteur sur deux axes différents (variations lentes sur l'axe $x$ et variations rapides sur l'axe $y$).

### 4.4.2 Modulations expressives

Dans cette section, les modulations expressives du Tableau 4.1 considérées sont les suivantes : yodel, vibrato, glissando, effort vocal et tension vocale. Toutes ces variations étant liées à la phonation (hauteur et qualité vocales), leur contrôle simultané sera plus intuitif s'il est effectué par un seul et même membre.

Dans le cas de la **tablette graphique**, la main qui tient le stylet pourra respecter cette condition, selon quatre modalités (entre crochets) :

- [axe $x$] note, amplitude et fréquence du vibrato, glissando
- [axe $y$] octave (yodel)
- [pression] effort vocal

Tableau 4.1 – *Comparaison des capacités de la tablette graphique, des claviers MIDI et des PMC pour le contrôle de la hauteur et de la qualité vocale.*

| | **Tablette** | **Claviers MIDI** | **PMC** |
|---|---|---|---|
| **Mélodies monophoniques** | Informations audiovisuelles uniquement (suffisantes)<br>+<br>Contrôle continu<br>+ | Informations audiovisuelles, tactiles et kinesthésiques<br>++<br>Contrôle discret<br>- | Informations audiovisuelles, tactiles et kinesthésiques<br>++<br>Contrôle continu<br>+ |
| **Yodel (octave)** | Axe $y$<br>- | Axe $x$<br>+ | Axe $x$<br>+ |
| **Vibrato (amplitude et fréquence)** | Contrôle direct<br>+ | Potentiomètres<br>- | Contrôle direct<br>+ |
| **Glissando** | Contrôle direct<br>+ | Pré-réglages<br>- | Contrôle direct<br>+ |
| **Effort Vocal** | Pression<br>+ | Aftertouch<br>+ | Pression<br>+ |
| **Tension Vocale** | Axe $y$ ou inclinaison<br>+ | Potentiomètre<br>- | Axe $y$<br>+ |
| Nombre minimal de membres nécessaires | 1<br>+ | 4<br>- | 1<br>+ |
| Nombre minimal de modalités nécessaires | 4<br>+- | 5<br>- | 3<br>+ |
| **Mélodies polyphoniques** | Informations visuelles<br>-<br>Modulations expressives indépendantes<br>+<br>Contrôle expressif différent<br>- | Informations tactiles<br>+<br>Modulations expressives communes<br>- | Informations tactiles<br>+<br>Modulations expressives indépendantes<br>+<br>Contrôle expressif identique<br>+ |

- [inclinaison] tension vocale

Pour les **PMC**, un seule main sera nécessaire, et seulement trois modalités, grâce au fait que l'octave puisse être contrôlée rapidement sur l'axe $x$ pour le yodel :

- [axe $x$] note, amplitude et fréquence du vibrato, glissando, octave (yodel)
- [pression] effort vocal
- [axe $y$] tension vocale

Pour le cas des **claviers MIDI**, quatre membres et cinq modalités seront nécessaires :

- une main : - [axe $x$] note, yodel
  - [pression (vélocité & aftertouch)] effort vocal
- l'autre main : - [*modulation wheel*] amplitude du vibrato
- un pied : - [pédale d'expression] fréquence du vibrato
- l'autre pied : - [pédale d'expression] tension vocale

Selon ces trois listes, le contrôle simultané de toutes les variations expressives considérées semble bien moins intuitif au clavier MIDI, qui nécessite quatre membres, qu'avec des PMC ou des tablettes graphiques, qui n'en nécessitent qu'un seul. Les PMC semblent posséder un léger avantage par rapport aux tablettes graphiques, qui demandent une modalité supplémentaire. Cependant, bien que la pratique du yodel soit répandue géographiquement, elle reste relativement rare dans de nombreux styles de chant et sera sans doute ignorée par certains interprètes. Dans ce cas, les PMC et tablettes graphiques semblent rivaliser : la tension vocale pourra être contrôlée par la position du stylet sur l'axe $y$ de la tablette, et trois modalités seront nécessaires au total pour les deux types d'interfaces.

Néanmoins, la question de la puissance expressive des gestes est intéressante à soulever. En effet, notre pratique des différentes interfaces nous laissent penser que les mouvements d'écriture de la tablette graphique sont plus intuitifs et plus expressifs que les mouvements de toucher des PMC. Il serait intéressant pour de futures recherches de vérifier cette hypothèse de façon formelle.

### 4.4.3 Mélodies Polyphoniques

Nous avons vu dans la section 4.1.5 que la fonction tactile de la tablette graphique permettait de détecter la position de contact d'un doigt pour le contrôle d'une seconde voix de synthèse. La hauteur sera contrôlée par la position du doigt sur l'axe $x$, l'effort vocal par la surface en contact avec la tablette et la tension vocale (ou l'octave pour le yodel) sur l'axe $y$. Le regard jouant un rôle très important dans le contrôle de la hauteur à la tablette graphique, les musiciens qui ont pratiqué ce mode avaient tendance à conserver la note d'une des deux voix fixes, alors que la seconde voix effectuait la mélodie.

Les claviers MIDI et les PMC ont l'avantage d'offrir des retours tactiles, non présents sur la tablette. L'importance amoindrie du regard dans le contrôle de la mé-

lodie permet alors une plus grande indépendance de contrôle des deux voix, comme le montre la vidéo `Ex08`. Par ailleurs, dans le cas des PMC, toutes les modulations expressives sont effectuées exactement de la même manière par chacune des voix, ce qui devrait rendre le contrôle indépendant plus intuitif que dans le cas de la tablette graphique. Dans le cas du clavier MIDI, il sera impossible de contrôler les modulations expressives de façon indépendante pour chacune des deux voix, le nombre de membre faisant défaut.

### 4.4.4 Discussion

Selon les sections précédentes, résumées dans le Tableau 4.1, les claviers MIDI semblent moins adaptés au contrôle expressif du chant de synthèse que les PMC ou les tablettes graphiques. D'une manière générale, les PMC semblent posséder quelques avantages par rapport aux tablettes graphiques, notamment pour ce qui est du contrôle polyphonique. Cependant, si le contrôle rapide de l'octave pour le yodel est ignoré, et si seule la monophonie est considérée, les deux types d'interfaces semblent rivaliser. Nous supposons d'ailleurs que les mouvements d'écriture de la tablette graphique permettent un contrôle plus expressif, et plus simple à apprendre que ceux des PMC, en raison de l'expertise quasi-universelle du maniement du stylo. Certains gestes de contrôle mélodique effectués au stylet seront sans doute difficilement reproduits avec un PMC.

D'autre part, dans le Tableau 4.1, tous les éléments jugés comme négatifs pour la tablette graphique sont liés à la vitesse du contrôle de la hauteur ou encore à la polyphonie. Ces éléments interviennent peu dans le contrôle de l'intonation dans le cas de la parole. En effet, la polyphonie est une pratique exclusivement réservée au chant – le cas de la parole simultanée relève plutôt de la cacophonie. De plus, le contrôle de l'intonation ne devrait pas nécessiter de variations rapides de la hauteur de type yodel.

Nous pensons donc que les PMC sont sans doute mieux adaptées au contrôle polyphonique du chant, mais que les tablettes graphiques sont plus appropriées aux mélodies monophoniques et au contrôle de l'intonation dans le cas de la parole. Par ailleurs, la tablette graphique devrait être plus accessible aux débutants. Quoi qu'il en soit, le choix entre ces deux types d'interfaces devrait dépendre de l'intention musicale des interprètes.

## 4.5 Conclusion

Dans ce chapitre, nous avons présenté différentes stratégies et interfaces de contrôle de la hauteur vocale et des modulations expressives associées. Nous avons établi une comparaison des capacités de trois types d'interfaces : les claviers MIDI, les PMC (*Polyphonic Multidimensionnal Controllers*) et les tablettes graphiques. Cette comparaison, informelle du point de vue scientifique, s'est appuyée sur leurs différences matérielles ainsi que sur notre pratique de chacune. Elle a permis d'ouvrir

une réflexion sur leur utilisabilité dans le domaine du contrôle expressif des paramètres vocaux liés à la phonation. Les claviers MIDI semblent moins adaptés que les deux autres types d'interfaces, en raison de leur nature discrète, mais également à cause du nombre de membres nécessaires augmentant avec le nombre de paramètres vocaux à contrôler. Pour le contrôle d'une mélodie monophonique, les PMC et les tablettes graphiques semblent rivaliser. Les PMC semblent mieux adaptées au contrôle de variations mélodiques rapides de type yodel. Cependant, notre pratique nous laisse penser que les mouvements d'écriture de la tablette graphique ont un pouvoir expressif plus important que les mouvements tactiles des PMC (nous avons d'ailleurs mené une expérience, présentée en Annexe A, qui nous a permis de montrer que le contrôle chironomique de la hauteur permettait d'améliorer l'expressivité de signaux de parole issus d'un système de synthèse HMM expressive.) De plus, nous pensons que la maîtrise experte quasi-universelle du maniement du stylo rend les tablettes graphiques plus rapides à prendre en main pour des débutants. Quoi qu'il en soit, chacune de ces d'interfaces offre différentes possibilités de mouvements, et le choix d'utilisation devrait dépendre des intentions musicales des interprètes. Pour ce qui est du contrôle précis et simultané de deux voix de synthèse, les retours tactiles et kinesthésiques qu'offrent certains PMC constituent un avantage par rapport aux tablettes graphiques, pour lesquelles la vision joue un rôle très important [Perrotin & D'alessandro 2016a].

Chapitre 5

# Vokinesis

## Sommaire



Le développement du logiciel Vokinesis a été central dans ce travail de thèse. Enfant de Calliphony, il a directement hérité des méthodes de modification de la hauteur et de la durée qui y étaient déjà implémentées (voir le Chapitre 3). Nous y avons ajouté les méthodes de contrôle que nous avons présentées dans les Chapitres 3 et 4, ainsi que de la méthode de traitement de signal VoPTiQ présentée dans le Chapitre 2.

Ce travail de développement a suivi plusieurs objectifs. Tout d'abord, nous avons développé des méthodes de configuration matérielle permettant d'optimiser la modularité du logiciel : « *Un outil puissant, c'est un outil qu'on peut détourner, adapter,*

*façonner à sa main* » [Beaudouin-Lafon 2016]. En effet, comme nous avons déjà pu le voir dans les Chapitres 3 et 4, cette modularité nous a permis d'assigner et de tester un très grand nombre d'interfaces pour nos différentes méthodes de contrôle performatif de la voix. Nous avons également souhaité rendre Vokinesis utilisable dans des situations de concert. Pour cela, nous avons mis en place un fonctionnement sous formes de *projets* : plusieurs fichiers audio originaux peuvent être chargés dans un projet et organisés à la guise des interprètes pour faciliter leur sélection lors d'une représentation. Enfin, nous voulions également développer une interface permettant d'utiliser Vokinesis à des fins expérimentales, et permettre ainsi à des chercheurs de concevoir des tâches de production spécifiques.

Dans ce chapitre, nous allons présenter le fonctionnement du logiciel sous différents points de vue. Nous présenterons son fonctionnement général dans un première section. Nous détaillerons ensuite son architecture afin de définir la façon dont les différents éléments qui le composent communiquent les uns avec les autres. La section suivante se concentrera sur le *mapping*, c'est à dire sur la façon dont les données émises par les différentes interfaces de contrôle sont transmises et transformées en paramètres vocaux qui permettront d'indiquer au système comment modifier la voix originale. Enfin, nous présenterons le fonctionnement de Vokinesis du point de vue des utilisateurs, en détaillant les différentes interfaces graphiques du système.

## 5.1 Fonctionnement général

### 5.1.1 Aperçu du système

La Figure 5.1 présente l'architecture générale de Vokinesis. Les flèches vertes indiquent des réglages effectués dans des interfaces graphiques, les flèches oranges des données contrôlées en temps-réel, les flèches rouges des signaux audio et la flèche bleue des données d'analyse.

Les données qui doivent accompagner chaque signal de parole original sont leurs marqueurs périodiques (indiquant au système l'emplacement de chaque période dans les zones voisées) et leur étiquetage phonétique (permettant au système d'une part de placer automatiquement les FCP selon les règles présentées section 3.4.2, et d'autre part de différencier les phases nucléiques et les phases de liaison).

De multiples interfaces de contrôle peuvent être connectées au système. Sur la figure, celles-ci sont représentées par une tablette graphique et deux pédales d'expression. Les données qu'elles émettent sont capturées et transformées en paramètres de contrôle suprasegmental (cadre violet dans la figure) selon les réglages effectués dans l'interface graphique correspondante (cadre vert). La partie *Traitement de signal* (cadre rouge) pourra alors transformer un signal vocal original selon ses données d'analyse (cadre bleu), les données de contrôle gestuel, et les réglages effectués dans les interfaces graphiques. Un fichier audio peut être sélectionné à partir d'une interface graphique, et ses données d'analyse peuvent y être éditées.

Pour résumer, un signal original, accompagné de ses données d'analyse, est modifié en temps-réel selon des paramètres vocaux et temporels, calculés à partir (1)

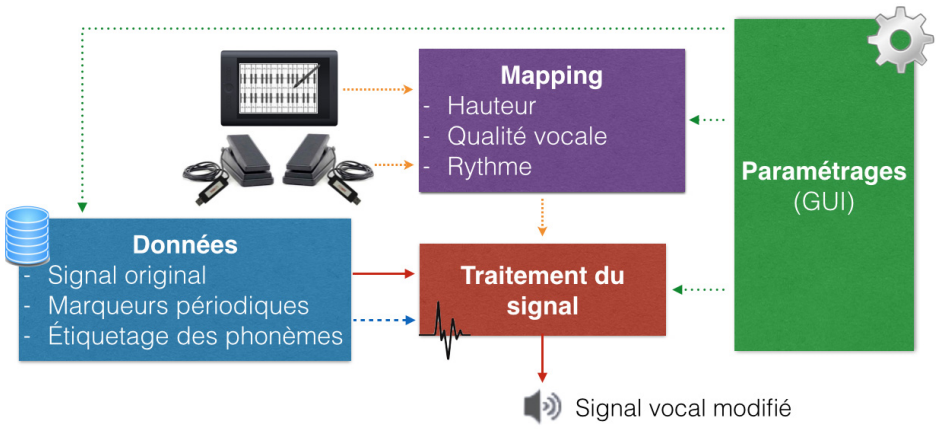


FIGURE 5.1 – *Vue d'ensemble de l'architecture de Vokinesis. Flèches vertes : réglages effectués dans une interface graphique. Oranges : données contrôlées en temps-réel. Bleue : données d'analyse du signal original. Rouges : signaux audio.*

des données émises par les contrôleurs et (2) des réglages effectués dans les interfaces graphiques.

### 5.1.2 Gestion du logiciel

Les deux interfaces graphiques principales de Vokinesis sont présentées FIGURE 5.2. Pour faciliter son usage lors de représentations publiques, nous avons mis en place un fonctionnement sous la forme de *projets* : une fois le logiciel démarré, la première chose à faire est l'ouverture ou la création d'un projet (bouton *Create/-Load project* dans le fenêtre *Project Editor*, à droite dans la figure). Ceci consiste simplement à sélectionner un dossier sur l'ordinateur. Une fois le projet chargé, tous les fichiers audio qu'il contient sont affichés dans le *tableau du projet* (tableau blanc dans la fenêtre *Project Editor*). Une utilisatrice pourra sélectionner l'un des fichiers audio qu'il contient. Ceci fait, le signal correspondant sera affiché dans la fenêtre *Vokinesis* (à gauche sur la figure). L'utilisatrice pourra alors le modifier en temps-réel selon les méthodes que nous avons présentées dans les CHAPITRES 3 et 4.

Un *dossier projet* contient toutes les données relatives à ce projet : les chemins d'accès aux fichiers audio qu'il contient, leurs données d'analyse (marqueurs périodiques et étiquetage des phonèmes), et les différents réglages / paramétrages qui auront été sauvegardés. Au sein d'un projet, un certain nombre de réglages peuvent être effectués et sauvegardés. Certains réglages sont *globaux*, ils concernent tout le projet, d'autres sont *spécifiques*, ils peuvent être différents pour chaque fichier audio. Les réglages globaux concernent le paramétrage des contrôleurs : tel paramètre d'une interface contrôlera tel paramètre vocal (voir la section 5.5.3). Les réglages spécifiques sont... plus spécifiques : mode de contrôle temporel, marquage périodique, étiquetage des phonèmes, réglages des effets audio, etc... (section 5.5.2).

La structure d'un dossier projet est présentée dans la FIGURE 5.3. Le dossier

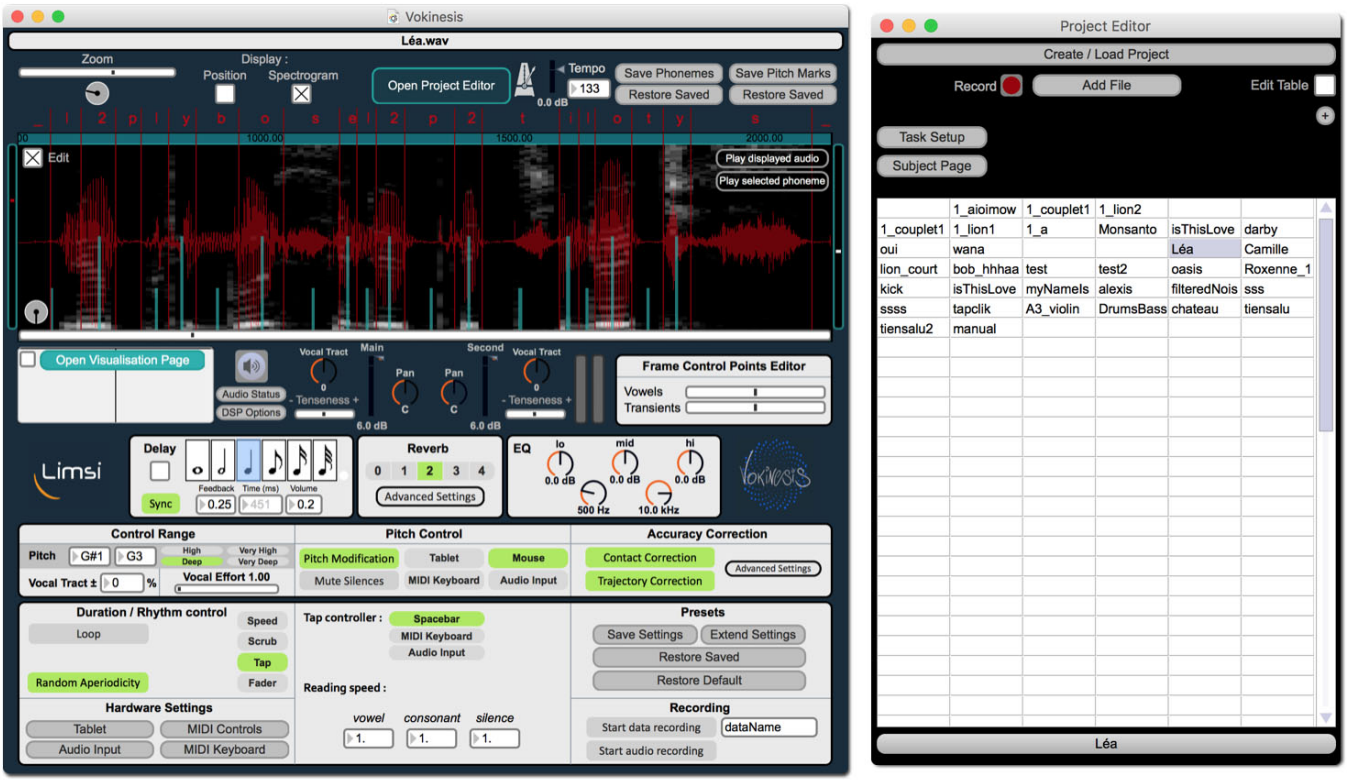


FIGURE 5.2 – *Interfaces graphiques principales de Vokinesis. À gauche : fenêtre principale. À droite : Project Editor*

`data` et le fichier `ProjectMatrix.txt` sont créés automatiquement lors de la création d'un projet. Tant qu'aucun fichier audio n'est ajouté au projet (bouton *Add File* dans le *Project Editor*), ce fichier et ce dossier resteront vides. Lorsqu'un fichier est ajouté, son chemin d'accès et son emplacement dans le tableau du projet sont ajoutés au fichier `ProjectMatrix.txt`. L'exemple ci-dessous représente le fichier `ProjectMatrix.txt` d'un projet auquel auraient été ajoutés 3 fichiers audio, chacun situé à un emplacement différent (chemins 1, 2 et 3), et positionnés dans les cases 1, 2 et 3 du tableau du projet :

(chemin1)/fichier1.wav 1
(chemin2)/fichier2.wav 2
(chemin3)/fichier3.wav 3

De plus, chaque nouveau fichier audio chargé subit une analyse de périodicité, et un fichier `.gci` est alors créé (voir la section 2.2.1) et sauvegardé dans le dossier `data`. Une portion d'un fichier `.gci` est représentée dans la FIGURE 5.4, à gauche. Chaque ligne indique l'emplacement d'un marqueur périodique (en numéro d'échantillon). Les emplacements précédés d'un signe négatif représentent les parties non-voisées du signal original. Par ailleurs, lorsque l'étiquetage des phonèmes est effectué (voir les sections 3.4.4 et 5.5.2.4), les fichiers d'étiquetage des phonèmes `.phon` sont éga-

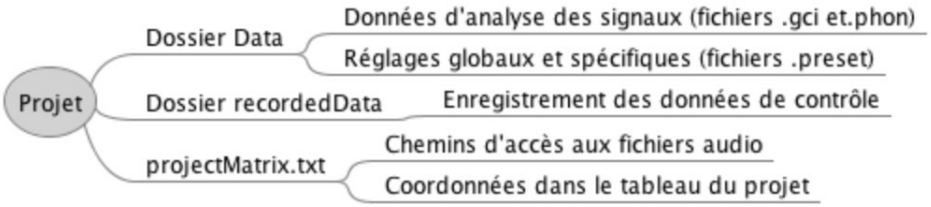


FIGURE 5.3 – *Structure d'un dossier projet.*

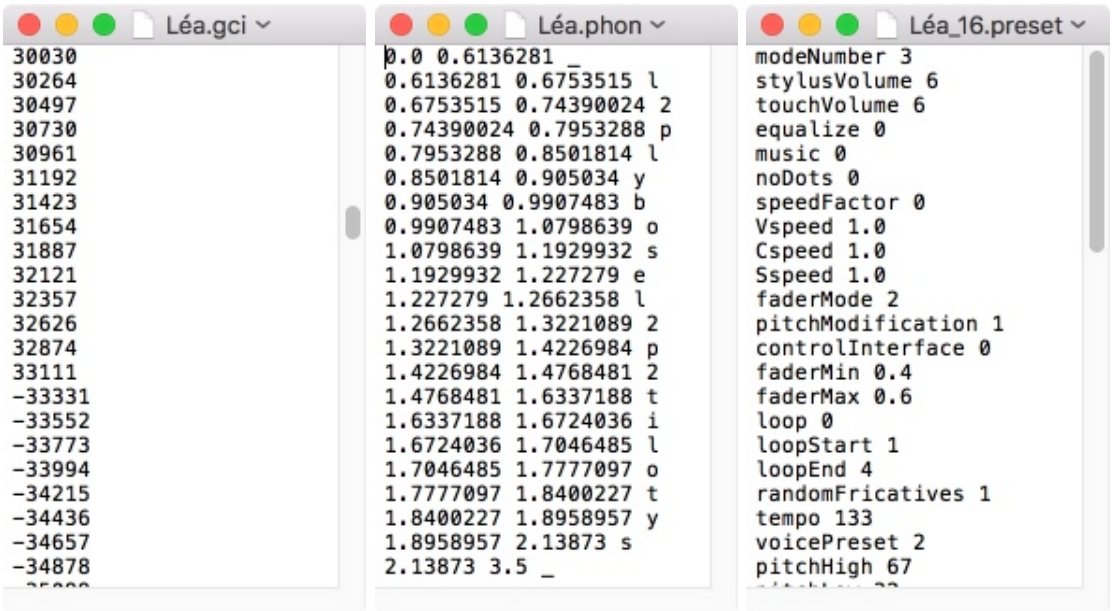


FIGURE 5.4 – *Exemples de fichiers* `.gci` *(à gauche),* `.phon` *(au milieu) et* `.preset` *(à droite).*

lement sauvegardés dans le dossier `data`. Un exemple de fichier `.phon` est représenté au centre de la FIGURE 5.4. De même, lorsqu'un ensemble de réglages spécifiques (volume, effets audio, modes de contrôle, etc...) sont sauvegardés (bouton *Save Settings* dans le cadre *Presets* de la fenêtre *Vokinesis* à gauche de la FIGURE 5.2), un fichier `.preset` est enregistré dans le dossier `Data`. À droite de la FIGURE 5.4 se trouve un exemple de fichier `.preset`. Chaque ligne représente un paramètre réglé dans la fenêtre *Vokinesis* et sa valeur correspondante. Les lecteurs auront sans doute remarqué que le nom des fichiers `.gci` et `.phon` conserve celui du fichier audio original sélectionné dans le tableau du projet présenté FIGURE 5.2. Le nom du fichier `.preset` contient en plus le numéro de la case sélectionnée dans le tableau du projet : cela permet d'utiliser plusieurs fois le même fichier audio dans un seul projet, et de lui assigner différents réglages spécifiques.

Comme nous l'avons déjà évoqué, et comme nous le verrons dans la section 5.5.4, nous avons souhaité permettre à des chercheurs de faire usage de Vokinesis dans un cadre expérimental. Il leur est donc possible d'activer la sauvegarde des données de contrôle lors d'une performance (par exemple la valeur d'un ou plusieurs contrôleurs ou paramètres vocaux au cours du temps). Les données enregistrées seront sauvegardées dans un dossier `recordedData` (voir la FIGURE 5.3).

## 5.2 Architecture

Dans cette section, nous allons présenter les détails de l'architecture de Vokinesis, en nous appuyant sur la FIGURE 5.5. Le code couleur qui y est utilisé est identique à celui de la FIGURE 5.1. Nous avons représenté les différentes interfaces de contrôle dans le cadre orange afin de souligner la multiplicité des interfaces qui peuvent être assignées à Vokinesis.

### 5.2.1 Gestion des fichiers audio

Nous avons déjà eu un aperçu de la fenêtre principale et de l'éditeur de projet dans la FIGURE 5.2. Celles-ci sont représentées par des cadres verts dans la FIGURE 5.5. L'éditeur de projet permet de gérer le contenu d'un projet, c'est à dire les différents fichiers audio qui pourront y être utilisés. Les noms des fichiers d'un projet sont affichés dans le tableau du projet, et leur emplacement peut être réorganisé pour préparer un ordre de sélection pour un concert ou pour une procédure expérimentale. Une modification du tableau (ajout ou déplacement d'un fichier audio) enregistre les données d'organisation dans le fichier `projectMatrix.txt`, qui seront réutilisées pour afficher les différents signaux au bon emplacement lors de l'ouverture d'un projet (double flèche verte *contenu du projet* dans la FIGURE 5.5). L'éditeur de projet permet également d'enregistrer de nouveaux fichiers audio et de les ajouter au projet (flèche rouge *Enregistrement de signaux vocaux*).

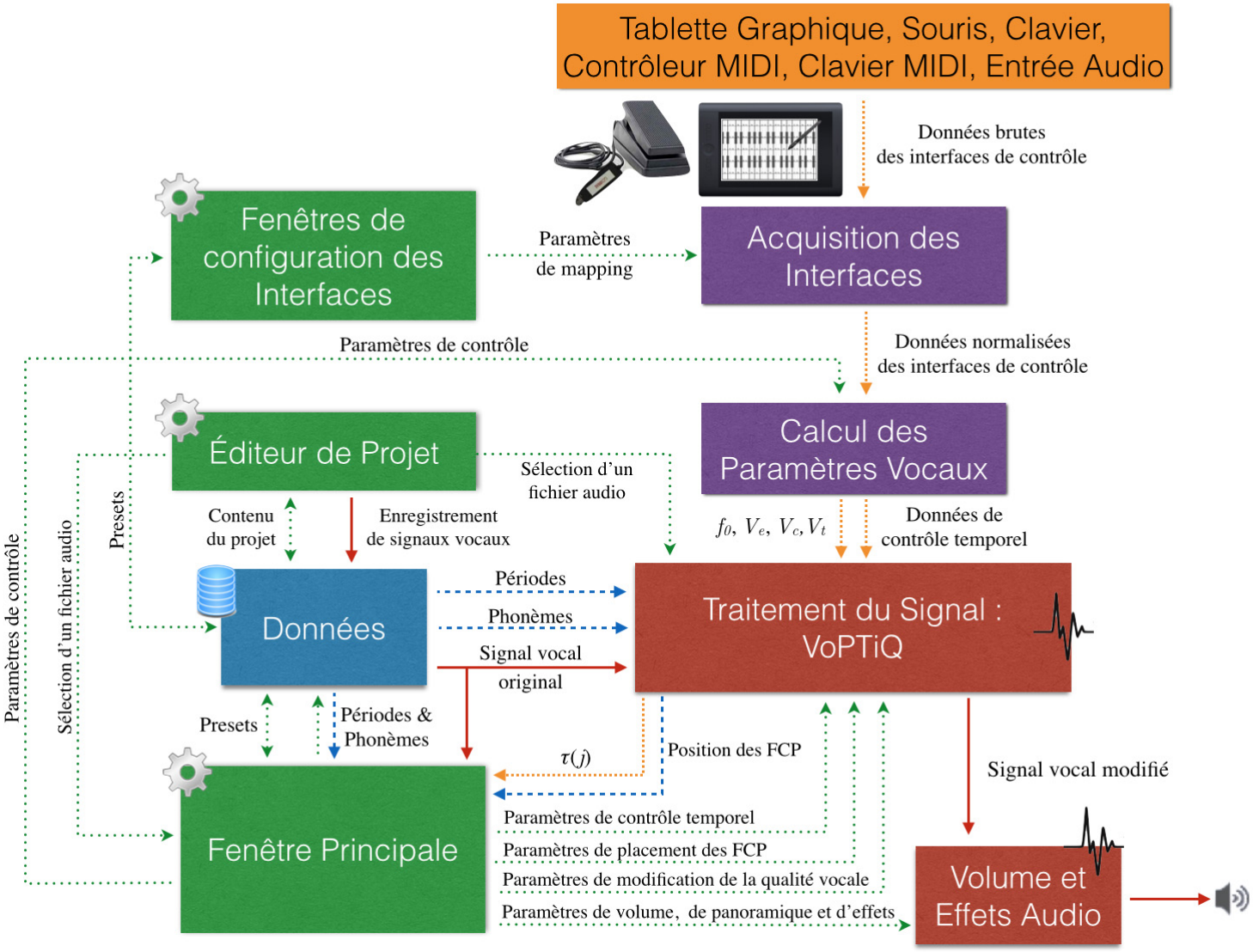


FIGURE 5.5 – *Vokinesis : Architecture du Logiciel. Le cadre orange représente les différentes interfaces de contrôle assignables à Vokinesis. Flèches rouges : signaux audio. Bleues : données d'analyse (marquages périodiques, phonétiques et position des FCP). Vertes : réglages effectués à partir d'interfaces graphiques. Oranges : données contrôlées en temps-réel.* $f_0$ *: fréquence fondamentale,* $V_e$ *: effort vocal,* $V_t$ *: tension vocale,* $V_c$ *: taille du conduit vocal,* $\tau(j)$ *: instant cible.*

### 5.2.2 Affichage du signal et édition de ses données d'analyse

La sélection d'un fichier audio dans l'éditeur de projet permet à la fenêtre principale d'afficher son contenu (flèche verte *Sélection d'un fichier audio*). Dans la FIGURE 5.2, le fichier *Léa* a été sélectionné dans le tableau du projet, et sa forme d'onde est affichée en haut de la fenêtre principale, ainsi que son spectrogramme et que ses données d'analyse (phonèmes et FCP). Le signal vocal original et ses données périodiques et phonétiques (fichiers `.gci` et `.phon`) sont récupérées à partir du disque dur (flèche rouge *Signal vocal original* et flèche bleue *Périodes et phonèmes*). Les données d'analyse peuvent être modifiées à la main à partir de la fenêtre principale, auquel cas les fichiers `.phon` et `.gci` correspondants seront remplacés (flèche verte *périodes & phonèmes*). Le positionnement des FCP est calculé directement par la partie *Traitement du Signal*, selon les étiquetages phonétiques et leurs paramètres de placement effectués dans la fenêtre principale (flèche bleue *position des FCP*).

Un autre paramètre qui peut être affiché sur la forme d'onde du signal est l'instant cible $\tau(j)$, c'est à dire l'instant temporel du signal original visé par un interprète à la $j^e$ période de synthèse (flèche orange $\tau(j)$). Une ligne verticale blanche indiquera alors l'instant du signal original actuellement re-synthétisé.

### 5.2.3 Paramétrages spécifiques et globaux

La fenêtre principale joue également un rôle de paramétrage très important : elle permet de régler les paramètres de placement des FCP, de modification de la qualité vocale, de volume, de panoramique et d'effets audio (flèches vertes à droite du cadre *fenêtre principal*), mais également les paramètres de contrôle (flèche verte de gauche), tels que les plages de contrôle de différents paramètres vocaux, les modes de contrôle temporel, etc... Tous ces paramétrages peuvent être sauvegardés dans des fichiers `.preset`, et récupérés lors du prochain chargement du fichier audio correspondant (double flèche verte *presets*). Ce sont des paramétrages spécifiques : ils peuvent être sauvegardés de façon indépendante pour chaque fichier du projet.

En bas à gauche de la fenêtre principale se trouvent les quatre boutons représentés au centre de la FIGURE 5.6. Un clic sur l'un de ces boutons permet d'ouvrir la fenêtre de configuration des interfaces correspondante. C'est dans ces fenêtres que le mapping entre les paramètres de contrôle des interfaces et les paramètres vocaux est effectué. Par exemple, si une utilisatrice souhaite faire en sorte que la hauteur soit contrôlée par la position du stylet sur l'axe $x$ de la tablette, elle devra cliquer sur le bouton *Tablet* pour ouvrir la fenêtre *Tablet Settings*, puis sélectionner l'option *Stylus X* dans la ligne *Pitch* (toutes les fenêtres de configuration matérielle seront présentées en détails dans la section 5.5). Tous ces réglages sont effectués de manière globale : ils concernent la totalité du projet. Ils peuvent être sauvegardés dans des fichiers `.preset`, et récupérés lors de l'ouverture d'un projet (double flèche verte *presets* dans la FIGURE 5.5). Ils sont émis à la partie *Acquisition des Interfaces* (flèche verte *paramètres de mapping*), afin de lui indiquer quelles données de tel ou tel contrôleur doivent être assignées à tel ou tel paramètre vocal.

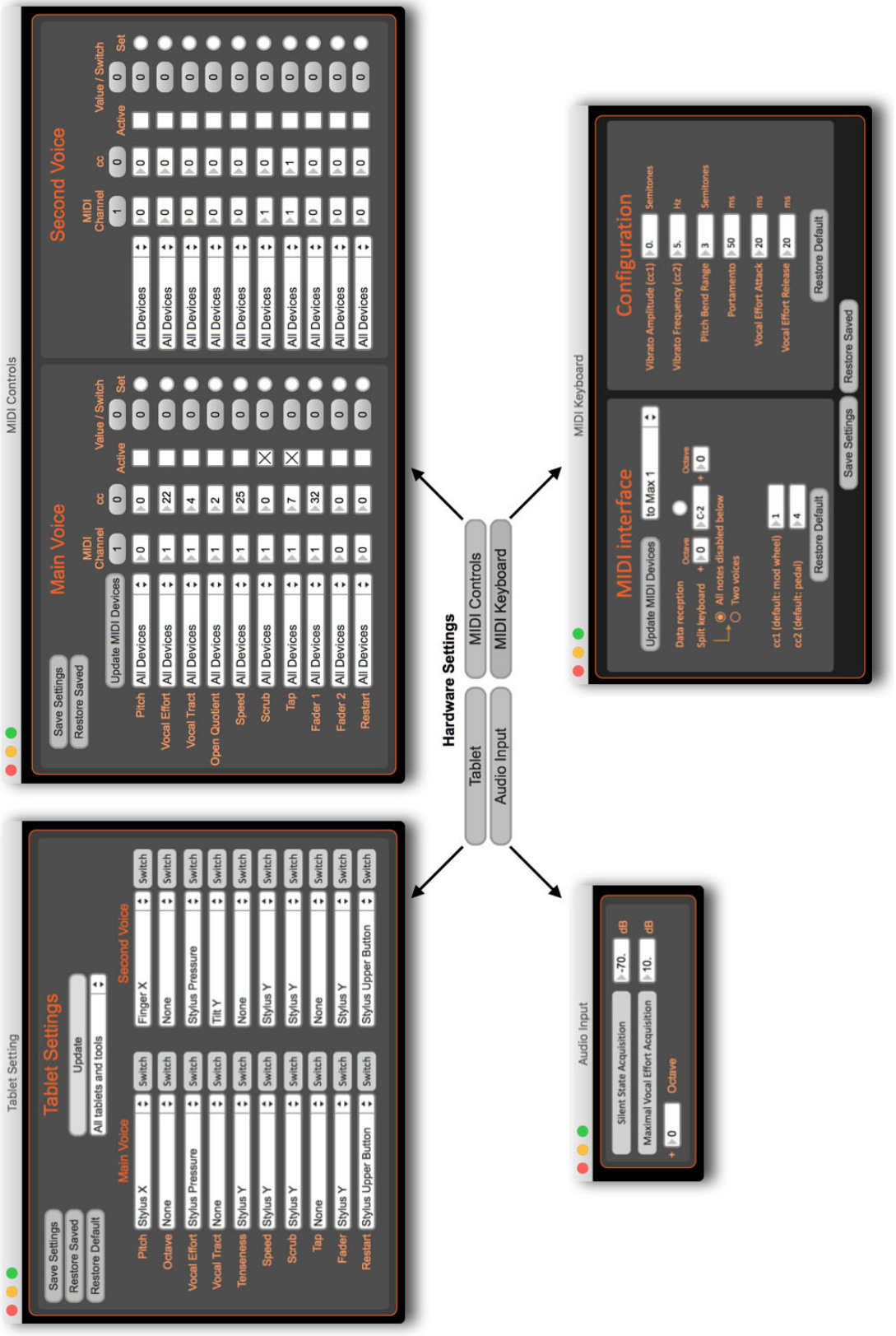


FIGURE 5.6 – *Ouverture des différentes fenêtres de configuration des contrôleurs à partir des boutons du coin inférieur droit de la fenêtre principale.*

### 5.2.4 Normalisation des données de contrôle

La partie *Acquisition des Interfaces* a pour rôle de normaliser les données brutes des contrôleurs. En effet, différents contrôleurs n'émettront pas forcément des données comprises dans les mêmes plages de valeurs. Par exemple, un contrôleur MIDI émet des données comprises entre 0 et 127, la pression appliquée au stylet est comprise entre 0 et 1, alors que son inclinaison correspond à une valeur comprise entre -90° et 90° sur chaque axe. La normalisation consiste donc à convertir toutes les données de contrôle en valeurs comprises entre 0 et 1. Ainsi, quelle que soit la configuration des interfaces de contrôle, la partie *Calcul des Paramètres Vocaux* recevra toujours le même type de données. C'est cette étape de normalisation qui donne à Vokinesis sa grande modularité : le logiciel interprétera toujours les données de contrôle de la même manière, quelle que soit la nature des interfaces qui y seront connectées.

### 5.2.5 Calcul des paramètres de contrôle suprasegmental et re-synthèse du signal original

Les données normalisées des contrôleurs sont ensuite converties en paramètres vocaux, selon les paramètres de contrôle réglés dans la fenêtre principale (flèche verte *paramètres de contrôle*). Les paramètres vocaux tels que la fréquence fondamentale, l'effort vocal, la taille du conduit et la tension vocale ($f_0$, $V_e$, $V_c$, $V_t$) ainsi que les données de contrôle temporel peuvent alors être transmis à la partie *Traitement du Signal*, qui modifiera le signal original en conséquence. Les données de contrôle temporel transmises dépendent du mode de contrôle temporel (voir le CHAPITRE 3), qui peut être sélectionné dans la fenêtre principale. En effet, le calcul de l'instant cible $\tau$ s'effectue dans la partie *Traitement du Signal*. La liste ci-dessous indique le type de données transmises pour chaque mode de contrôle temporel :

- Mode *Speed* (continu) : Vitesse de lecture.
- Mode *Scrub* (continu) : $\begin{cases} 0 \rightarrow \text{début du signal} \\ 1 \rightarrow \text{fin du signal} \end{cases}$
- Mode *Tap* (binaire) : $\begin{cases} 0 \rightarrow \text{bouton de contrôle relâché} \\ 1 \rightarrow \text{bouton de contrôle maintenu} \end{cases}$
- Mode *Fader* (continu) : Valeur comprise entre 0 et 1 représentant la position du potentiomètre dans sa plage de contrôle active, réglée dans la fenêtre principale (voir la section 5.5.2.2).

Pour les détails sur la calcul de l'instant cible, se référer au CHAPITRE 3.

La partie *Traitement du Signal*, qui contient la méthode VoPTiQ (CHAPITRE 2), est le cœur de Vokinesis : c'est ici que toutes les données d'analyse, de paramétrage et de contrôle se rejoignent pour la fabrication d'un signal re-synthétisé. Ainsi, nous

avons déjà eu l'occasion d'évoquer toutes les données qui y sont émises.

Le signal vocal re-synthétisé est émis en temps-réel à la partie *Volume et Effets Audio*, qui modifie le volume et la panoramique des voix de synthèse principale et secondaire, et y applique des effets audio (écho, réverbération et égalisation) selon les paramètres réglés dans la fenêtre principale. Les signaux ainsi modifiés sont ensuite émis à la sortie audio.

## 5.3 Mapping

Le *mapping* représente les stratégies de liaison entre les sorties des interfaces de contrôle et les paramètres d'entrée de l'algorithme de synthèse, et constitue une étape essentielle dans la conception d'un instrument numérique [Hunt *et al.* 2003]. Dans la section précédente, nous avons pu voir qu'une étape de normalisation des données de contrôle permettait l'assignation d'une multitude d'interfaces gestuelles aux différents paramètres de contrôle de la voix. Afin de rendre notre système modulable, nous avons souhaité faire en sorte que le mapping puisse être entièrement remodelé à la guise des interprètes.

Dans cette section, nous allons nous pencher sur les règles de mapping que nous avons mises en place, c'est à dire sur la façon dont les données émises par les interfaces de contrôle sont transformées et acheminées jusqu'aux entrées de la partie *Traitement du signal*, selon les réglages effectués dans les différentes interfaces graphiques. Pour ce faire, nous nous appuierons sur le schéma de la Figure 5.7, qui se lit de haut en bas.

Le code couleur reste très proche de celui utilisé pour les Figure 5.1 et 5.5. En voici les différences : la couleur verte représente toujours des interfaces graphiques (cadres) ou des réglages qui peuvent y être effectués (flèches), mais les fenêtres de configuration matérielle sont représentées en vert clair, et la fenêtre principale en vert foncé. Les numéros sur les différents cadres *Fenêtre principale* sont utilisés pour simplifier leur référencement. Les lignes continues représentent un paramètre unique, les lignes pointillées un ensemble de paramètres. Les points oranges indiquent la fusion de plusieurs paramètres. Les points noirs représentent des options d'interfaces ou de modes de contrôle et les flèches vertes représentent le choix de l'une de ces options. Les lignes vertes (sans flèche) correspondent à des pré-réglages de paramètres vocaux ou de plages de contrôle. Les valeurs entre crochet sont des valeurs normalisées (comprises entre 0 et 1), celles précédées d'un + sont des pré-réglages, et le symbole ⇔ signifie « plage de contrôle ». La position de chaque commutateur du schéma représente les réglages par défaut de Vokinesis.

### 5.3.1 Stratégies de mapping

Trois stratégies de mapping sont proposées par [Rovan *et al.* 1997]. Le mapping *one-to-one* ou *direct* consiste à assigner une donnée de contrôle gestuel à une donnée de synthèse. Le mapping *divergent* consiste à assigner plusieurs paramètres de synthèse à un paramètre de contrôle gestuel. Le mapping *convergent* consiste quant-à

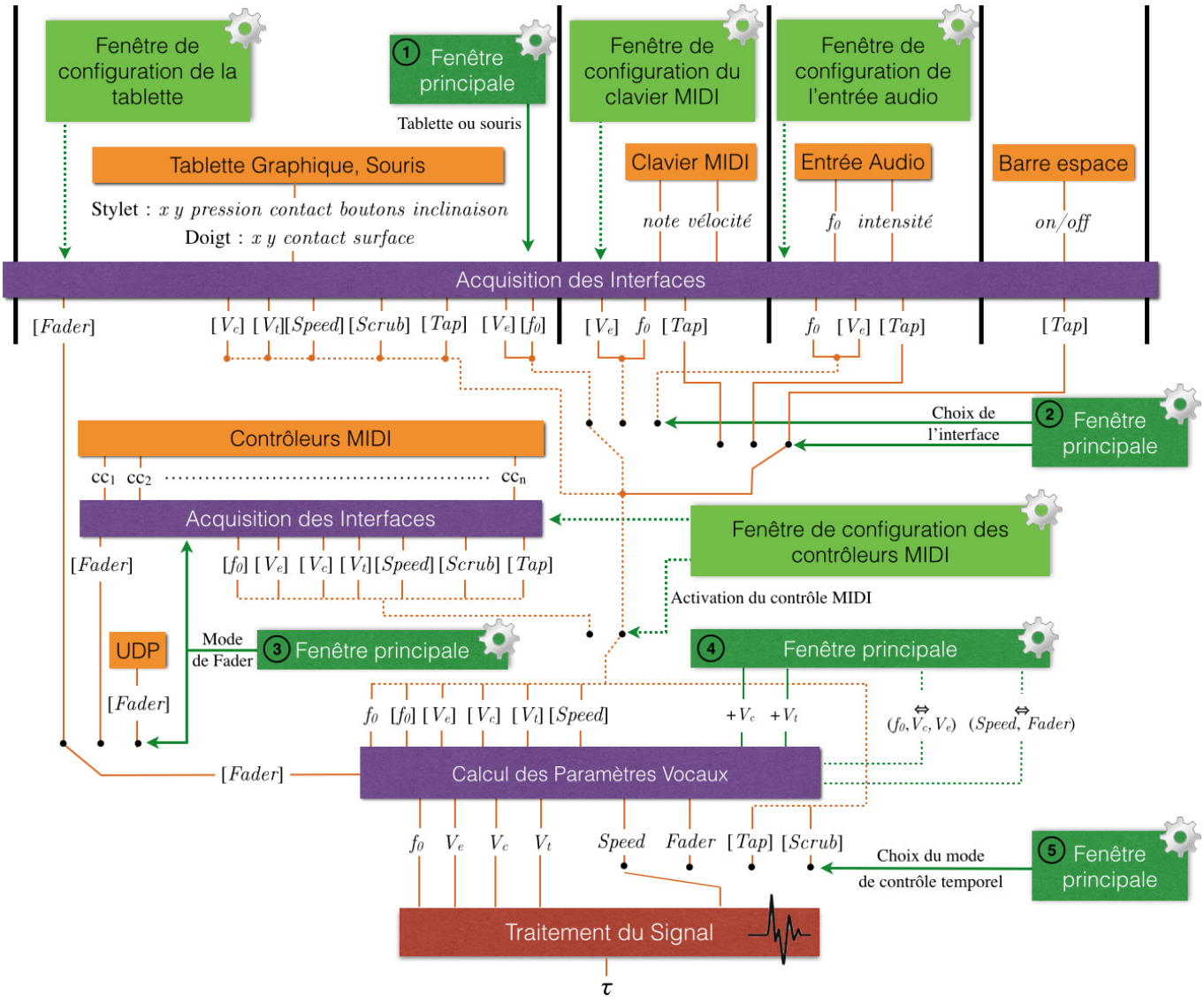


FIGURE 5.7 – *Architecture de paramétrage de Vokinesis. Lignes oranges : données contrôlées en temps-réel. Vertes : données prédéfinies dans l'une des interfaces graphiques. Lignes continues : paramètre unique. Lignes pointillées : ensemble de paramètres.* [valeur] *: valeur normalisée (comprises entre 0 et 1).* ⇔ *: plage de contrôle.* $f_0$ *: fréquence fondamentale,* $V_e$ *: effort vocal,* $V_c$ *: taille du conduit vocal,* $V_t$ *: tension vocale, AC : correction de justesse, TC : paramètres de correction temporelle des transitoires,* $\tau$ *: instant cible.*

lui à assigner le contrôle d'un seul paramètre de synthèse à plusieurs paramètres de contrôle gestuel.

Nous avons décidé de laisser la possibilité aux interprètes de choisir entre un mapping direct et un mapping divergent : chaque donnée de contrôle gestuel peut être assignée à plusieurs paramètres vocaux.

Les données de contrôle gestuel peuvent provenir d'une tablette graphique (ou d'une souris), d'un clavier et de contrôleurs MIDI, de la barre espace, d'un signal UDP ou encore d'une entrée audio (cadres oranges sur la FIGURE 5.7). Tous ces contrôleurs n'offrent pas le même nombre de degrés de liberté : une barre espace offre moins de possibilités que le stylet d'une tablette graphique. Ainsi, certaines interfaces qui offrent peu de degrés de liberté ne permettent pas de contrôler certains paramètres de synthèse. Par exemple, la barre espace ne peut contrôler que le séquencement binaire du cadre rythmique (*Tap*). Un clavier MIDI ou une entrée audio ne pourront contrôler que la fréquence fondamentale ($f_0$), l'effort vocal ($V_e$) ou le séquencement binaire du rythme. La tablette graphique, quant-à elle, offre assez de degrés de libertés pour permettre le contrôle simultané de la durée (*Tap*, *Fader*, *Speed* ou *Scrub*), de la $f_0$ et des paramètres de qualité vocale (effort, tension $V_t$ et taille du conduit $V_c$). De même, puisque le nombre de contrôleurs MIDI utilisables est quasiment illimité, ils peuvent êtres assignés à n'importe quel paramètre de synthèse. À ce jour, nous n'avons pas développé d'interface de paramétrage des signaux UDP, et nous ne l'utilisons que pour le contrôle continu des liaisons rythmiques (*Fader*) avec une Leap Motion (voir la section 3.3.6). Cependant, tout comme pour les contrôleurs MIDI, les possibilités d'utiliser de multiples signaux UDP est tout à fait envisageable, et les développements futurs devraient permettre de contrôler n'importe quel paramètre vocal avec un signal UDP.

La partie *Acquisition des Interfaces* se charge d'assigner les données de contrôle aux paramètres vocaux sélectionnés dans les fenêtres de configuration des interfaces. Par exemple, une utilisatrice pourrait choisir d'assigner le $\text{cc}_1$ d'un contrôleur MIDI à l'effort vocal $V_e$, et la position $y$ du stylet à la tension vocale $V_t$. Elle effectuerait les réglages nécessaires dans la fenêtre de configuration des contrôleurs MIDI d'une part, et dans la fenêtre de configuration de la tablette d'autre part. Les données de configuration alors transmises aux parties *Acquisition des Interfaces* (flèches vertes) lui indiquent à quel paramètre vocal assigner tel ou tel signal de contrôle gestuel. Si elle souhaitait utiliser une stratégie de mapping divergent, elle pourrait par exemple assigner la pression du stylet à la fois à l'effort vocal et à la tension.

### 5.3.2 Choix des contrôleurs

Nous allons expliquer ici la façon dont les données de contrôle gestuel sont acheminées jusqu'à la partie *Traitement du Signal*, en détaillant les différentes règles de priorité que nous avons mises en place afin d'empêcher l'utilisation d'une stratégie de mapping convergent. En effet, sauf pour le cas du mode *Fader Duo*, qui fait usage de deux pédales d'expression pour le contrôle continu des liaisons rythmiques, nous pensons qu'un paramètre vocal ne doit pas être contrôlé par plus d'un paramètre

gestuel, car son contrôle en deviendrait moins intuitif.

#### 5.3.2.1 Priorité aux contrôleurs MIDI

La priorité est accordée aux contrôleurs MIDI. Ceci est représenté sur la Figure 5.7 par un commutateur, enclenché par la fenêtre de configuration des interfaces MIDI. L'activation du contrôle MIDI pour un paramètre vocal enclenche le commutateur pour ce paramètre uniquement, et il ne sera donc plus contrôlable que par l'interface MIDI assignée. Dès que le contrôle d'un paramètre vocal par un cc MIDI est activé, les données correspondantes des autres interfaces de contrôle sont désactivées. Les sections suivantes expliqueront les règles de priorité que nous avons mises en place pour les données émises en amont de ce commutateur.

#### 5.3.2.2 Fréquence fondamentale et effort vocal

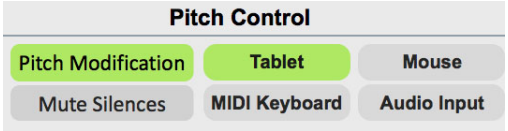


Figure 5.8 – *Choix de l'interface de contrôle de la hauteur et de l'effort vocal dans la fenêtre principale (boutons Tablet, Mouse, MIDI Keyboard et Audio Input).*

La Figure 5.8 montre la zone de la fenêtre principale permettant de choisir quel sera le contrôleur de la fréquence fondamentale et de l'effort vocal (*Tablet*, *Mouse*, *MIDI Keyboard* ou *Audio Input*). Intéressons nous pour l'instant aux choix *Tablet* et *Mouse* (flèche *Tablette ou souris* en sortie du cadre *Fenêtre principale* (1) sur la Figure 5.7). Le contrôle à la souris n'a été mis en place que pour permettre à des utilisateurs de simuler une tablette graphique s'ils n'en possèdent pas, et de tester ainsi les possibilités du logiciel. Il n'a donc pas vocation à être utilisé dans un cadre plus élaboré. Un clic sur le choix *Mouse* ouvre la fenêtre présentée Figure 5.9. Nous pouvons y voir le masque apposé à la tablette graphique. Un clic dans cette fenêtre émettra la position $(x, y)$ de la souris, considérée comme la position $(x, y)$ du stylet sur la tablette graphique, et une valeur de pression égale à 1 (pression maximale). Ainsi, tous les réglages liés à la position $(x, y)$ et à la pression du stylet dans la fenêtre de configuration de la tablette seront valables pour celles simulées par la souris. Par exemple, si un utilisateur assigne la taille du conduit vocal à la position $y$ du stylet, alors la position $y$ de la souris contrôlera également la taille du conduit vocal si l'option *Mouse* est sélectionnée.

Si l'un des choix *MIDI Keyboard* ou *Audio Input* est sélectionné (flèche verte supérieure du cadre *Fenêtre principale* (2)), alors les données de fréquence fondamentale et d'effort vocal émises par la tablette ou la souris seront désactivées. Le contrôle de la fréquence fondamentale et de l'effort vocal sont toujours liés, sauf si l'un de ces deux paramètres se voit assigner un cc MIDI : une utilisatrice pourrait vouloir continuer à contrôler la hauteur avec la position $x$ du stylet, mais assigner

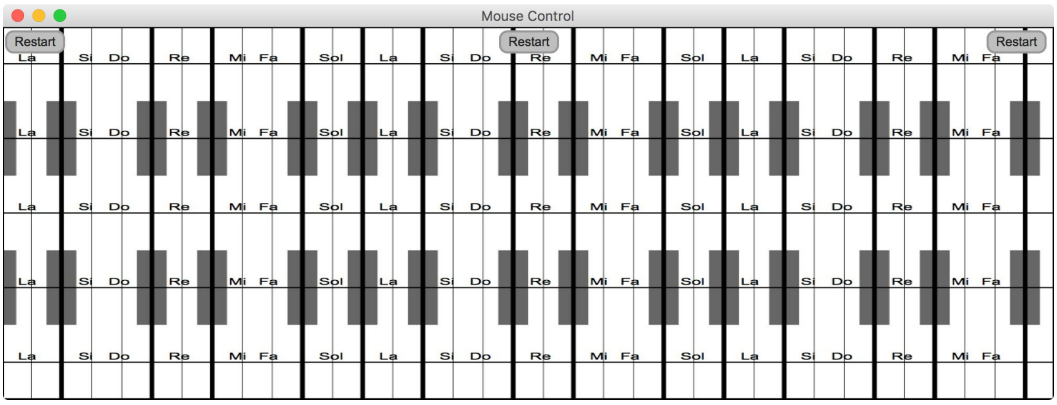


FIGURE 5.9 – *Fenêtre de simulation de la tablette graphique à la souris.*

le contrôle de l'effort vocal à une pédale MIDI, par exemple. Cependant, il semble peu probable qu'elle veuille contrôler la hauteur à la tablette graphique et l'effort vocal au clavier MIDI, ou inversement.

En observant la FIGURE 5.7, les lecteurs auront pu remarquer que le signal de contrôle de la fréquence fondamentale est normalisé pour la tablette, mais pas pour le clavier MIDI ou l'entrée audio. En effet, ces deux interfaces émettent directement une valeur de fréquence fondamentale correspondant à la note jouée. Il serait donc inutile de les normaliser pour les reconvertir plus tard. Les signaux émis par la tablette graphique, quant-à eux, ne correspondent pas directement à des notes. La partie *Calcul des Paramètres Vocaux* sélectionnera donc le signal de contrôle de la fréquence fondamentale adéquat, selon l'interface sélectionnée.

#### 5.3.2.3 Séquencement binaire du cadre rythmique (Tap)

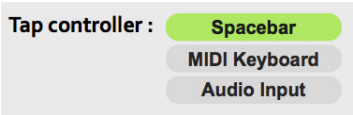


FIGURE 5.10 – *Choix de l'interface de contrôle binaire du séquencement du cadre rythmique dans la fenêtre principale.*

La FIGURE 5.10 montre la zone de la fenêtre principale permettant de sélectionner l'interface de contrôle binaire du séquencement du cadre rythmique (flèche verte *Choix de l'interface* inférieure en sortie du cadre *Fenêtre principale* (2) de la FIGURE 5.7). Par défaut, le séquencement binaire du cadre est assigné à la barre espace, et aucun signal de contrôle de la tablette n'y est assigné. Cependant, un utilisateur peut décider à partir de la fenêtre de configuration de la tablette d'y assigner un bouton, ou l'état de contact du stylet, par exemple. Dans ce cas, le séquencement du cadre pourra être effectué et par la tablette, et par le contrôleur sélectionné dans la fenêtre principale. Encore une fois, ces contrôles effectués en

amont seront désactivés si un contrôleur MIDI y est assigné en aval. Ce signal étant toujours compris entre 0 (position off) et 1 (position on), il est toujours représenté comme un signal normalisé dans le schéma. Si un paramètre continu de la tablette ou d'un contrôleur MIDI y est assigné, sa valeur sera quantifiée afin de conserver un signal binaire. Si l'entrée audio est sélectionnée comme interface de contrôle binaire du rythme, un seuil d'intensité permettra de définir l'état du signal de contrôle. Si l'intensité du signal d'entrée est inférieur à ce seuil, le signal de contrôle sera mis à 0. Sinon, il sera mis à 1.

#### 5.3.2.4 Contrôle continu des liaisons rythmiques (Fader)

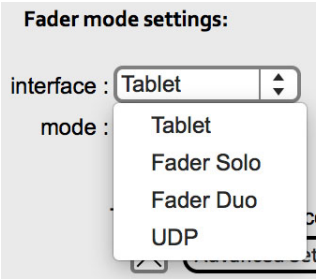


Figure 5.11 – *Choix de l'interface de contrôle continu du séquencement du cadre rythmique dans la fenêtre principale.*

La Figure 5.11 représente la zone de la fenêtre principale permettant de sélectionner l'interface de contrôle continu des liaisons rythmiques (flèche verte *Mode de Fader* en sortie du cadre *Fenêtre principale* (3) sur la Figure 5.7). Les options *Tablet* et *UDP* permettent d'assigner le contrôle à la tablette et au signal UDP. Les options *Fader Solo* et *Fader Duo* assignent le contrôle aux contrôleurs MIDI. Même si le mode *Fader Duo* est sélectionné, la partie *Traitement du signal* recevra toujours une seule valeur normalisée : les valeurs des deux contrôleurs MIDI assignés seront traitées dans la partie *acquisition des interfaces* et converties afin qu'elles soient considérées comme une valeur de potentiomètre unique.

### 5.3.3 Réglage des paramètres acoustiques et temporels du signal de synthèse

Le calcul des paramètres de synthèse s'effectue d'une part à partir des signaux de contrôle temps-réel, et d'autre part à partir des pré-réglages et des plages de contrôle configurés dans la fenêtre principale. Sur la Figure 5.12, La zone permettant d'effectuer les pré-réglages de taille du conduit vocal et de tension vocale se trouve à gauche, et les paramètres de plage de contrôle de la hauteur, de la taille du conduit vocal et de l'effort vocal se trouvent à droite. Cette figure correspond au cadre *Fenêtre principale* (4) de la Figure 5.7.

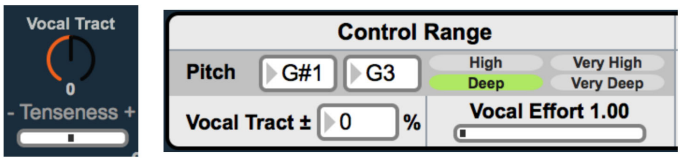


FIGURE 5.12 – *Zones de pré-réglage des paramètres vocaux et de leurs plages de contrôle dans la fenêtre principale. À gauche : pré-réglages de la taille du conduit vocal (Vocal Tract) et de la tension vocale (Tenseness). À droite : réglage des plages de contrôle de la hauteur (Pitch), de la taille du conduit vocal (Vocal Tract) et de l'effort vocal (Vocal Effort).*

#### 5.3.3.1 Fréquence fondamentale

Les signaux normalisés de contrôle de la hauteur sont convertis en valeur de fréquence fondamentale selon la plage de contrôle définie par les paramètres *Pitch* du cadre *Control Range* de la FIGURE 5.12, puis émis à la partie *Traitement du Signal*. La plage de contrôle peut être réglée à partir des presets *Very Deep* , *Deep*, *High* et *Very High*, correspodnant respectivement aux plages $[G\#0\,;\ G2]$, $[G\#1\,;\ G3]$, $[G\#2\,;\ G4]$ et $[G\#3\,;\ G5]$. Cependant, les valeurs minimale et maximale de cette plage de contrôle peuvent être réglées de façon indépendante.

#### 5.3.3.2 Effort vocal

La plage de contrôle de l'effort vocal peut être réglée par le paramètre *Vocal Effort* du cadre *Control Range* de la FIGURE 5.12. Ce paramètre peut prendre des valeurs comprises entre 1 et 2. Si un potentiomètre est assigné au contrôle de l'effort, et que sa valeur normalisée est comprise entre 0 et 1, alors la valeur d'effort vocal transmise au synthétiseur sera le résultat de la multiplication de la valeur du potentiomètre à celle du paramètre *Vocal Effort Control Range*. L'effort vocal du signal de synthèse correspond à l'effort vocal du signal original multiplié par la valeur d'effort vocal transmise au synthétiseur. Si le paramètre *Vocal Effort Control Range* est laissé à 1, alors l'effort vocal du signal de synthèse ne sera jamais supérieur à celui de l'original. Si ce paramètre est supérieur à 1, alors l'effort vocal pourra être augmenté.

#### 5.3.3.3 Tension vocale

Comme nous l'avons vu dans la section 2.5.1, un changement de tension vocale est simulé par l'augmentation ou la diminution de l'amplitude de l'harmonique fondamental. La plage de contrôle de la modification de la tension vocale est par défaut comprise entre $V_t = -1$ (diminution maximale) et $V_t = 1$ (augmentation maximale).

Cependant, dans les signaux originaux, le niveau de tension vocale peut varier fortement d'un enregistrement à l'autre. Voilà pourquoi Vokinesis offre la possibilité d'effectuer une modification préalable de la tension vocale, grâce au paramètre *Ten-*

*seness* en haut à gauche de la FIGURE 5.12. Ce paramètre peut prendre des valeurs allant de -1 à 1, et est ajouté à la plage de contrôle par défaut. Par exemple, si le pré-réglage *tenseness* est réglé à -1, la plage de modification de tension vocale sera $-2 \leq V_t \leq 0$.

#### 5.3.3.4 Taille du conduit vocal

Le paramètre *Vocal Tract* du cadre *Control Range* de la FIGURE 5.12 permet de configurer la plage de modification du conduit vocal, qui s'exprime en pourcentage. Si cette valeur est réglée à 20%, alors le conduit vocal pourra être allongé ou raccourci au maximum de 20%. À gauche de la figure se trouve un autre paramètre nommé *Vocal Tract* (au dessus du paramètre *tenseness*). C'est un paramètre de modification préalable de la taille du conduit vocal, qui fonctionne de la même manière que celui de la tension : sa valeur sera ajoutée à la valeur finale de modification de taille du conduit vocal émise au synthétiseur.

#### 5.3.3.5 Contrôle temporel

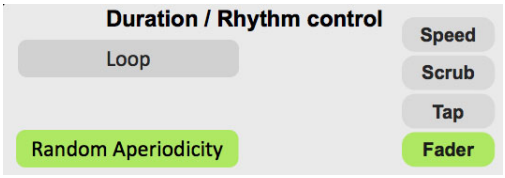


FIGURE 5.13 – *Choix du mode de contrôle temporel dans la fenêtre principale.*

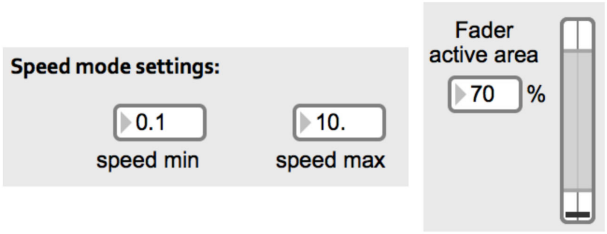


FIGURE 5.14 – *À gauche : réglage de la plage de contrôle de la vitesse de lecture. À droite : réglage de la plage active du fader.*

La FIGURE 5.13 correspond au cadre *Fenêtre principale* (5) sur la FIGURE 5.7. Elle représente la zone qui permet de choisir le mode de contrôle temporel dans la fenêtre principale. La FIGURE 5.14 correspond au cadre *Fenêtre principale* (4) de la FIGURE 5.7. Elle représente la zone permettant de régler la plage de contrôle de la vitesse de lecture en mode speed (à gauche), et celle permettant de définir la plage active du potentiomètre en mode *Fader* (à droite).

Le calcul de l'instant cible s'effectue toujours dans la partie *Traitement du Signal*, selon le mode de contrôle temporel sélectionné et les pré-réglages effectués. Pour plus

de précision, les lecteurs pourront se référer aux sections 3.1 et 3.3.

## 5.4 Programmation

Le développement de Vokinesis a été principalement effectué sous Max/MSP, un environnement graphique destiné à des applications multimédia. La programmation sous Max/MSP consiste à assembler des *objets* qui accomplissent des fonctions diverses (calculs numériques, traitement des signaux audio, éléments graphiques, etc...) et à les faire communiquer d'une façon adaptée à l'application recherchée. Un objet peut transmettre des données audio ou numériques. Elles sont acheminées d'un objet à un autre soit par un câble qui les relie, soit par une procédure d'émission/réception de messages. Les objets sont assemblés les uns aux autres dans une fenêtre de programmation, appelée *patch*. Il est possible de développer de nouveaux objets, soit en utilisant le langage Max/MSP, soit en utilisant un langage de programmation classique (C, C++, Java, Python, JavaScript). Les objets créés en Max/MSP sont nommés *sous-patchs*, ceux créés dans un autre langage sont nommés *externals*.

Dans cette section, nous allons détailler le cœur du système. Nous présenterons tout d'abord le sous-patch `VoPTiQ`, qui effectue les modifications des signaux originaux que nous avons présentées dans le Chapitre 2, puis nous donnerons des détails sur l'external Java `sd.VRTPSOLA`, qui joue un rôle central en terme de traitement du signal, mais également d'édition des données d'analyse. Enfin, nous présenterons brièvement les externals et sous-patch tiers que nous avons utilisés.

### 5.4.1 Sous-patch VoPTiQ

Le sous-patch `VoPTiQ`, présenté dans la Figure 5.15, correspond à la partie *Traitement du Signal* des Figures 5.5 et 5.7, et représente donc le cœur du système. Il possède un certain nombre de sous-patchs (indiqués par une lettre initiale `"p"`), mais également des objets de réception (`"r"`) et d'émission (`"s"`) de données numériques, et un external audio programmé en Java (`"mxj~"`), utilisé deux fois. Le code couleur respecte celui des Figures 5.5 et 5.7. Le vert désigne les messages provenant des interfaces graphiques (réglages), ainsi que ceux qui leur sont émis (données d'affichage). L'orange correspond à des données de contrôle temps-réel. Le rouge représente les différents éléments de VoPTiQ.

Ce sous-patch contient deux externals `sd.VRTPSOLA`, qui ont pour rôle principal d'effectuer les modifications de durée, de hauteur et de taille du conduit vocal d'un signal original (voir la section 2.3), selon les données de contrôle. L'external de gauche produit le signal de synthèse de la voix principale, et celui de droite produit celui de la seconde voix dans le cas d'un contrôle polyphonique. Les sous-patchs `vocalEffort` et `tenseness` effectuent les modification d'effort vocal et de tension indépendamment pour les deux voix de synthèse (voir les sections 2.5.2 et 2.5.1), selon les données de contrôle. Tous ces éléments de traitement de signal constituent la méthode VoPTiQ.

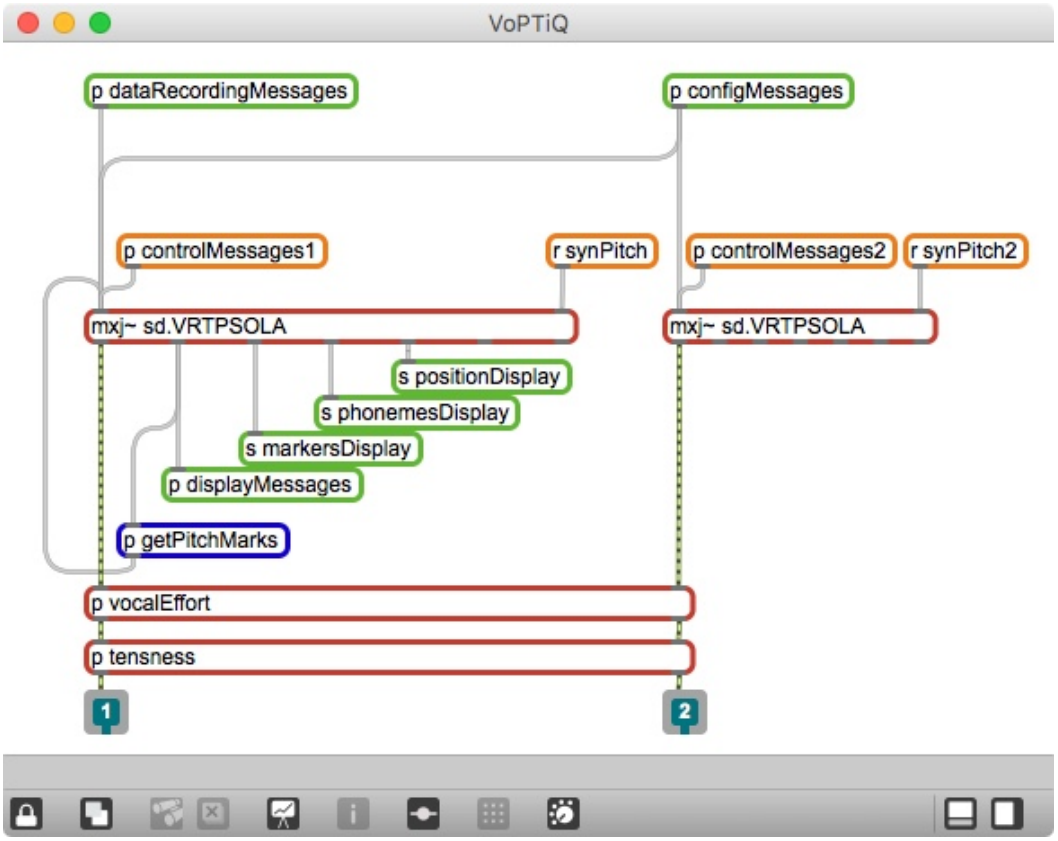


Figure 5.15 – *Sous-patch Max/MSP* `VoPTiQ`. *Vert : messages de communication avec les interfaces graphiques. Orange : données de contrôle temps-réel. Rouge : éléments de la méthode VoPTiQ. Bleu foncé : analyse de périodicité des signaux originaux.*

Le sous-patch `configMessages` a pour rôle de rassembler toutes les données de réglages spécifiques et de les émettre aux synthétiseurs des deux voix. Par contre, les messages de contrôle temps-réel leur sont transmis de façon indépendante par les récepteurs `synPitch` et `synPitch2` (fréquence de synthèse de chacune des voix), ainsi que par les sous-patchs `controlMessages1` et `controlMessages2` (tous les autres paramètres de contrôle).

Un rôle secondaire de l'external `sd.VRTPSOLA` est d'émettre des données concernant l'affichage de certaines informations sur la zone du signal dans la fenêtre principale (position de l'instant cible $\tau$, étiquetage des phonèmes, position des FCP, marqueurs périodiques, début et fin de la boucle). Les quatre objets verts se trouvant sous l'external `sd.VRTPSOLA` de gauche ont pour rôle d'envoyer ces messages aux éléments graphiques correspondants. Seul l'external de la voix principale envoie ces données : il n'est pas utile d'émettre plusieurs fois des informations identiques.

Vokinesis offre la possibilité de manipuler les étiquetages phonétiques et périodiques directement sur l'affichage du signal (plus de détails dans la section 5.5.2.4). L'external `sd.VRTPSOLA` a également pour rôle de gérer l'édition manuelle de ces marqueurs : des données de modification des étiquetages sont transmises par le sous-patch `configMessages`, et `sd.VRTPSOLA` met directement à jour leur affichage sur la zone du signal.

Dans la section 5.5.4, nous présenterons une interface graphique permettant de sélectionner des données à enregistrer lors d'une performance. Par exemple, il est possible d'indiquer au système d'enregistrer l'évolution d'un paramètre de contrôle gestuel ou d'un paramètre vocal au cours du temps. Le sous-patch `dataRecordingMessages` a pour rôle d'indiquer au synthétiseur de la voix principale les messages concernant la configuration de ces données à enregistrer. Pour l'instant, seules les données de la voix principale peuvent être enregistrées. Les futurs développements pourront remédier à cette lacune.

Le sous-patch `getPitchMarks` a pour rôle de lancer la procédure d'analyse de périodicité lorsqu'un nouveau signal est ajouté au projet, par l'exécution d'un script Praat présenté ci-dessous. L'exécution de ce script permet d'obtenir un fichier `.PointProcess` (un format reconnu par Praat), qui est directement converti en un fichier `.gci` par un code Java.

### Commande shell pour lancer le script Praat

```
[path_to_praat] [path_to_script] [path_to_audiofile]
```

### Script Praat

---

```
form Read a wav file in given directory
   sentence SourceDirectory /Users/delalez/Programmation/
   sentence FileName file
```

```
endform


## Below: loop through the list of files, extracting each name and reading it into the Objects list


   file$ = sourceDirectory$ + fileName$ + ".wav"
        Read from file: file$


# Create Point Process

        soundName$ = replace$ fileName$, " ", "_", 0

        select Sound 'soundName$'
                    To Pitch... 0 50 600
        select Sound 'soundName$'
        plus Pitch 'soundName$'
                    To PointProcess cc

        select PointProcess 'soundName$'_'soundName$'
    savepath$ = sourceDirectory$ +"/"+ fileName$ + ".PointProcess"
        Write to short text file... 'savepath$'
        To TextGrid vuv... 0.02 0.01

         select TextGrid 'soundName$'_'soundName$'
         savepath$ = sourceDirectory$ +"/"+ fileName$ + ".vuv"
        Write to short text file... 'savepath$'
```

---

### 5.4.2 External sd.VRTPSOLA

Comme nous venons de le voir, l'external `sd.VRTPSOLA` joue plusieurs rôles importants dans le fonctionnement de Vokinesis. Il permet d'une part de gérer les transformations de durée, de hauteur et de taille du conduit vocal, mais également l'édition des étiquetages phonétiques et périodiques. Ces deux rôles sont illustrés dans la Figure 5.16. Les cadres rouges représentent des classes Java incluses dans `sd.VRTPSOLA`. Les cadres gris représentent des listes de donnés. Les données de communication entre Max/MSP et `sd.VRTPSOLA` sont représentées par des flèches noires, et les données de communication entre `sd.VRTPSOLA` et ses classes sont indiquées par des flèches rouges.

#### 5.4.2.1 Obtention du signal de synthèse

`sd.VRTPSOLA` permet de modifier la hauteur, la durée et la taille du conduit vocal d'un signal original, selon les données de contrôle vocal émises par Max/MSP. Dans les modes de contrôle temporel *Tap* et *Fader*, le calcul de l'instant cible $\tau$ dépend des positions des FCP, dont les emplacements en échantillon sont stockés dans la liste

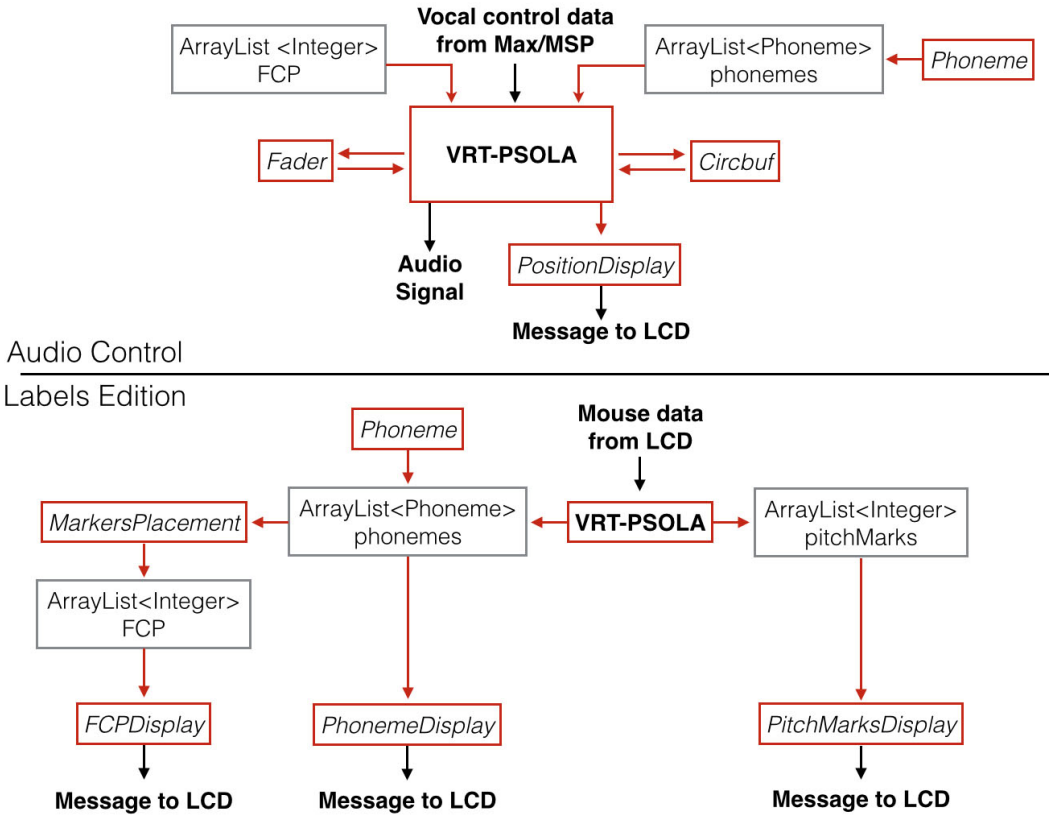


FIGURE 5.16 – *Schéma structurel de* `sd.VRTPSOLA` *pour le contrôle du signal audio (en haut) et pour l'édition des étiquetages périodiques et phonétiques (en bas). Cadres rouges : classes incluses dans* `sd.VRTPSOLA`*. Cadres gris : listes de données. Flèches noires : données de communication entre Max/MSP et* `sd.VRTPSOLA`*. Flèches rouges : données de communication entre* `sd.VRTPSOLA` *et ses classes.*

d'entiers `FCP`. La classe `Phoneme` permet de définir l'identité d'un phonème (nom du phonème, échantillons de début et de fin), et tous les phonèmes d'une phrase originale sont contenus dans la liste `phonemes`. En mode *Fader*, les données émises par le potentiomètre de contrôle sont parfois traitées de façon relativement complexe. Nous avons donc créé une classe `Fader` consacrée à cette tâche. Le calcul du signal de synthèse s'effectue dans la classe `Circbuf`, qui contient le buffer circulaire (voir la section 2.3.4). Une fois les données audio calculées, elles sont retransmises à `sd.VRTPSOLA`, qui se chargera de le transmettre à Max/MSP. Max/MSP possède un objet nommé `LCD`, qui permet d'afficher des éléments graphiques en lui envoyant des messages qu'il pourra interpréter. Le rôle de la classe `PositionDisplay` est d'émettre des messages à un objet `LCD` concernant l'affichage de la position de l'instant cible.

#### 5.4.2.2 Édition des étiquetages périodique et phonétique

L'objet `LCD` peut également émettre des messages concernant la position relative et l'état (cliqué ou non) de la souris. Les phonèmes, les FCP et les marqueurs périodiques sont affichés dans la zone du signal grâce à un objet `LCD`. Il est possible d'éditer l'étiquetage des phonèmes et la position des marqueurs périodiques en cliquant sur leur affichage et en les faisant glisser (voir la section 5.5.2.4). Les données de la souris sont alors transmises à `VRT-PSOLA`, qui mettra à jour la liste d'entiers `pitchMarks` comportant la position des marqueurs périodiques, ainsi que la listes des phonèmes selon les modifications effectuées. Lorsque la liste des phonèmes est mise à jour, la classe `MarkersPlacement` se charge de calculer le positionnement des FCP selon les règles que nous avons présentées section 3.4, et met alors à jour la liste `FCP`. Les messages à envoyer à l'objet `LCD` concernant la position des FCP, l'étiquetage des phonèmes et des marqueurs périodiques. Ces données sont transmises par les classes `FCPDisplay`, `PhonemeDisplay` et `PitchMarksDisplay`.

### 5.4.3 Externals et sous-patchs tiers

Les données de la plupart des interfaces de contrôle (clavier, souris, interfaces MIDI, signaux UDP) sont très facilement récupérées grâce à des objets inclus dans Max/MSP. Cependant, la tablette graphique est une interface assez spécifique, initialement destinée au dessin assisté par ordinateur, et non à des application audio. Il a donc fallu programmer des externals permettant de récupérer les données du stylet d'une part, et des fonctions tactiles d'autre part. Ce travail a été fait au LMA, et les externals `s2m.wacom` et `s2m.wacomtouch` qui y ont été développés sont disponibles à l'adresse [1]. Un autre external que nous utilisons dans Vokinesis nous permet de détecter en temps-réel la fréquence fondamentale d'une entrée audio. Cet external créé par [Puckette *et al.* 1998] se nomme `fiddle`, et est disponible à l'adresse [2]. C'est lui que nous utilisons pour le contrôle de la hauteur de signal de synthèse avec une entrée audio.

1. http://www.maxobjects.com/?v=libraries&id_library=163
2. http://www.maxobjects.com/?v=objects&id_objet=977

Vokinesis a hérité de quelques fonctionnalités du Cantor Digitalis. Le sous-patch permettant de traduire les données émises par un clavier MIDI en fréquence et en effort vocal de synthèse que nous utilisons est une adaptation du celui créé par [Feugère *et al.* 2017]. Le sous patch de correction dynamique de la hauteur qui permet d'améliorer la justesse du jeu mélodique a été développé par [Perrotin 2015].

## 5.5 Emploi du logiciel

Sortons à présent des détails intrinsèques, et observons Vokinesis du point de vue utilisateur. Dans cette section, nous détaillerons chaque élément des différentes interfaces graphiques permettant de paramétrer le système. Comme nous avons eu l'occasion de le constater au cours de ce chapitre, le logiciel contient différentes fenêtres, que nous distinguerons par trois types principaux :

— L'*éditeur de projet*, qui permet d'ouvrir ou de créer un projet, d'y ajouter et de sélectionner des fichiers audio originaux.

— La *fenêtre principale*, qui permet de visualiser le fichier sélectionné et d'effectuer les réglages spécifiques.

— Les *fenêtres de paramétrage*, qui permettent d'effectuer les réglages globaux.

Les sections suivantes détailleront l'utilisation de chacune des interfaces graphiques de Vokinesis.

### 5.5.1 Éditeur de projet

L'éditeur de projet est représenté FIGURE 5.17. Il contient des boutons permettant de sélectionner ou de créer un projet (*Create / Load Project*), d'ajouter un fichier audio au projet (*Add File*) ou d'enregistrer un nouveau fichier audio (*Record*). Le rôle des boutons *Task Setup* et *Subject Page* sera détaillé dans la section 5.5.4. Une fois un fichier ajouté ou enregistré, son nom sera affiché dans le tableau blanc. L'option *Edit Table* permet, lorsqu'elle est activée, de déplacer les fichiers dans le tableau, de les dupliquer ou de les supprimer. Le fait de dupliquer un fichier peut s'avérer utile dans le cas où un utilisateur souhaiterait lui assigner plusieurs réglages spécifiques, et ainsi disposer de plusieurs manières de le contrôler au sein d'un même projet.

### 5.5.2 Paramétrages spécifiques : fenêtre principale

La fenêtre principale, présentée dans la FIGURE 5.18 est divisée en trois zones : La *Zone du Signal*, la *Zone des Effets* et la *Zone de Contrôle*. Chacune de ces zones sera détaillée ci-dessous. Notez que tous les réglages qui s'effectuent à partir de cette fenêtre sont des réglages spécifiques : ils peuvent être sauvegardés indépendamment pour chaque élément du tableau du projet.

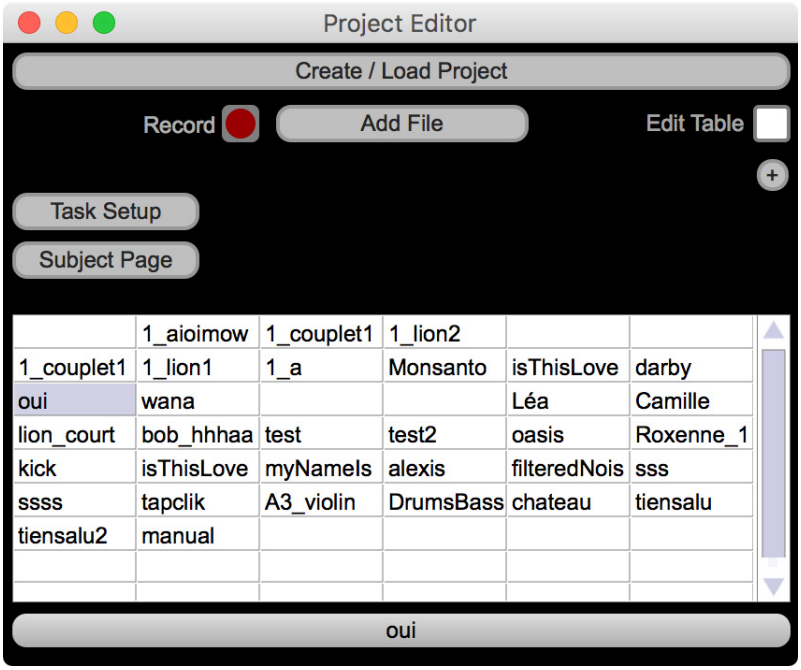


FIGURE 5.17 – *Fenêtre de Projet de Vokinesis. Les noms des fichiers contenus dans le projet sont indiqués dans le tableau du projet. Le fichier actuellement sélectionné se nomme « oui ».*

#### 5.5.2.1 Zone du signal

La zone du signal a différents rôles d'affichage et de réglage. Elle est détaillée dans la FIGURE 5.19. Elle permet d'abord d'afficher la forme d'onde du signal sélectionné (20), ainsi que son nom (1). Son spectrogramme sera également affiché si l'option (18) est activée. Les options (4) permettent de gérer l'intensité de l'affichage de la forme d'onde et du spectrogramme. Il est possible de zoomer (2) ou de déplacer le début de l'affichage (5). Si l'option *Edit* (3) est activée, il est possible d'éditer les étiquetages périodiques et phonétiques du signal original (les procédures seront détaillées dans la section 5.5.2.4). Les boutons (14) & (15) permettent de sauvegarder les modifications appliquées aux étiquetages, ou de restituer la dernière sauvegarde. Un clic sur l'un des boutons de sauvegarde aura pour effet de créer dans le dossier `data` un fichier `.gci` (périodes) ou `.phon` (phonèmes) qui sera assigné à l'élément sélectionné dans le tableau du projet, et restitué lors de sa sélection. Si l'option *Position* est activée (19), alors l'instant cible sera indiqué par un ligne verticale blanche sur la forme d'onde du signal, et le phonème actuellement prononcé sera mis en évidence. Dans la figure, l'instant cible se trouve au centre du phonème noté E. Elle permet aussi d'afficher le spectre instantané des signaux de synthèse (8). Le bouton *Open Visualisation Page* (7) permet d'ouvrir la page de visualisation représentée FIGURE 5.20, qui offre un affichage de la forme d'onde (en haut) et du spectre (en bas) des voix de synthèse principale (à gauche) et secondaire (à droite).

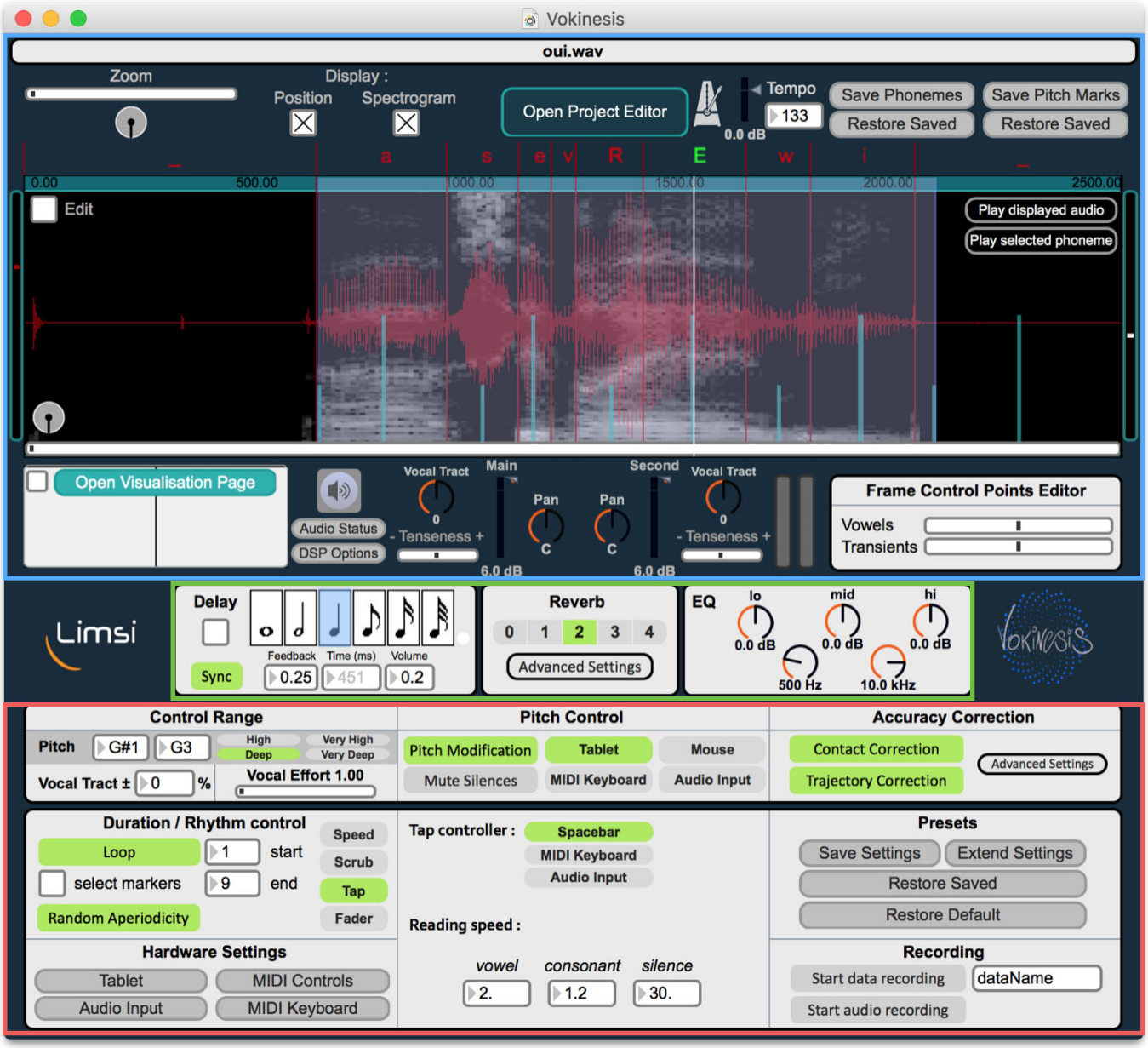


FIGURE 5.18 – *Fenêtre principale de Vokinesis ; Cadre supérieur : zone du signal. Cadre central : zone des effets. Cadre inférieur : zone de contrôle.*

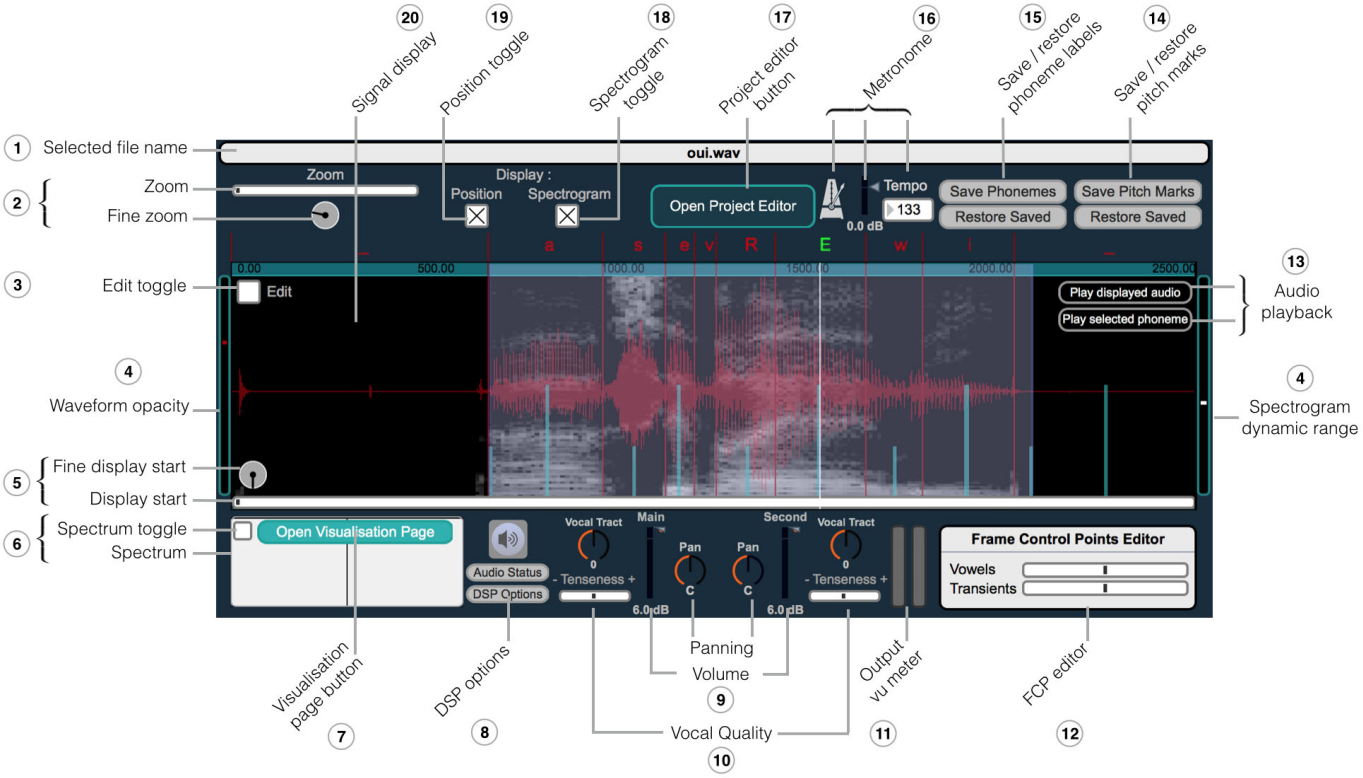


FIGURE 5.19 – *Zone du signal dans la fenêtre principale de Vokinesis*

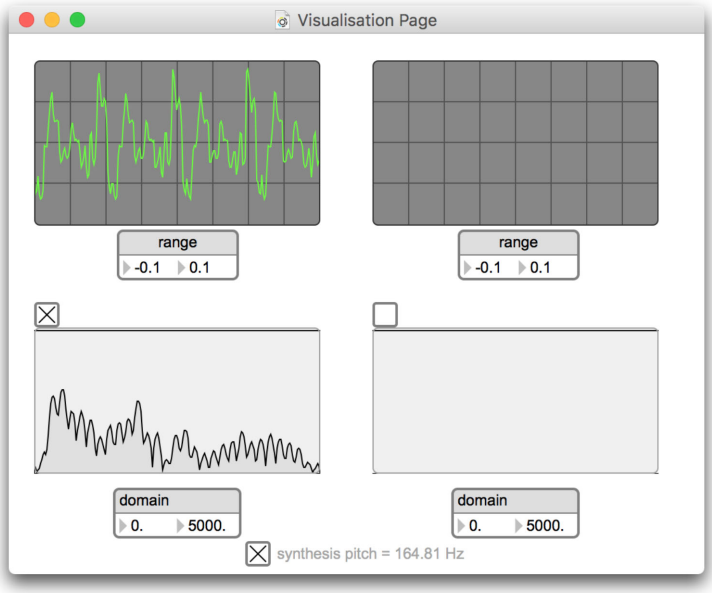


FIGURE 5.20 – *Page de visualisation des signaux et spectres des voix principale (à gauche) et secondaire (à droite).*

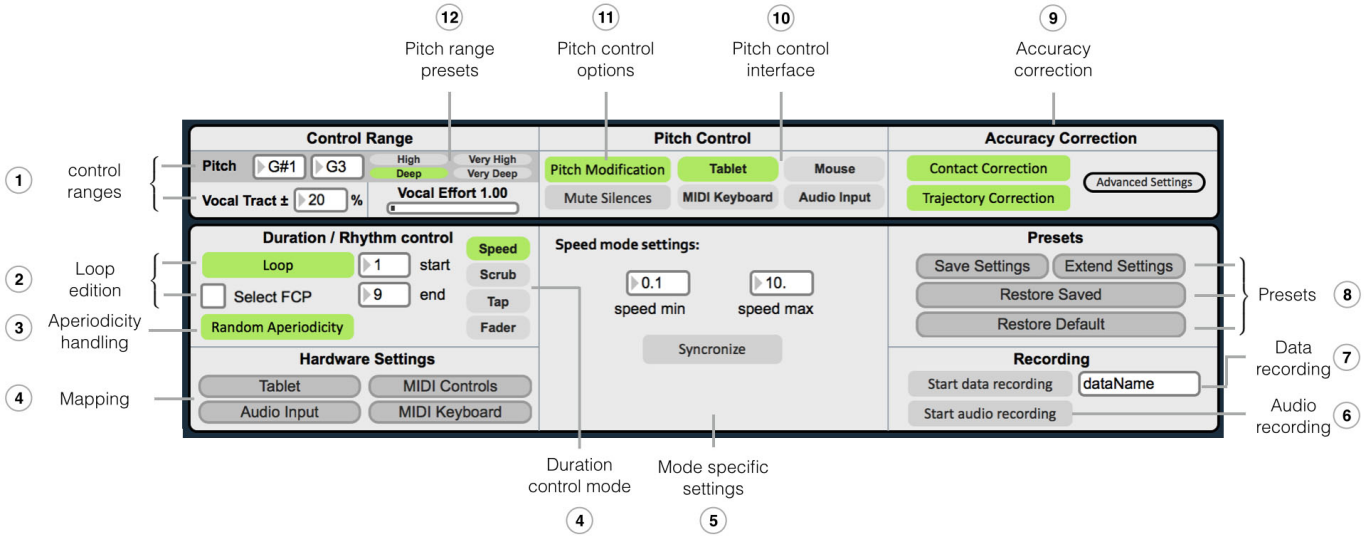


FIGURE 5.21 – *Zone de contrôle dans la fenêtre principale de Vokinesis*

Le cadre *Frame Control Points Editor* (12) permet de définir l'emplacement des FCP au sein d'une voyelle (*Vowels*) et de la dernière consonne des parties transitoires (*Transients*). Il est également possible d'effectuer les réglages relatifs aux signaux de synthèse (9) & (10) : panoramique (*Pan*), volume des voix principale (*main*) et secondaire (*second*), taille du conduit vocal (*Vocal Tract*) et tension vocale (*Tenseness*). Les paramètres de taille du conduit et de tension permettent de modifier à l'avance le timbre du signal sélectionné. Un clic sur l'icône du métronome (16) active ou désactive un battement isochrone. La valeur indiquée par le paramètre *Tempo* indique le nombre de battements que le métronome effectuera en une minute. Le bouton *DSP options* (8) permet d'ouvrir une fenêtre de paramétrage du traitement du signal (choix de la fenêtre d'analyse, taille du buffer circulaire, etc...) Cette fenêtre est utilisée à des fins de développement, et nous ne la détaillerons pas. le bouton *Audio Status* permet d'ouvrir la fenêtre de configuration des entrées/sorties audio (taille des buffers, choix de la carte son, etc...) L'icône du haut parleur permet d'activer / désactiver l'audio. Enfin, le bouton *Open Project Editor* (17) permet d'ouvrir l'éditeur de projet.

### 5.5.2.2 Zone de contrôle

La zone de contrôle permet d'effectuer tous les réglages relatifs au contrôle des paramètres vocaux. Elle est détaillée dans la FIGURE 5.21. Elle contient plusieurs cadres que nous présenterons ci-dessous. Pour simplifier les explications, nous imaginerons que tous les paramètres vocaux sont assignés à un potentiomètre, dont nous considérerons les valeurs minimale, centrale et maximale. Toutes les valeurs comprises entre ces trois cas seront interpolées linéairement.

Le cadre **Control Range** permet de régler les plages de contrôle des paramètres vocaux (hauteur, taille du conduit vocal, tension et effort vocal) qui seront dispo-

nibles sur les potentiomètres assignés (1). Dans la ligne du haut (*Pitch*), Les deux paramètres de gauche indiquent les valeurs minimale et maximale de contrôle de la hauteur. À leur droite, les options (12) *Very Deep, Deep, High* et *Very High* permettent respectivement de leur assigner les plages [*G#0* ; *G2*], [*G#1* ; *G3*], [*G#2* ; *G4*] et [*G#3* ; *G5*]. Cependant, ces valeurs peuvent être réglées manuellement et de façon indépendante. En bas à gauche de ce cadre, le paramètre *Vocal Tract* ± permet de régler la plage de contrôle de modification de la longueur du conduit vocal. Si le potentiomètre assigné est dans sa position centrale, la taille du conduit sera modifiée de 0%. S'il est dans sa position minimale, le conduit vocal sera ici rapetissé de 20%. S'il est dans sa position maximale, il sera agrandi de 20%. Notez que ces valeurs sont ajoutées à celle indiquée par le paramètre *Vocal Tract* de la zone du signal (FIGURE 5.19, (10)). Enfin, en bas à droite, le paramètre *Vocal Effort* permet de définir la valeur maximale assignée au potentiomètre, sa valeur minimale étant toujours 0 (pas de voix). Ainsi, si ce paramètre est laissé à 1, l'effort vocal maximal correspondra à l'effort vocal original. S'il est supérieur à 1, alors l'effort vocal pourra être augmenté.

Le cadre **Pitch Control** permet tout d'abord d'activer ou de désactiver la modification de la hauteur grâce au bouton *Pitch Modification* (11). S'il est actif, alors la fréquence de synthèse correspondra à la fréquence désignée par l'interface de contrôle de la hauteur. S'il est inactif, elle correspondra à la fréquence du signal original. Le bouton *Mute Silences* (11) permet, lorsqu'il est en fonction, de désactiver toutes les parties du signal dont l'étiquetage de phonèmes correspond au silence, et d'éviter ainsi certains bruits liés par exemple à la respiration du locuteur original. Enfin, ce cadre permet de sélectionner l'interface de contrôle de la hauteur (10) : une tablette graphique (*Tablet*), une souris (*Mouse*), un clavier MIDI (*MIDI Keyboard*) ou une entrée audio (*Audio Input*). Si le contrôle à la souris est sélectionné, la fenêtre de contrôle présentée plus haut FIGURE 5.9 s'ouvrira. Le cadre **Accuracy Correction** (9) permet d'activer ou de désactiver l'algorithme de correction de justesse proposé par [Perrotin & d'Alessandro 2013].

Le cadre **Duration / Rhythm Control** permet d'effectuer les réglages relatifs au contrôle du rythme et de la durée. Tout d'abord, il permet de choisir le mode de contrôle temporel (4) (*Speed*, *Scrub*, *Tap* ou *Fader*, voir le CHAPITRE 3). Les options (2) concernent l'édition des boucles. Si l'option *Loop* est activée, le signal de synthèse bouclera entre les FCP indiqués par *start* et *end*. Si l'option *Select FCP* est activée, il sera possible de définir les FCP *start* et *end* en les sélectionnant directement sur la zone du signal (FIGURE 5.19, (20)). Notez que les option *start*, *end* et *Select FCP* ne sont affichées que si l'option *Loop* est activée. Le cadre (5) change d'apparence selon le mode de contrôle temporel sélectionné. Ceci est illustré dans la FIGURE 5.22.

En mode *Speed*, il est possible de définir les vitesses minimale et maximale assignées au potentiomètre. Le paramètre *Speed min* peut prendre des valeurs positives ou négatives. S'il est positif, la position centrale du potentiomètre correspondra à une vitesse de 1, égale à celle du signal original, et le signal sera ralenti dans la partie inférieure du potentiomètre, accéléré dans sa partie supérieure. S'il est négatif, la position centrale correspondra à une vitesse de 0 (l'instant cible n'évolue plus) et le

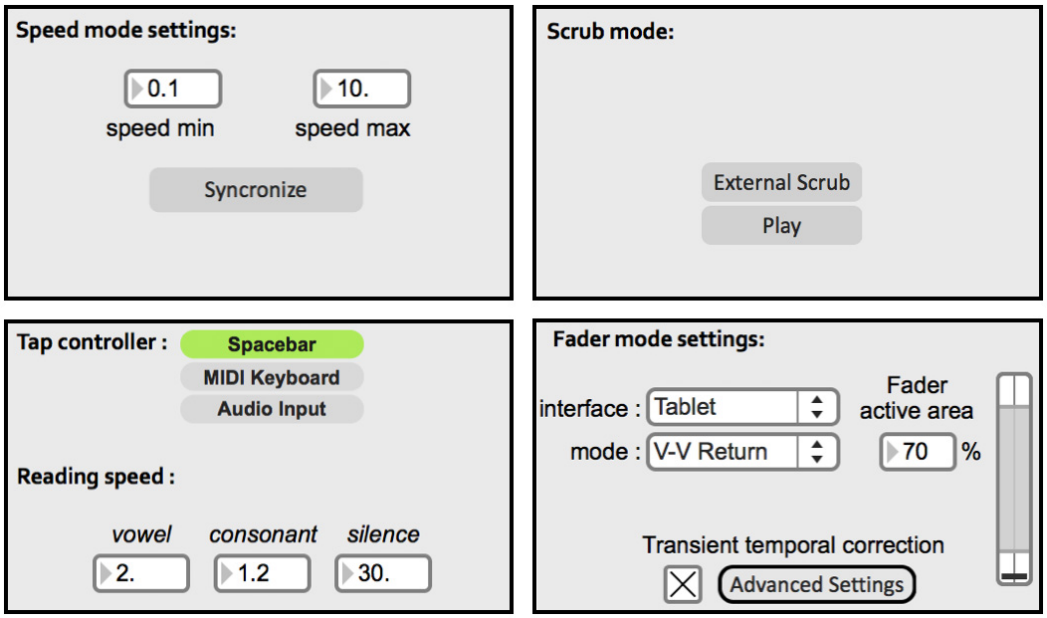


FIGURE 5.22 – *Cadre central de la zone de contrôle pour les différents modes de contrôle temporel. Dans l'ordre de lecture : Modes* Speed, Scrub, Tap, *et* Fader.

signal sera lu à l'envers dans la partie inférieure du potentiomètre.

Lorsque le mode *Tap* est sélectionné (voir la section 3.3.2), le cadre central permet de choisir le contrôleur binaire du rythme syllabique : la barre espace (*Spacebar*), les touches d'un clavier MIDI (*MIDI Keyboard*) ou un signal d'entrée (*Audio Input*). Il permet également de régler la vitesse à laquelle seront lues les voyelles, consonnes et silences lors des transitions entre deux FCP. Les valeurs indiquées sur la figure sont celles que nous préconisons. La lecture rapide des voyelles n'est pas un problème pour la qualité de la synthèse, et permet de produire un rythme plus rapide que l'original. La lecture un peu plus lente des consonnes permet d'en conserver l'intelligibilité. La lecture très rapide des silences permet d'éviter un temps d'attente dans le cas où le signal original contiendrait de longues pauses.

Lorsque le mode *Fader* est sélectionné (voir la section 3.3.4), le cadre central permet de choisir le contrôleur continu du rythme syllabique ; le menu déroulant *interface* permet de choisir entre une tablette graphique (*Tablet*), un contrôleur MIDI continu (*Fader Solo*), deux contrôleurs MIDI continus (*Fader Duo*) ou un contrôleur externe dont la valeur est émise sur le réseau local (UDP). Il permet ensuite de sélectionner le mode de contrôle ; le menu déroulant *mode* permet de choisir entre *T-V*, *V-V* et *V-V Return*. Le paramètre *Fader active area* permet de définir la plage active du contrôleur continue. Ici, seulement 70% de la plage du potentiomètre est active. Enfin, le paramètre *transient temporal correction* permet d'activer ou de désactiver la correction temporelle des phases transitoires. La fenêtre *Advanced Settings*, présentée FIGURE 5.23, permet de sélectionner la vitesse de lecture maximale des transitoires, et la vitesse de rattrapage lors des parties vocaliques.

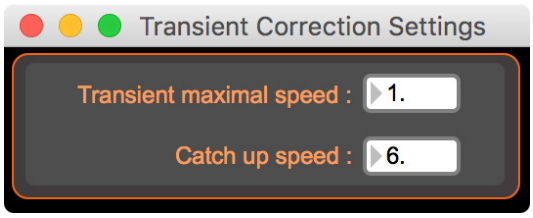


FIGURE 5.23 – *Fenêtre de paramétrage de la correction temporelle des transitoires.*

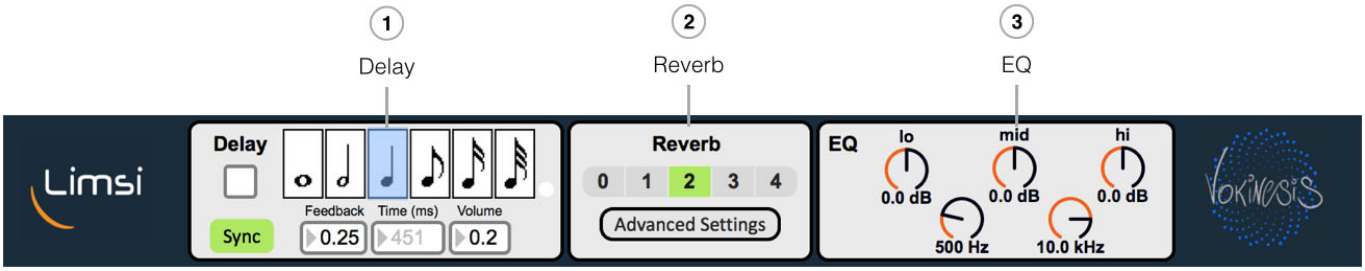


FIGURE 5.24 – *Zone des effets dans la fenêtre principale de Vokinesis*

Retournons à présent sur la FIGURE 5.21. Le cadre **Presets** (8) permet d'enregistrer les réglages spécifiques grâce au bouton *Save Preset*. Cela aura pour effet de créer dans le dossier `data` un fichier `.preset` qui sera assigné à l'élément sélectionné dans le tableau du projet. À chaque fois que cet élément sera sélectionné à nouveau, tous les réglages effectués dans la fenêtre principale pourront alors être restitués. Le bouton *Restore Saved* permet de restituer les derniers réglages sauvegardés, et le bouton *Restore Default* permet de restituer les réglages originaux.

Enfin, le cadre **Recording** permet d'enregistrer le signal de synthèse dans un fichier audio (bouton *Start audio recording* (6)), mais également les données de contrôle (bouton *Start data recording* (7)). Pour plus d'information sur l'enregistrement des données de contrôle, se référer à la section 5.5.4.

#### 5.5.2.3 Zone des effets

La zone des effets permet d'appliquer des effets audio aux signaux de synthèse. Celle-ci comporte tout d'abord un *Delay* (1), qui correspond à un effet d'écho. Le choix de la figure de note permet de choisir l'intervalle de temps qui sépare deux échos, selon la valeur indiquée par le paramètre *Tempo* dans la zone de signal (FIGURE 5.19, (16)). Elle comporte également une *Reverb*, qui correspond à un effet de réverbération de salle. Ici, il est possible de sélectionner le niveau de réverbération, de 0 (pas de réverbération) à 5 (très forte réverbération). Ces options agissent sur le paramètre de la fenêtre de configuration de la reverb présentée FIGURE 5.25, qui s'ouvrira lors d'un clic sur le bouton *Advanced Settings*. Cette effet peut être retrouvé dans les exemples fournis par Max/MSP 6, patcher `reverb_example.maxpat`. Enfin, elle fournit un *EQ*, ou égaliseur, qui permet d'amplifier ou d'atténuer certaines bandes de fréquences.

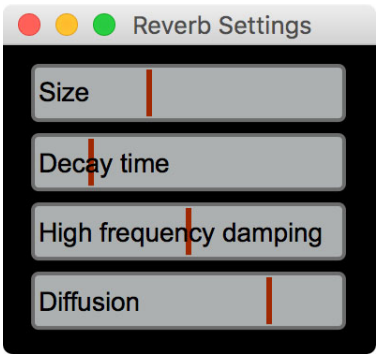


FIGURE 5.25 – *Fenêtre de réglages avancés de l'effet de Reverb*

#### 5.5.2.4 Marquages périodique et phonétique

Cette section présente la façon dont le marquage périodique et phonétique peut s'effectuer à la main, directement sur la forme d'onde du signal original. Nous rappelons ici que nous préconisons l'utilisation d'algorithmes de détection des périodes, telles que ceux fournis par le logiciel *Praat* [Boersma *et al.* 2002]. De même pour l'étiquetage des phonèmes de longs signaux de parole, nous préconisons l'utilisation du plugin *easyalign* [Goldman 2011], destiné à *Praat*. Cependant, ces algorithmes n'étant pas infaillibles, une correction manuelle est généralement nécessaire pour l'obtention d'un signal de synthèse de qualité. De plus, pour des signaux courts, il est sans doute plus rapide d'effectuer directement l'étiquetage phonétique de façon manuelle. Nous avons donc mis en place les procédures que nous présenterons ci-dessous.

Sur la zone d'affichage de la forme d'onde du signal original (FIGURE 5.18), l'option *Edit* permet d'activer l'édition de l'étiquetage des phonèmes et du marquage périodique. Si la durée affichée est supérieure à $250ms$, l'étiquetage des phonèmes pourra alors être édité. Dans le cas contraire, ce seront les marqueurs périodiques qui pourront être modifiés. Ils seront indiqués par des croix, comme le montre la FIGURE 5.26. Toute modification des étiquettes de phonèmes ou des marqueurs périodiques sera perdue si l'utilisateur omet de cliquer sur les boutons *Save Phonemes* et *Save Pitch Marks* avant de sélectionner un autre fichier audio dans le tableau du projet.

Commençons par l'étiquetage des phonèmes (durée affichée $> 250ms$). Un clic sur une délimitation de phonèmes permet de la faire glisser, et de changer sa position. Un clic au centre de deux délimitations permet de sélectionner un phonème : son caractère sera affiché en blanc, et il sera possible de le modifier avec les touches du clavier, en utilisant l'alphabet phonétique SAMPA. Un clic accompagné de la touche ALT enfoncée permet d'ajouter ou de supprimer une délimitation de phonèmes.

Le marquage périodique manuel s'effectue de façon relativement similaire. Un clic sur un marqueur périodique permet de le déplacer. Si deux marqueurs périodiques sont placés au même endroit, alors l'un des deux sera supprimé. Sur la FIGURE 5.26, les marqueurs périodiques verts sont assignés aux parties voisées du signal original,

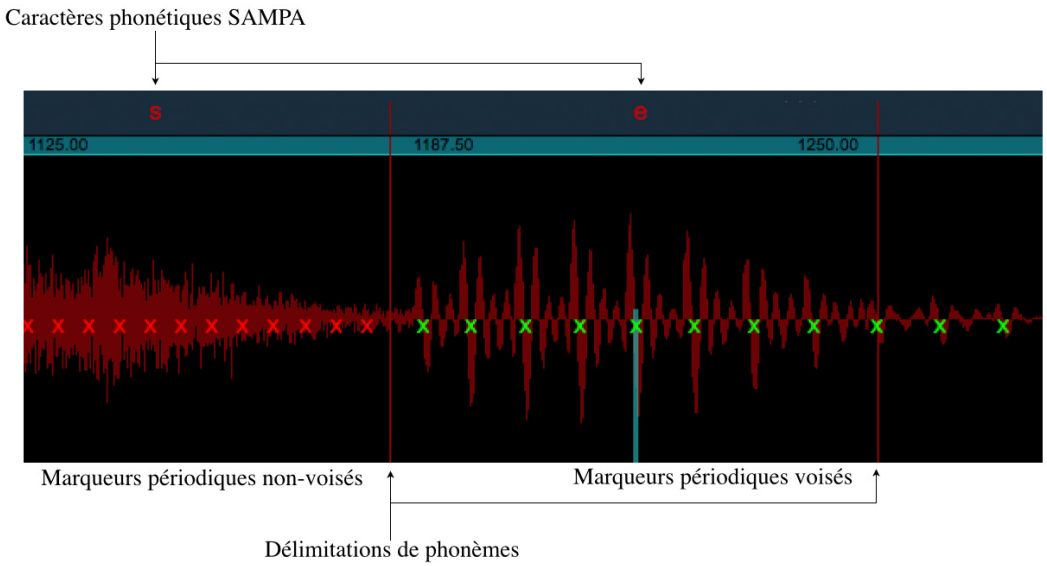


FIGURE 5.26 – *Édition des marqueurs périodiques. La durée affichée est inférieure à 250ms, ce qui permet l'édition de ces marqueurs. Les croix vertes sont assignées aux parties voisées, les croix rouges aux parties non-voisées.*

les marqueurs rouges aux parties non-voisées. L'un des soucis majeurs de qualité est lié à une détection imparfaite de frontières entre parties voisées et non voisées (VNV). Ici, il est possible de changer l'état VNV d'un marqueur en cliquant dessus tout en maintenant la touche CTRL enfoncée. Enfin, il est possible de rajouter ou de supprimer un marqueur en cliquant avec la touche ALT enfoncée. L'état VNV d'un nouveau marqueur périodique sera identique à celui du marqueur périodique précédent.

### 5.5.3 Paramétrages globaux : configuration des contrôleurs

Dans la fenêtre principale (FIGURE 5.18), le cadre *Hardware Settings* contient quatre boutons. Chacun de ces boutons permettra d'ouvrir une fenêtre de paramétrage des contrôleurs. Chacune de ces fenêtres sera présentée dans cette section. Les réglages qui y sont effectués sont des paramétrages globaux : ils sont communs à tous les éléments du tableau du projet.

#### 5.5.3.1 Tablette graphique

Un clic sur le bouton *Tablet* ouvrira la fenêtre représentée FIGURE 5.27. À la gauche de cette figure, les paramètres qui peuvent être contrôlés par la tablette graphique sont listés. Chaque paramètre vocal (de *Pitch* à *Tenseness*) et de contrôle temporel (de *Speed* à *Restart*) peut se voir assigner un attribut de la tablette graphique, indépendamment pour les deux voix de synthèse. Les attributs de la tablette qui peuvent être sélectionnés dans les menus déroulants sont pré-

sentés FIGURE 5.28. Les boutons *Switch* permettent d'inverser la plage de contrôle du paramètre correspondant. Un clic sur le bouton *Save Settings* créera un fichier `tabletSettings.preset` dans le dossier `data`. Ainsi, dès que le projet sera chargé, les réglages effectués seront restitués.

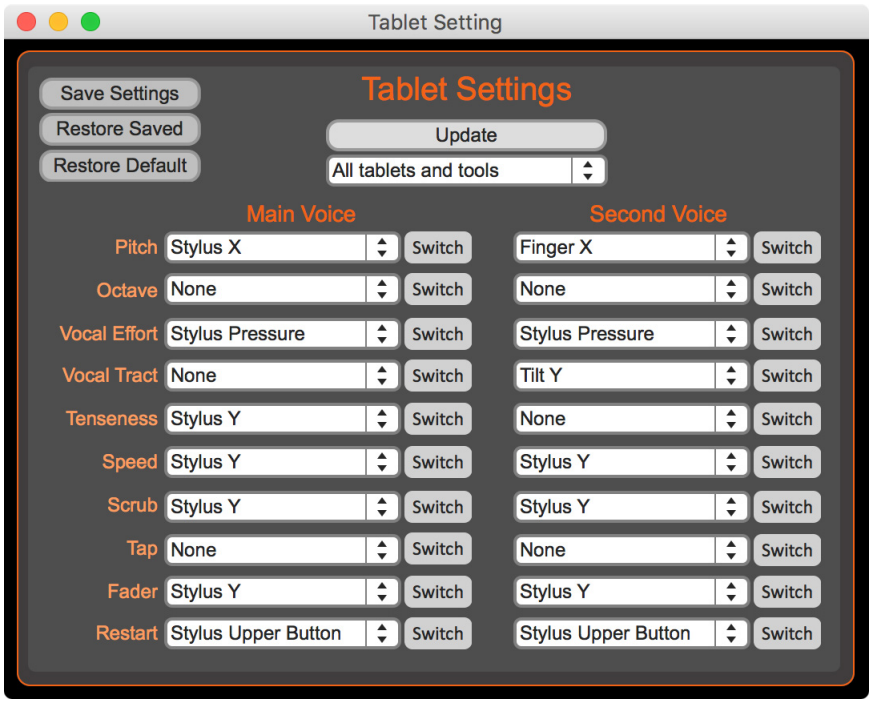


FIGURE 5.27 – *Fenêtre de configuration de la tablette graphique*

#### 5.5.3.2 Contrôleurs MIDI

Un clic sur le bouton *MIDI Controls* dans le cadre *Hardware Settings* de la fenêtre principale (FIGURE 5.18) ouvrira la fenêtre de configuration des contrôleurs MIDI, présentée FIGURE 5.29. La seule différence dans la liste des paramètres contrôlables par rapport à celle présentée dans la FIGURE 5.27 pour la tablette graphique se tient au niveau des faders. En effet, seul le mode *Fader Solo* peut être utilisé à la tablette graphique, alors que les interfaces MIDI permettent également d'utiliser le mode *Fader Duo*.

Chacun de ces paramètres peut se voir assigner un contrôle continu (cc) d'une interface MIDI. Lorsqu'un cc est actif, par exemple si la position d'un potentiomètre MIDI est modifiée, alors son numéro de canal (*MIDI Channel*), son numéro de contrôle continu (*cc*) et sa valeur (*Value / Switch*) seront affichés dans les zones grises correspondantes. Il suffira ensuite de cliquer sur le bouton *set* d'un des paramètres vocaux pour lui assigner le dernier cc actif. Le fait de cliquer sur le bouton *set* coche automatiquement la case *Active* correspondante, ce qui activera le contrôle du paramètre en question par le cc assigné. Comme indiqué dans la section 5.3, les contrôles MIDI prennent la priorité sur les autres interfaces. Par exemple, si la

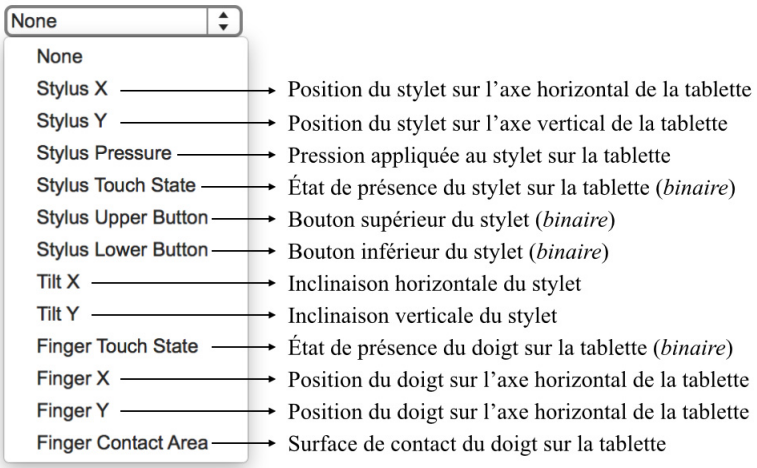


Figure 5.28 – *Attributs de la tablette graphique. Si un attribut est indiqué comme étant binaire, alors il ne peut prendre que deux valeurs : 0 ou 1. Tous les autres attributs peuvent prendre n'importe quelle valeur comprise entre 0 et 1.*

case *active* du paramètre *Vocal effort* est cochée, alors la tablette graphique ne sera plus en mesure de contrôler l'effort vocal. Cela permet d'éviter les conflits entre les différentes contrôleurs.

Un clic sur le bouton *Save Settings* créera un fichier `midiSettings.preset` dans le dossier `data`. Ceci permettra lors de la prochaine ouverture du projet de restituer les réglages préalablement effectués.

#### 5.5.3.3 Clavier MIDI

Un clavier MIDI peut être utilisé pour contrôler d'une part la hauteur et l'effort vocal du signal de synthèse, qui correspondront à la note jouée et à sa vélocité (voir la section 4.2.3), et d'autre part le rythme syllabique en mode *Tap* (voir la section 3.3.2) si cela a été configuré dans les cadres centraux de la zone de contrôle (Figure 5.21 et 5.22). La configuration d'un clavier MIDI s'effectue dans la fenêtre présentée Figure 5.30, qui s'ouvrira si le bouton *MIDI Keyboard* dans le cadre *Hardware Settings* de la zone de contrôle (Figure 5.21) est cliqué.

La fenêtre de configuration du clavier MIDI (Figure 5.30) est divisée en deux parties : la partie *MIDI interface*, qui permet de configurer le comportement du clavier MIDI, et la partie *Configuration* qui permet de configurer la façon dont les paramètres de hauteur et d'effort vocal réagiront aux commandes du clavier MIDI.

Dans la partie *MIDI interface*, le menu déroulant permet de sélectionner un clavier MIDI parmi la liste des claviers connectés à l'ordinateur. Le bouton *Data reception* clignote lorsqu'une commande envoyée par le clavier MIDI sélectionné est reçue. Le paramètre central de *Split keyboard*, que nous appellerons *note de séparation*, permet de séparer le clavier en deux. Si l'option *All notes disabled below* est choisie, alors toutes les notes inférieures à la note de séparation seront désactivées.

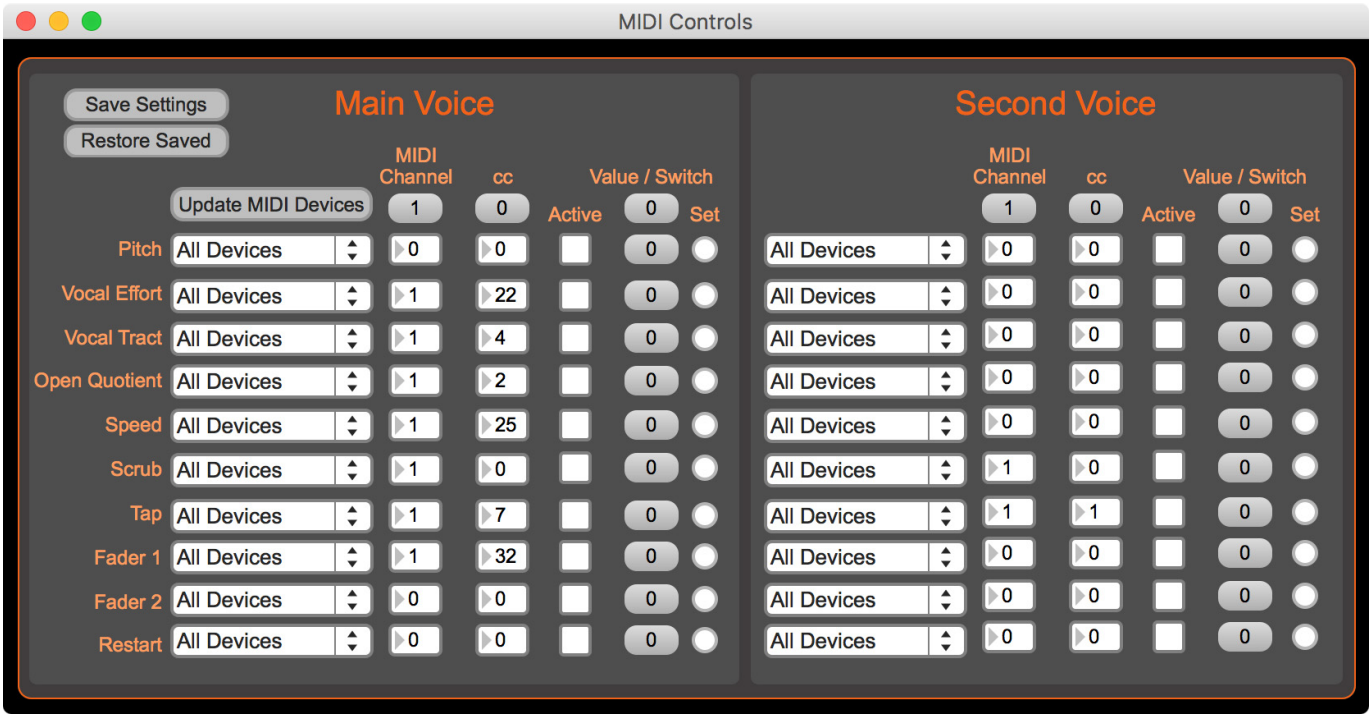


FIGURE 5.29 – *Fenêtre de configuration des contrôleurs MIDI*

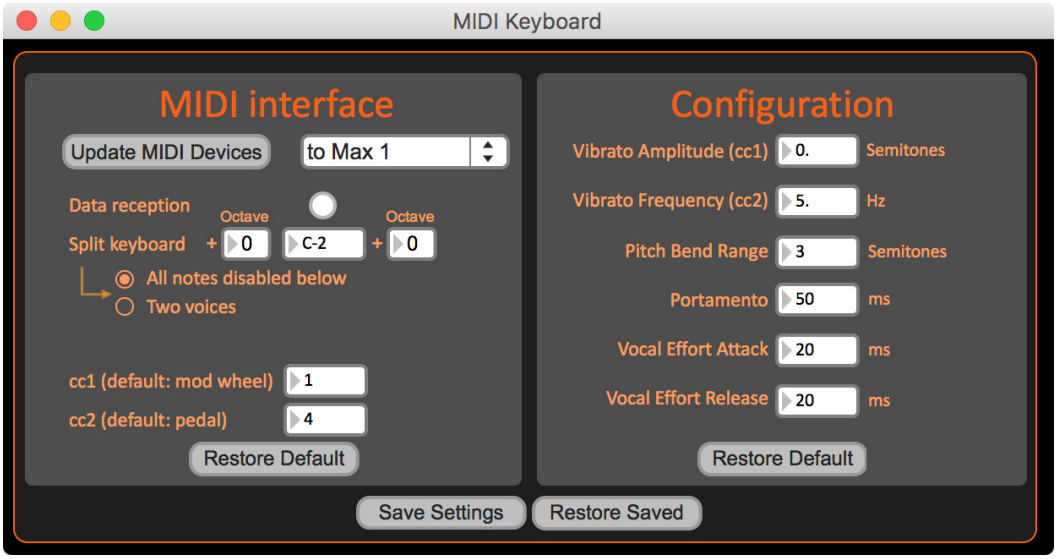


FIGURE 5.30 – *Fenêtre de configuration d'un clavier MIDI*

Cela permet à l'utilisateur de contrôler un autre synthétiseur avec la partie gauche du clavier MIDI sélectionné. Si l'option *Two voices* est sélectionnée, alors toutes les notes inférieures à la note de séparation seront assignées à une deuxième voix de synthèse. Il sera alors possible d'augmenter ou de diminuer l'octave de chacune des voix grâce aux paramètres *Octave* qui entourent la note de séparation. Enfin, le numéro assigné aux paramètres *cc1* et *cc2* permettra d'assigner l'un des cc du clavier MIDI à l'amplitude (*cc1*) et à la fréquence (*cc2*) du vibrato.

Dans la partie *Configuration*, les paramètres *Vibrato Amplitude* et *Vibrato Frequency* permettent de régler les valeurs maximales que pourront prendre l'amplitude et la fréquence du vibrato. Le paramètre *Pitch Bend Range* permet de régler l'étendue de contrôle de la molette de Pitch-bend du clavier MIDI. Le paramètre *Portamento* permet de définir la durée que prendra la fréquence fondamentale à atteindre une seconde note lorsqu'une première est maintenue. Enfin, les paramètres *Vocal Effort Attack* et *Vocal Effort Release* permette de définir les durées d'attaque et de relâchement de l'effort vocal. Pour plus d'informations sur le contrôle de la hauteur avec un clavier MIDI, se référer à la section 4.2.3.

Un clic sur le bouton *Save Settings* créera un fichier `keyboardSettings.preset` dans le dossier `data`. Ainsi, les réglages effectués seront restitués à chaque ouverture du projet.

#### 5.5.3.4 Entrée audio

Si l'option *Audio Input* a été sélectionnée dans les cadre centraux de la zone de contrôle (FIGURE 5.21 et 5.22), une entrée audio peut être utilisée pour contrôler d'une part la hauteur et l'effort vocal du signal de synthèse, et d'autre part le rythme syllabique en mode *Tap* (voir la section 3.3.2) .

Le paramètre *Silent State Acquisition* permet de régler l'intensité sonore minimale de l'environnement : lorsqu'il est actif, sa valeur conserve la moyenne de l'intensité du signal reçu par l'entrée audio. le paramètre *Maximal Vocal Effort Acquisition* permet de régler l'intensité maximale que pourra atteindre le signal d'entrée. Lorsqu'il est activé, ce paramètre conserve la valeur maximale de l'intensité du signal d'entrée. Ces paramètres peuvent également être réglés à la main. Ils servent d'une part à obtenir un seuil d'intensité pour le mode *Tap*, et d'autre part à définir une plage de contrôle de l'effort vocal par l'intensité du signal d'entrée. Le paramètre *Octave* permet de définir la fréquence fondamentale du signal de synthèse, qui correspondra à celle du signal d'entrée augmentée du nombre d'octaves indiqué (ou diminuée si ce nombre est négatif).

### 5.5.4 Vokinesis en tant qu'outil expérimental

En tant qu'instrument parlant et chantant permettant d'externaliser la production vocale, Vokinesis peut être utilisé comme un outil expérimental. Pour ce faire, nous avons développé deux interfaces graphiques. L'une d'entre elles, représentée FIGURE 5.32, est destinée à être présentée aux sujets d'une expérience. Elle est acces-

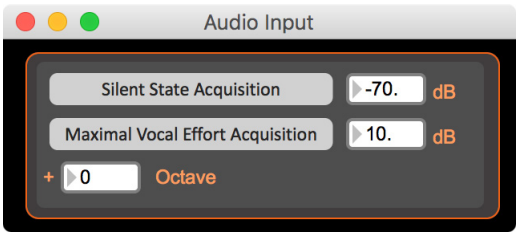


FIGURE 5.31 – *Fenêtre de configuration de l'entrée audio*

sible en cliquant sur le bouton *Subject Page* dans l'éditeur de projets (FIGURE 5.17). L'autre, présentée FIGURE 5.33, et destinée à l'opérateur, permet de mettre en place une procédure expérimentale. Elle est accessible grâce au bouton *Task Setup*.

Penchons-nous sur la fenêtre du sujet (FIGURE 5.32). Tout d'abord, l'opérateur entre le nom du sujet dans la zone de texte correspondante. Lorsqu'il cliquera sur le bouton *DÉMARRER LE TEST*, ceci aura pour effet de créer un dossier portant le nom du sujet dans le dossier `recordedData`. Le sujet pourra alors démarrer le test. Ici, nous pouvons voir que le sujet aura 27 performances à produire. À chaque performance, le sujet pourra écouter la phrase originale en cliquant sur le bouton correspondant. Il devra modifier cette phrase avec les interfaces qui lui seront fournies et selon les consignes qui lui seront données. Il devra éventuellement enregistrer sa propre voix si cela lui aura été demandé. La procédure d'enregistrement est identique pour la synthèse et pour la voix naturelle : un clic sur le bouton *enregistrer l'essai* (pour la synthèse) ou *Enregistrer voix* (pour le naturel) démarre l'enregistrement, un deuxième clic y met fin. Une fois un essai ou une voix enregistrés, ceux-ci peuvent être réécoutés, réenregistrés ou validés. Une fois l'essai et la voix validés, l'utilisateur pourra passer à la phrase suivante en cliquant sur le bouton correspondant, qui sera alors dégrisé.

Passons à présent à la partie gauche de la fenêtre de configuration de l'expérience (FIGURE 5.33). Elle permet d'abord de configurer la façon dont les éléments du tableau du projet seront sélectionnés par le sujet. Si l'option *Ordered Selection* est choisie, alors le sujet se verra présenter les éléments du tableau dans l'ordre dans lequel ils sont organisés, le nombre de fois indiqué par le paramètre *times*. Si l'option *Random selection* est choisie, alors le sujet se verra présenter chaque élément du tableau du projet le nombre de fois indiqué par le paramètre *times* dans un ordre aléatoire. Le bouton *Restart File Selection* permet de redémarrer la procédure de sélection des éléments du tableau. La partie gauche de la fenêtre permet également de configurer les touches du clavier qui permettront au sujet d'effectuer les actions indiquées. L'appui sur une touche du clavier enclenche l'affichage du caractère et du cade ASCII correspondants sous les attributs *Character* et *ASCII Code*, et un appui sur le bouton *set* assigne la dernière touche pressée à l'action indiquée. Ceci permet d'éviter au sujet d'avoir à se servir de la souris lors d'une procédure expérimentale.

La partie droite de la fenêtre de configuration de l'expérience permet de configurer les données à enregistrer. Un fichier par paramètre coché sera enregistré pour

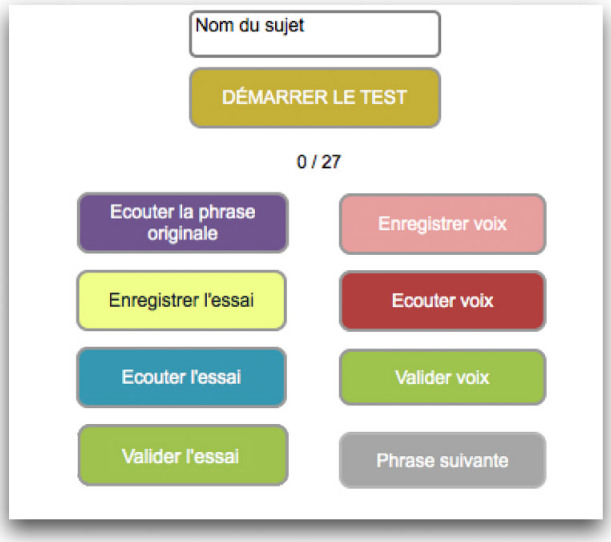


FIGURE 5.32 – *Fenêtre présentée aux sujets lors d'une expérience.*


FIGURE 5.33 – *Fenêtre de configuration d'une procédure expérimentale. La partie de gauche permet de configurer la façon dont les éléments du tableau du projet seront sélectionnés par les sujets. La partie de droite permet de configurer les paramètres à enregistrer.*

chaque essai dans le dossier portant le nom du sujet, lui même situé dans le dossier `recordedData`. Le nom du fichier enregistré aura la forme suivante :

```
[nom original] _ [emplacement dans le tableau] . [extension explicite]
```

Par exemple, si le fichier original s'appelle « *oui.wav* » et qu'il est situé dans la première case du tableau, alors le fichier qui contiendra les données relatives au contrôle de la hauteur portera le nom suivant : `oui1.pitch`. Enfin, si l'option *Record voice* est activée, le sujet devra, en plus de la synthèse, enregistrer sa propre voix à chaque essai, et si l'option *Metronome* est activée, il entendra un métronome lors de chaque enregistrement (synthèse et naturel).

## 5.6 Futurs développements

Dans ce chapitre, nous avons présenté Vokinesis, un système permettant la modification performative de paramètres suprasegmentaux de signaux de voix préenregistrés. Les interfaces graphiques de configuration matérielle que nous avons développées offrent à Vokinesis une modularité certaine. Cependant, celle-ci comporte encore certaines limites. Tout d'abord, il serait utile de permettre aux interprètes de régler la plage de contrôle des données gestuelles indépendamment pour chaque paramètre vocal auquel ils sont assignés. En effet, cela permettrait de rendre la stratégie de mapping divergent plus puissante : un utilisateur pourrait par exemple assigner le contrôle de l'effort et de la tension au même contrôleur et rendre la plage de contrôle moins importante pour la tension. Par ailleurs, la prise en charge des signaux UDP n'a pour l'instant été utilisée que pour le mode *Fader* contrôlé par Leap Motion. Il serait sans doute utile de permettre un paramétrage des signaux UDP similaire à celui des signaux des contrôleurs MIDI, et de permettre ainsi le contrôle de tous les paramètres vocaux par des signaux UDP. D'autre part, l'external `s2m.wacom`, permettant de récupérer les données du stylet sur la tablette graphique, ne fonctionne que sur Mac, et constitue le seul frein à une prise en charge sur PC. Nous nous sommes brièvement penchés sur cette question au début de ce travail de thèse, et la complexité du fonctionnement des interfaces de type tablette graphique sous Windows et le temps imparti nous en ont dissuadés. Enfin, le système est aujourd'hui limité à deux voix de synthèse. Les futurs développements devraient réfléchir à une façon de choisir le nombre de voix jouées. Ainsi, avec l'utilisation de PMC, il serait possible de jouer jusqu'à 15 voix de manière simultanée. Il faudrait également permettre l'enregistrement des données de contrôle de chacune des voix de synthèse.

L'utilisation d'une entrée audio comme interface de contrôle est pour l'instant très prototypique, et pourrait subir des améliorations : remplacer le seuil d'intensité par une détection d'attaque et de relâchements pour le contrôle temporel, trouver des méthodes de contrôle continu du rythme en convertissant le signal d'entrée en un signal de fader (suivi d'intensité, suivi de formants pour un signal vocal...)

Pour la programmation des externals, nous avons fait le choix du langage Java

en raison de la rapidité de développement qu'il permet comparé au C ou au C++. Cependant, nous avons remarqué de fortes baisses de performances lorsque de nombreuse données sont échangées entre l'external et l'environnement Max/MPS. C'est pourquoi nous avons fait en sorte de permettre l'activation ou la désactivation de l'affichage de l'instant cible lors de la synthèse : la voix de synthèse est assez retardée par rapport aux gestes de contrôle lorsque l'affichage est activé. D'après cette discussion [3], il semblerait que l'utilisation du langage C (ou C++) permettrait d'améliorer ces performances.

Lors de ces travaux de développement, nous avons mis en place le prototype d'une version « synthèse par concaténation en temps-réel » de Vokinesis. Dans cette version, une utilisatrice peut d'une part entrer un texte et en contrôler le rythme syllabique en mode *Tap* ou *Fader*, et d'autre part contrôler directement l'articulation avec les touches d'un clavier d'ordinateur. Le principe est de remplacer l'utilisation d'un signal d'entrée unique par celle d'une base de donnée étiquetée destinée à la synthèse par concaténation (la base dont nous disposons est la même que celle utilisée par [Ardaillon *et al.* 2015], et a été enregistrée à l'IRCAM dans le cadre du projet ANR ChaNTeR). Les futurs travaux devraient également se pencher sur la recherche de méthodes de contrôle permettant d'improviser un texte, ou du moins des syllabes, sans avoir à contrôler le séquencement de tous les phonèmes.

Dans le chapitre suivant, nous verrons comment Vokinesis peut être utilisé comme instrument chanteur, mais également comme un outil de création sonore unique en son genre.

3. https://cycling74.com/forums/java-javascript-or-c/

Chapitre 6

# Chanter avec Vokinesis, et au-delà...

**Sommaire**



Ce chapitre présente un aperçu des utilisations musicales que peut permettre Vokinesis. Nous y présentons dans un première section les différentes représentations de l'ensemble de voix de synthèse Chorus Digitalis qui en ont fait usage, puis nous montrerons les possibilités de créations sonores qu'offre ce logiciel, pouvant aller parfois bien au-delà du chant.

## 6.1 Représentations du Chorus Digitalis

Le Chorus Digitalis est un ensemble de voix de synthèse qui a vu le jour avec l'apparition du Cantor Digitalis. L'idée était de réunir des musiciens intéressés par le contrôle performatif de la voix de synthèse et d'explorer les possibilités musicales offertes par ce nouvel instrument. Un certain nombre de représentations ont été données dans diverses situations, et avec différentes formations (la page internet [1] contient des vidéos ainsi qu'une liste des représentations publiques de l'ensemble, et fournit des informations sur les différentes formations). La pratique du Chorus Digitalis a permis une exploration profonde des possibilités offertes par le modèle vocal et les méthodes de contrôle du Cantor Digitalis, et a été une étape importante pour la finalisation du logiciel [Perrotin 2015]. L'intégration de Vokinesis au

1. https://cantordigitalis.limsi.fr/chorusdigitalis_fr.php

Chorus Digitalis est alors apparue essentielle pour son évaluation et pour son développement. Dans cette section, nous présenterons les différentes représentations du Chorus Digitalis lors desquelles nous avons fait usage de Vokinesis.

### 6.1.1 CURISOTas 2015 et JAP-TALN-RECITAL 2016

La toute première représentation publique de Vokinesis a eu lieu lors du festival arts/sciences CURIOSITas 2015. Il portait à l'époque encore le nom de son prédécesseur, Calliphony, mais permettait déjà le contrôle binaire du cadre rythmique (voir la section 3.3.2). C'était un concert de l'ensemble Chorus Digitalis, qui était composé pour lors de Christophe d'Alessandro, Boris Doval, Hélène Meynar, Anelies Bratford, Lionel Feugère, Olivier Perrotin et moi-même. La vidéo `Ex13` présente deux extraits de cette représentation, que nous commenterons ci-dessous, en conservant l'ordre de l'exemple (nous garderons dans le texte l'appellation Vokinesis).

Le premier morceau musical faisant usage de Vokinesis était une reprise de la chanson *Circle Song* de Bobby Mc Ferrin. La formation était composée pour ce morceau d'une percussionniste corporelle, de 5 Cantor Digitalistes choristes et d'un Vokinésiste soliste. Ce dernier utilisait un fichier original comportant l'enchainement d'onomatopées suivant : [bambədimbigorammam]. Ce signal était bouclé sur les consonnes finale et initiale (la section 3.3.8 présente la façon dont le bouclage fonctionne), et l'interprète improvisait le rythme et la mélodie. Ce morceau a permis de démontrer la capacité d'un Vokinésiste à se synchroniser avec une percussionniste.

Les capacités de synchronisation entre plusieurs Vokinésistes ont été démontrées lors de la chanson « *Le lion est mort ce soir* ». L'ensemble était alors composé d'une cajoniste, de trois Cantor Digitalistes et de trois Vokinésistes. Dans l'exemple, nous avons sélectionné le seul passage lors duquel les Vokinésistes chantaient en trio, qui ne comportait pas d'accompagnement musical.

Ce concert a fait l'objet d'une première évaluation positive de notre méthode de contrôle binaire du cadre rythmique en situation musicale. Nous avons également donné une représentation assez similaire lors de la conférence JAP-TALN-RECITAL 2016, dont un extrait peut être visionné ici [2]. Dans cet extrait, l'interprète de droite contrôle les liaisons rythmiques des paroles avec une pédale. Pour les concerts suivants, nous avons utilisé des signaux originaux de bien meilleure qualité, autant du point de vue de la production vocale que de l'enregistrement, ce qui nous a permis d'améliorer la qualité de la synthèse de Vokinesis.

### 6.1.2 Festival aCROSS 2017 et Colloque Voix et Psychanalyse 2017

En 2017, nous avons effectué deux représentations dans le cadre du festival aCROSS, et une autre pour le $8^e$ colloque voix et psychanalyse. Une version récente de Vokinesis y était utilisée. Du point de vue du style musical, nous nous sommes tournés vers un côté plus expérimental que pour les précédentes représentations. Les deux représentations peuvent être visualisées dans les vidéos `Ex14`, `Ex15`

2. https://youtu.be/RXR9ivA-h6w

et `Ex16`.

La première représentation pour le festival aCROSS a eu lieu le 5 mai 2017 au conservatoire de Vitry-sur-Seine. L'ensemble était composé de quatre musiciens : Christophe d'Alessandro, Boris Doval, Victor Wetzel et moi-même. La pièce comportait deux actes, dont l'un faisait exclusivement usage de Vokinesis. C'était une improvisation préparée basée sur le conte algérien *Brirouch*. La phrase originale « l'histoire de Brirouch » était utilisée pour une introduction et un final musicaux, chantés sur l'air de *miserere* de Lotti. Le texte de Brirouch était ensuite récité : un Vokinésiste jouait le rôle de la mère, et les trois autres se partageaient les différents personnages restants. Les Vokinésistes jouaient en mode *Speed*, *Tap* ou *Fader* selon l'intention musical désirée. Des réglages spécifiques avaient été effectués pour chaque personnage. Nous pouvons par exemple relever le cas du rat, pour lequel le conduit vocal fût fortement rétréci. Les gestes de contrôle étaient également adaptés à chaque personnage : des mouvements sinueux pour l'eau, tranchants pour le couteau, etc... Un exemple intéressant que nous pouvons donner ici concerne la phrase « passez-moi la lame ! » prononcée par le forgeron. En effet, une erreur d'analyse avait étiqueté certaines voyelles comme non-voisées. Notre système de synthèse y appliquait la technique d'allongement des signaux non-voisés présentée section 2.2.7, résultant alors en un signal perçu à la fois comme chuchoté et fortement crié. Cet acte a constitué une bonne démonstration des capacités de transformation vocale offertes par Vokinesis, et la partie chantée a été un preuve de plus des possibilités de synchronisation entre plusieurs Vokinésistes. Le deuxième acte faisait exclusivement usage du Cantor Digitalis, et consistait en une improvisation musicale. L'idée était cette fois-ci de démontrer les capacités du Cantor Digitalis à produire des sons vocaux ou non, en poussant ses paramètres dans leurs valeurs extrêmes.

Les deux dernières représentations du Chorus Digitalis se sont déroulées le 14 mai 2017 à l'église Sainte-Élisabeth-de-Hongrie pour le festival aCROSS de nouveau (sans Victor Wetzel), puis le 10 juin 2017 à l'université Paris-Diderot pour le 8$^{e}$ colloque voix et psychanalyse (sans moi). L'ensemble s'est alors entouré d'un chanteur, Robert Expert [3], contre-ténor et professeur de chant lyrique au Conservatoire National Supérieur de Musique de Lyon. Nous avons alors adapté les actes de la représentation du 5 mai. Le rôle de la mère dans *L'histoire de Brirouch* lui a été confié. Il chantait le texte en improvisant et en adaptant ses productions au personnage auquel il s'adressait. Lors de l'acte d'improvisation, il expérimentait des placements de voix particuliers en essayant d'intégrer ses sons vocaux aux sons méta-vocaux produits par le reste de l'ensemble. Le 14 mai, cet acte s'est déroulé avec la participation d'Olivier Innocenti [4] équipé de son *EigenHarp*, au centre de la FIGURE 6.1.

En plus des actes que nous venons d'évoquer et qui avaient déjà été joués le 5 mai, la pièce comportait un acte musical préparé, basé sur une adaptation d'un texte de Kafka, *Devant la loi*. L'adaptation du texte et la composition musicale

3. www.robertexpert.net
4. www.olivierinnocenti.com/

Figure 6.1 – *Acte d'improvisation lors de la représentation du 14 mai à l'église Sainte-Élisabeth-de-Hongrie. De gauche à droite : Samuel Delalez (Cantor Digitalis), Boris Doval (Cantor Digitalis), Olivier Innocenti (EigenHarp), Robert Expert (Voix), Christophe d'Alessandro (Cantor Digitalis).*

ont été effectuées par Christophe d'Alessandro. Cet acte comportait deux Vokinésistes, un Cantor Digitaliste et un chanteur. Les Vokinésistes disposaient de voix pré-enregistrées par Robert Expert dans sa voix de baryton et de contre-ténor. Lors de la représentation, la voix humaine jouait le rôle des deux personnages (un portier et un homme du pays) et les voix re-synthétisées se partageaient la narration. Le Cantor Digitalis jouait ici un rôle d'accompagnement sonore. Encore une fois, les Vokinésistes contrôlaient certains passages en mode *Speed*, d'autres en mode *Tap* ou *Fader*. C'est la première pièce du Chorus Digitalis lors de laquelle le contrôle polyphonique individuel fût introduit. Sur la $9^e$ page de la partition, les deux Vokinésistes jouent des polyphonies en synchronie.

### 6.1.3 Retours de Robert Expert

Lors de la représentation du 10 juin ($8^e$ colloque Voix et Psychanalyse, université Paris-Diderot), dont la formation est représentée Figure 6.2, Robert Expert a pu nous faire part de ses impressions avant et après le concert. Les paragraphes que nous fournissons ci-dessous en sont une transcription, et sont extraits d'une future publication dans les actes de ce colloque, qui seront publiés dans le courant de l'année 2018 :

**Avant le concert**

Robert Expert : Mon expérience avec ces instruments est très récente et mon vécu par rapport à cette expérience également et donc je ne suis qu'au stade des questions. Mais je peux apporter un tout petit témoignage. C'est donc ma voix qui a été enregistrée, qui est traitée par les instruments et que vous allez entendre dans la deuxième pièce. Mais ma voix est multipliée par trois : je me retrouve avec quatre fois ma voix. C'est étrange, et ils avaient annoncé la couleur et donc je m'étais préparé psychologiquement, j'ai survécu à l'opération, et finalement assez facilement. La première chose qui m'a frappée dans le travail avec Samuel, Victor, Christophe et Boris c'est que je me suis retrouvé vraiment avec des instruments, avec une pratique d'instrumentistes. Et il est sûr qu'il va falloir répéter beaucoup pour continuer ce travail là. J'ai ressenti très fortement une frustration de ne pas avoir assez répété avec vous, et vous aussi certainement, parce qu'on sent bien qu'il


FIGURE 6.2 – *Représentation lors du Colloque Voix et Psychanalise à l'université Paris-Diderot. De gauche à droite : (Vokinesis et Cantor Digitalis) Boris Doval, Victor Wetzel, Christophe d'Alessandro, (Voix) Robert Expert.*

y a des possibilités avec ces instruments, qui sont absolument immenses. Le champ d'action de ces instruments me paraît très important et donc le champ d'erreur est également important, c'est ça qui en fait le prix. C'est-à-dire que ces instruments sont difficiles à jouer, ce qui m'a paru extrêmement précieux, parce que du coup, avec beaucoup d'entraînement et de répétitions, je pense qu'on a devant nous des grandes possibilités. Autre chose que je voulais dire, c'est pour le peu de pratique que nous avons eue, je n'ai pas du tout eu la sensation, en tant qu'être vivant capable d'une infinité de possibilités par rapport à ma voix, d'être inférieur ou supérieur, mais d'être dans une situation de jeu musical. Ce qui est sûr c'est que ça m'a ramené à mes limites, puisqu'il y a quelque chose de l'ordre d'une prothèse dans ce que vous avez proposé. Je suis limité dans l'aigu, je suis limité dans le grave, je suis limité dans la puissance, je suis limité dans tout un tas de choses : c'est le complexe du chanteur face à l'acte chanté, je ne vais pas vous en faire une dissertation c'est assez facile à imaginer, vous êtes tous passés par là. Mais là, tout d'un coup, j'ai des appendices dans tous les sens et je deviens incroyablement puissant performant, aigu, grave, vibrant, pas vibrant : tous mes fantasmes vocaux au fond semblent se réaliser.

**Après le concert**

Robert Expert : Si je peux encore donner un témoignage sur ce qu'on vient de faire à l'instant, j'aimerais parler de ce qu'on appelle une « improvisation générative », le premier morceau. J'en ai fait avant de vous rencontrer [les chanteurs synthétiques], pas énormément mais j'en ai fait et puis j'en ai fait faire à mes étudiants, comme partie du geste pédagogique. Je suis obligé de vous témoigner que je n'ai jamais eu autant de plaisir à faire une improvisation générative. Je ne suis pas expert en la

matière, mais vraiment, j'ai pu me saisir dans l'instant d'objets sonores qui m'ont vraiment énormément inspirés et avec lesquels je me suis senti très en phase ou en décalage, peu importe. Mais en tous les cas je trouve qu'il y a avait une matière que je n'ai pas rencontrée dans les autres expériences d'improvisation générative que j'ai eues.

Christophe d'Alessandro : on pourrait le faire avec ta voix, d'ailleurs.

Robert Expert : oui ! En tous les cas, que ce soit dans l'improvisation générative qu'on vient de faire ou dans ta pièce sur Kafka « Devant la Loi », l'influence de vous, en tant qu'instruments, sur ma propre émission est considérable ! Au même titre qu'on ne chante pas pareil accompagné par un clavecin, un piano ou des cordes. Mais là évidemment, ça va quand même plus loin parce que c'est ma propre voix traitée différemment. Je dois témoigner que ça m'a fait chercher des choses, et peut-être trouver des choses, que je n'avais jamais osé faire ou que je n'avais jamais rencontrées dans ma pratique de musicien.

xxx : est-ce que vous reconnaissez votre voix ? Les intonations, le timbre, est-ce que vous reconnaissez quelque chose de l'ordre de votre voix ? Puisque vous évoquez le fait que vous avez eu beaucoup de plaisir à entendre votre voix ?

Robert Expert : je la reconnais tout le temps ! Mais ça ne me surprend pas tout le temps. Il y a des moments où le traitement est très neutre par rapport à l'original. Mais à partir du moment où ils traitent et la hauteur et le rythme et les transitions, les consonnes, tous les bruits, tous les transitoires d'attaque, de fin, etc. ça crée... un autre moi, mais c'est moi. Je peux m'identifier sans aucun problème, à tous les instants. Avec une certaine limite quand même, peut-être que dans les moments très aigus, ou très dans les graves, que je ne me reconnaitrais pas moi même. Mais sinon, aujourd'hui sur ce que vous [les chanteurs synthétiques] avez fait, je me reconnaissais tout le temps. Bien sûr il y a un élément psychologique quand même fort, je sais que c'est ma voix. Il faudrait faire des tests en mélangeant ma voix et une autre voix pour savoir à quel niveau je peux me reconnaitre, et je me tromperais bien entendu.

xxx : je suis très sensible à la nouveauté de ce que vous nous présentez, c'est très intéressant. Cela me fait penser à deux choses, d'une part au Sprechgesang, et d'autre part à Debussy, Pélléas et Mélisande, où il y a une espèce de voix chantée... Je me demande de quelle façon chacun des musiciens fait intervenir sa personnalité, il semble que chacun a des intonations particulières, mais il y a des choses qui semblent identiques'

Robert Expert : je vais répondre à la question : « est-ce que leur personnalité intervient dans ce qui est restitué ? ». Sans regarder, je savais très bien qui faisait quoi, sans aucun problème. D'autant que je n'ai jamais répété avec Victor, et par rapport au répétitions, j'ai tout à fait identifié une forme de personnalité par rapport à ce que faisait Samuel. Quand à Christophe et Boris, même avec une nouvelle partition, sans parler du style d'improvisation, je crois que je pourrais les reconnaître, sur leur mode de jeu, sur leur façon de faire. Oui, je les identifie très bien.

xxx : c'est très rassurant tout ça [rires]

Robert Expert : c'est pour ça que je dis que ce sont de vrais instruments. Je crois que ça m'a frappé dès le départ du travail, ce sont de vrais instruments, la personnalité du musicien s'exprime complètement, même si c'est ma voix... c'est une impression bizarre.

## 6.2 Au delà du chant

Dans les sections précédentes, nous avons démontré les capacités de Vokinesis à être utilisé comme instrument chanteur. Dans cette section, nous verrons comment utiliser ce logiciel pour des créations sonores allant au delà de la voix, grâce à des transformations extrêmes de signaux originaux commandées par un séquenceur externe.

La modularité de Vokinesis permet de lui assigner une multitude de contrôleurs. Il est donc possible de lui assigner des données émises par un DAW (*Digital Audio Workstation*), tel que Ableton Live. Ainsi, un clip MIDI peut contrôler la hauteur et le cadre rythmique en mode *Tap*, et des automations (cc MIDI virtuels préprogrammés) peuvent contrôler des paramètres vocaux ou temporels. Ces méthodes de contrôle automatisées sont tout à fait adaptées à la création de sons vocaux destinées à la musique électronique. Dans cette section, nous allons expliquer comment l'exemple sonore `Ex09 electroContext` a été créé. Nous verrons d'abord comment Vokinesis a été paramétré, et nous verrons ensuite comment le logiciel Ableton Live a été utilisé comme clavier et interface de contrôle MIDI automatiques.

### 6.2.1 Configuration de Vokinesis

L'exemple `Ex09 electroContext` est une mise en contexte des modifications des signaux originaux contenus dans l'exemple sonore `Ex12bis electroOrig`. Nous avons créé les exemples sonores `Ex10 electroPorta200` et `Ex11 electroPorta0` en modifiant le premier élément contenu dans `electroOrig`, et `Ex12 electroVibra` en modifiant le second élément de `electroOrig`. La façon dont Vokinesis a été configuré est représentée Figures 6.3. Tout d'abord, le contrôle de la hauteur a été assigné à un clavier MIDI (Cadre *Pitch Control*, paramètre *MIDI Keyboard*), ainsi que celui du cadre rythmique en mode *Tap* (Cadre *Tap controller*, paramètre *MIDI Keyboard*). Les vitesses de lecture des consonnes et des voyelles ont été réglées à 2 et 3 fois la vitesse originale. En effet, le but recherché en musique électronique n'est pas forcément d'avoir un signal vocal intelligible et naturel, mais plutôt un signal qui contienne une couleur vocale synthétique / robotique. Utiliser de telles vitesses de lecture permet alors de produire des rythmes rapides. Le mode *Loop* a été activé (cadre *Duration / Rhythm Control*), et les éléments contenus dans `electroOrig` représentent les parties bouclées du signal original : les FCP *start* et *end* ont été réglés à 5 et 9 pour les exemples `electroPorta[...]`, et à 4 et 8 pour l'exemple `electroVibra`. La plage de contrôle de la taille du conduit vocal a été réglée à 98%


FIGURE 6.3 – *Configuration de Vokinesis pour les exemples sonores* `electro[...]`. *Pour l'exemple* `electroVibra`, *les FCP* start *et* end *du cadre* Duration / Rhythm Control *ont été réglés à 4 et 8.*

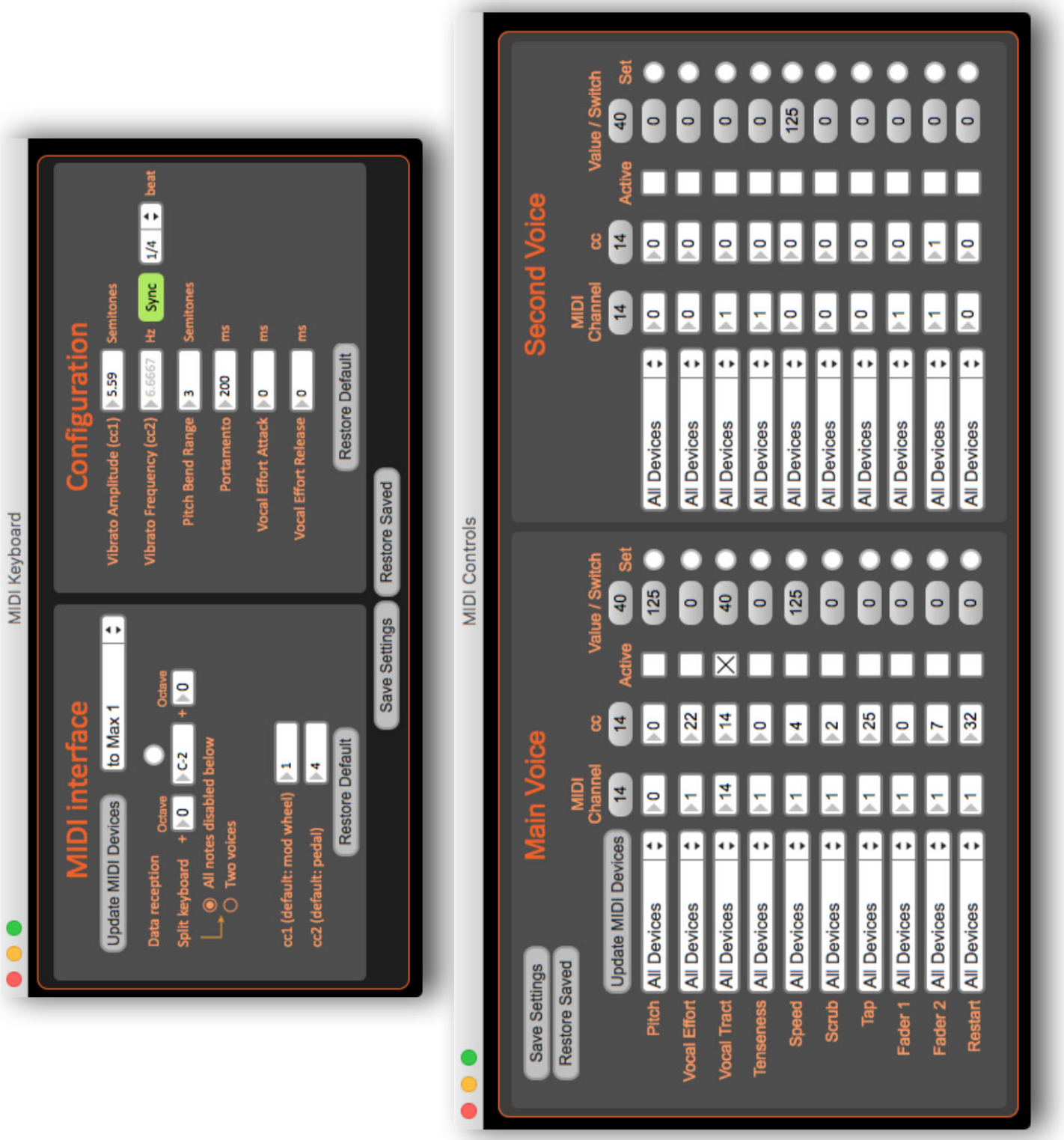


FIGURE 6.4 – *Configuration du clavier et des interfaces de contrôle MIDI pour les exemples sonores* `electro[...]`. *Pour l'exemple* `electroPorta0`, *le portamento de la fenêtre MIDI Keyboard a été réglé à 0 ms.*

(cadre *Control Range*, paramètre *Vocal Tract*). Le contrôleur assigné pourra donc agrandir ou rétrécir le conduit vocal au maximum de 98%. L'effet de réverbération (*Reverb*) a été activé, et une égalisation (*EQ*) permettant d'amplifier les basses fréquences et d'atténuer les hautes a été mise en place. Enfin, la tension vocale a été atténuée (paramètre *Tenseness* de gauche), et aucune modification préalable de la taille du conduit vocal n'a été effectuée.

La façon dont les interfaces de contrôle ont été paramétrées est présentée FIGURE 6.4. La durée du portamento a été réglé à 200 ms pour les exemples `electroPorta200` et `electroVibra`, mais à 0 ms pour l'exemple `electroPorta0` (paramètre *Portamento* dans le fenêtre *MIDI Keyboard*). Pour l'exemple `electroVibra`, nous avons réglé la fréquence maximale du vibrato à 1/4 de battement, et son amplitude maximale à 5.6 demi-tons. Le contrôle de l'amplitude et de la fréquence du vibrato ont été assignées aux cc 1 et 4 du clavier MIDI. Enfin, le contrôle de la taille du conduit vocal (paramètre *Vocal Tract* dans la fenêtre *MIDI Controls*) a été assigné au cc 14 du canal MIDI 14.

### 6.2.2 Modification du signal original avec Ableton Live

Ableton Live a été utilisé comme séquenceur MIDI pour produire les exemples `electro[...]`. Chacun d'entre eux a une durée de 4 mesures de 4 temps, et a donc été obtenu grâce à deux répétitions du clip MIDI représenté FIGURE 6.5, qui a une durée de 2 mesures. Le contrôle de la hauteur et du cadre rythmique est effectué par les notes MIDI préprogrammées présentées en haut de la figure. Cela fonctionne exactement de la même manière qu'avec un clavier MIDI ordinaire (voir la section 4.2.3). Si deux notes se chevauchent ou si elles ne sont pas espacées d'au moins 5 ms, alors le portamento est activé, et la touche n'est pas considérée comme relâchée pour le contrôle rythmique : l'instant cible n'évoluera pas du FCP actuel au suivant. L'effort vocal est directement lié à la vélocité de chaque note. Ainsi, pour couper le son de la synthèse, il faut appliquer une vélocité minimale, ce qui est le cas de la $5^e$ et de la $13^e$ note. Notez que ces notes ne sont pas espacées de plus de 5 ms de celles qui les précèdent, et l'instant cible n'évoluera donc que lorsqu'elles seront relâchées. Les exemples `electroPorta200` et `electroPorta0` ont été créés pour illustrer les différents rendus obtenus lorsque la durée du portamento a été réglée à 200 ms puis à 0 ms. Dans ce dernier cas, on entend des variations de hauteur très brusques au moment où les notes se chevauchant, alors qu'elles évoluent bien plus lentement dans le cas 200 ms.

Pour tous les exemples `electro[...]`, la taille du conduit vocal a été modifiée par l'automation correspondante (enveloppe temporelle rose *Taille du conduit vocal* sur la figure). Cette automation a été assignée au cc 14, comme l'indique la FIGURE 6.5 (paramètre à l'origine de la flèche qui pointe vers *Taille du conduit vocal*), et la piste MIDI qui contient ce clip a été assignée au canal MIDI numéro 14 (ceci n'est pas représenté sur la figure). Cela correspond bien à la configuration effectuée pour le paramètre *Vocal Tract* de la fenêtre *MIDI Controls* présentée FIGURE 6.4. Le conduit vocal est allongé lorsque cette automation est dans sa moitié inférieure,

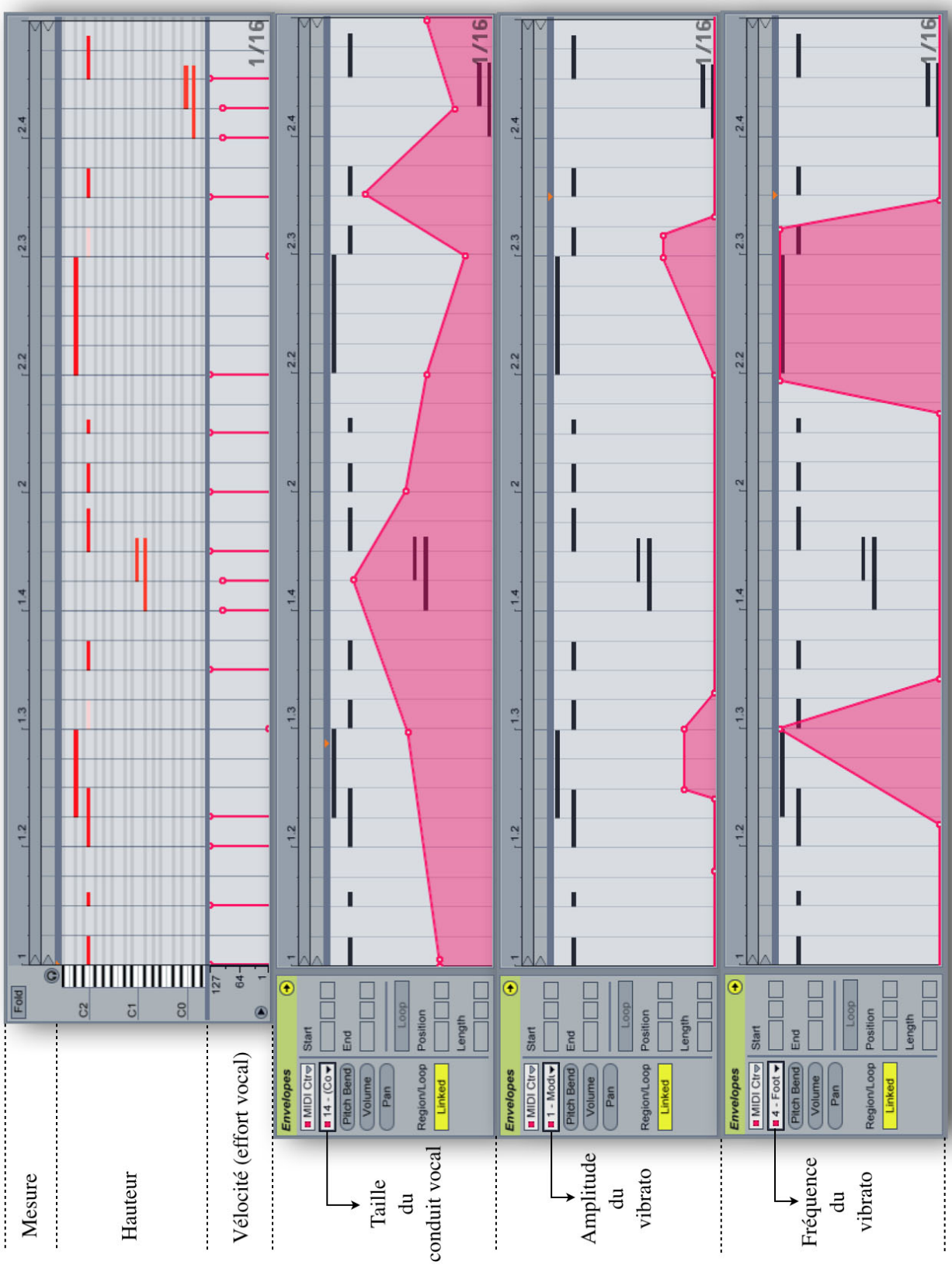


FIGURE 6.5 – *Motif mélodique et rythmique (cadre supérieur) et automations (3 cadres inférieurs) utilisés pour produire les exemples sonores* `electro[...]`. *Les automations liées au vibrato (2 cadres inférieurs) n'ont été utilisées que pour l'exemple* `electroVibra`.

raccourci lorsqu'elle est dans sa moitié supérieure. Les deux automations inférieures n'ont été utilisées que pour l'exemple `electroVibra`. Elles contrôlent l'amplitude et la fréquence du vibrato (cc 1 et 4 du clavier MIDI, comme configuré dans la fenêtre *MIDI Keyboard* de la FIGURE 6.4). L'amplitude maximale du vibrato ayant été réglée à une valeur assez élevée (5.6 demi-tons), les valeurs de son automation restent assez basses et cette amplitude maximale ne sera jamais atteinte. Le fréquence du vibrato, quant à elle, atteint sa valeur maximale à deux reprises. Lors de la première longue note (battement 1.2), la fréquence augmente petit à petit. Lors de la seconde longue note (battement 2.2), la fréquence du vibrato est constamment réglée à sa valeur maximale.

### 6.2.3 Mise en contexte des signaux modifiés

Nous avons souhaité placer les signaux modifiés dans un contexte de musique électronique. Nous avons donc créé un accompagnement rythmique grâce à un assemblage de signaux audio de grosse caisse, caisse claire, charleston, et crash (pistes *kick*, *snare*, *hh*, *hh open* et *crash* sur la FIGURE 6.6) ainsi qu'une ligne de basse électronique simple réalisée avec le synthétiseur Sylenth1 (piste *8 Sylenth1* sur la figure). Dans l'exemple sonore `electroContext`, les quatre premières mesures ne contiennent que les accompagnements. Les mesures 5 à 16 comportent l'enchainement des exemples `electroPorta200`, `electroPorta0` et `electroVibra` filtrés passe-haut (effet *EQ Eight* sur la figure). Chaque exemple dure 4 mesures, le démarrage d'un nouvel exemple étant indiqué par le retentissement d'une symbale crash. Les mesures 17 à 32 contiennent le même enchainement d'exemples (le dernier durant 8 mesures), avec cette fois-ci un effet d'écho (*Filter Delay*) activé par l'automation de la piste *9 Audio* sur la figure. Enfin, les mesures 27 à 32 contiennent une accélération de 100 bpm à 150 bpm (automation du tempo sur la figure). Les deux effets que nous venons d'évoquer sont fréquemment utilisés en musique électronique. Ils nous permettent d'illustrer la façon dont de tels signaux pourraient être utilisés dans ce contexte.

### 6.2.4 Limitations actuelles de Vokinesis pour ce type d'applications

La création de ces exemples sonores nous a permis de mettre en évidence les quelques lacunes encore présentes dans Vokinesis pour une utilisation automatisée. Tout d'abord, Vokinesis ne possède qu'un seul oscillateur à basse fréquence (ou *LFO* pour *Low Frequency Oscillator*), assigné à la fréquence fondamentale pour le vibrato. Or, il est très probable qu'un musicien souhaite assigner un LFO à un autre paramètre vocal ou temporel. Il serait donc utile d'ajouter à Vokinesis un jeu de plusieurs LFO (quatre, par exemple) qui puissent être assignés à n'importe quel paramètre vocal. Il serait également intéressant de permettre le contrôle des paramètres des LFO : chaque LFO pourrait se voir assigner un cc MIDI à son amplitude et à sa fréquence, permettant ainsi de les automatiser, comme nous l'avons fait pour les paramètres

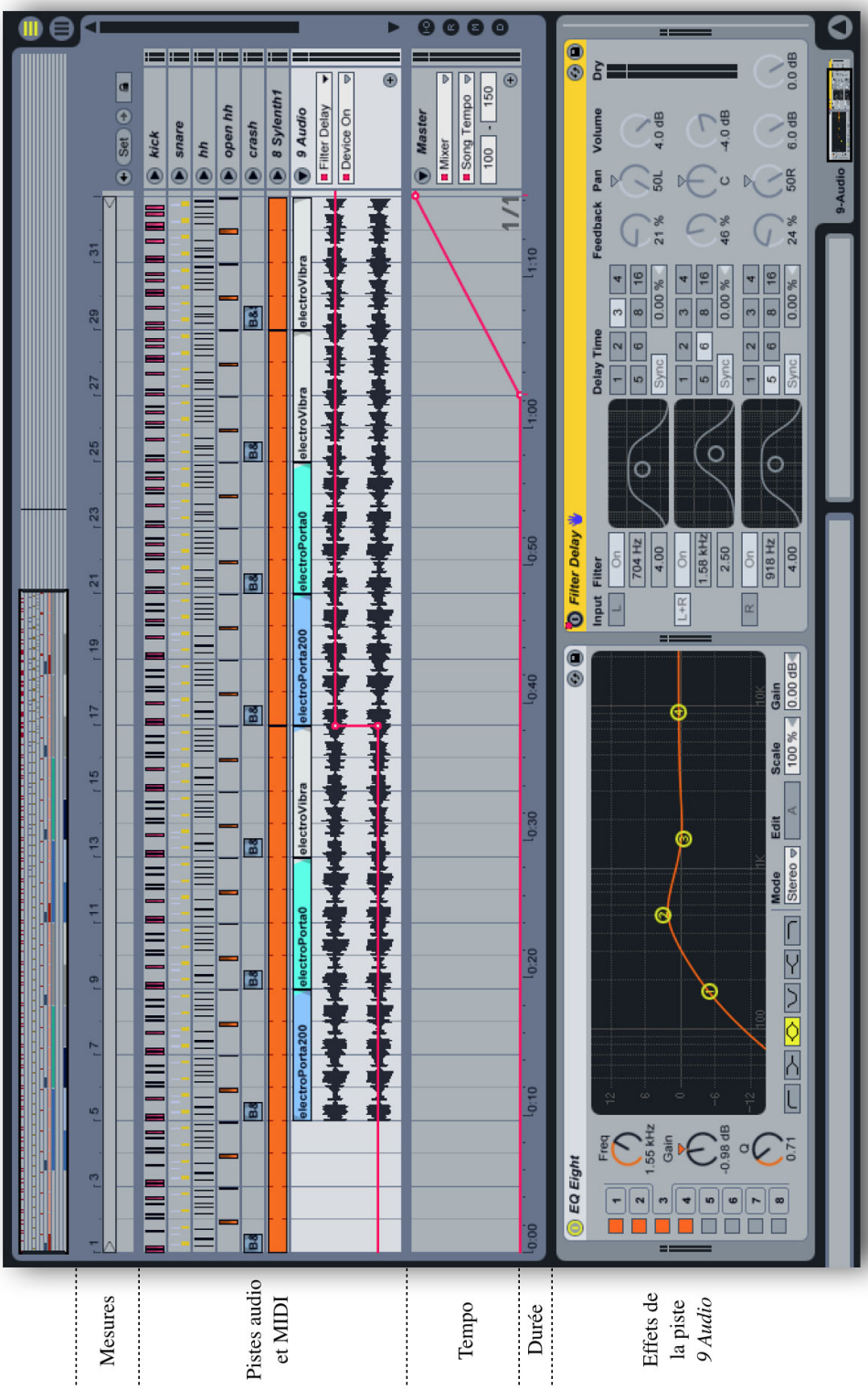


FIGURE 6.6 – *Session Ableton Live mise en place pour produire l'exemple* `electroContext`. *Les pistes* kick *à* crash *sont des pistes audio contenant les éléments percussifs. La piste* 8 Sylenth1 *contient le synthétiseur* Sylenth1 *qui produit la ligne de basse. La piste* 9 Audio *contient les enregistrements audio issus de Vokinesis (le nom de l'exemple sonore correspondant est indiqué sur chaque clip audio). L'automation sur la piste* 9 Audio *contrôle l'activation de l'effet* Filter Delay. *L'automation sur la piste* Master *contrôle le tempo, qui évolue de 100 bpm à 150 bpm sur les 6 dernières mesures.*

du vibrato. De même, la vélocité et l'enveloppe (durées d'attaque et de relâchement) ne permettent pour l'instant que de contrôler les variations d'effort vocal. Il serait utile de permettre l'assignation de plusieurs paramètres vocaux ou temporels à la vélocité. Il faudrait également fournir plusieurs enveloppes (quatre par exemple) qui puissent être assignées à n'importe quel paramètre vocal ou temporel. Enfin, la façon dont le portamento est programmé est pour l'instant quelque peu limitée. En effet, les synthétiseurs possèdent généralement 2 modes de portamento. L'un fonctionne de la même manière que le notre : si les notes se chevauchent, le portamento est actif ; sinon, la fréquence fondamentale d'une nouvelle note est directement émise au synthétiseur. Dans le mode manquant à Vokinesis, le portamento est tout le temps actif : lorsqu'une nouvelle note est jouée, la fréquence fondamentale évoluera de sa dernière valeur jouée à la nouvelle valeur ciblée, avec une durée définie par celle du portamento. Par ailleurs, il faudrait également offrir la possibilité d'assigner un contrôleur MIDI à la durée du portamento, ce qui en permettrait l'automatisation.

## 6.3 Conclusion

Dans ce chapitre, nous avons présenté les différentes représentations du Chorus Digitalis qui ont fait l'usage de Vokinesis. Nos méthodes de contrôle rythmique ont pu y être testées. Elles permettent aux musiciens de jouer en synchronie, aussi bien entre voix de synthèse qu'avec d'autres instruments. Par ailleurs, les retours de Robert Expert offrent une bonne évaluation de notre instrument de la part d'un professionnel de la voix chantée. Nous avons également démontré la puissance de la modularité du logiciel par son interfaçage avec le logiciel de création musicale Ableton Live.

Chapitre 7

# Conclusions et perspectives

**Sommaire**



## 7.1 Bilan

Vokinesis s'inscrit dans la lignée des systèmes d'externalisation vocale. En permettant le contrôle du rythme vocal, plus simple que le contrôle de tous les phonèmes, mais offrant plus de précision, de liberté et d'expressivité que le contrôle de la vitesse de lecture, nous avons franchi une nouvelle étape dans le domaine du contrôle performatif de la synthèse vocale. Les méthodes de contrôle rythmique que nous avons mises en place ont résulté d'une réflexion sur la nature du rythme vocal. Les règles de syllabification étant trop variables d'une langue à l'autre, l'unité rythmique ICPG (Inter-P-Center Group) proposée par [Barbosa & Bailly 1994] nous a paru mieux adaptée pour une définition d'un motif rythmique inter-linguistique invariable. Ces groupes rythmiques sont tous composés du motif *Noyau rythmique / Liaison rythmique*. Cette unité rythmique est par ailleurs mieux adaptée à l'écriture musicale : un événement rythmique d'une partition définit la durée qui sépare deux noyaux.

Deux méthodes de contrôle rythmique ont vu le jour. Le contrôle binaire du cadre rythmique (mode *Tap*) est une bonne analogie avec l'idée de la théorie Frame/Content de l'évolution de la parole [MacNeilage 1998] qui stipule que le séquencement rythmique des syllabes est effectué par des cycles d'ouverture/fermeture du conduit vocal : la durée d'une phase ouverte est commandée par le maintien enfoncé d'un touche de contrôle, la durée de la phase fermée par son maintien relâché. Pour le contrôle continu des liaisons rythmiques (mode *Fader*), les durées des phases ouvertes sont contrôlées par le maintien d'un potentiomètre dans une position extrême,

et les durées des phases de fermeture/ouverture sont gérées par les durées de transition d'une position extrême à l'autre. Le mode *Fader* constitue également une bonne analogie avec le modèle de gestes articulatoires de [Browman & Goldstein 1990a, Browman & Goldstein 1990b, Browman & Goldstein 1992]. Dans leur modèle, le geste vocalique, qui permet de passer d'une voyelle à la suivante, démarre au même instant que le premier geste consonantique d'un groupe de consonnes qui sépare les deux voyelles. Dans notre système, le déplacement d'un potentiomètre d'une position extrême à la suivante décrit une intention de passer d'un noyau rythmique au suivant (équivalent au geste vocalique), mais également de commencer la prononciation des consonnes qui séparent les deux noyaux (gestes consonantiques).

Les analyses des performances d'un groupe de sujets à effectuer une tâche d'imitation prosodique nous ont permis de montrer que le mode *Tap* permet une précision du contrôle du rythme de la parole remarquable. En effet, les durées des groupes rythmiques des phrases de synthèse différaient en moyenne de seulement 20 ms par rapport aux durées originales, alors que la JND pour une durée syllabique est d'environ 25ms [Wagner 2008]. Cependant, de nombreux sujets ont jugé la tâche fatigante, en raison de la rapidité des mouvements à effectuer. Nous pensons donc que cette méthode de contrôle serait mieux adaptée à des langues aux tempi plus lents telles que l'anglais, l'allemand ou le polonais [Wagner 2008]. Pour le français, l'utilisation d'un contrôle de type *Fader NLN* (voir section 3.3) avec des mouvements d'ouverture / fermeture de la main serait sans doute mieux adaptée : ce sont des mouvements très rapides et peu fatigants, qui ne semblent induire que peu d'effets de synergie avec la main qui contrôle la hauteur. Cependant, nous n'avons pas eu l'occasion de tester cette méthode avec une interface de contrôle qui soit assez réactive pour permettre une précision rythmique suffisante. L'utilisation d'un gant de contrôle tel que celui de [Fels & Hinton 1993, Fels & Hinton 1998] serait sans doute satisfaisante.

Par ailleurs, les deux modes de contrôle rythmique (*Tap* et *Fader*) ont été pratiqués à de nombreuses reprises dans le cadre des répétitions et des représentations du Chorus Digitalis. Les musiciens étaient en mesure de jouer en synchronie avec très peu d'entrainement, aussi bien entre voix de synthèse qu'avec d'autres instruments de musique.

Pour le contrôle des paramètres de hauteur et de qualité vocal, la puissance expressive qu'offre la tablette graphique semble avoir fait consensus [Wanderley *et al.* 2000, Kessous 2004b, D'Alessandro & Dutoit 2007, Astrinaki *et al.* 2012, Le Beux *et al.* 2007, d'Alessandro *et al.* 2011, d'Alessandro *et al.* 2014, Perrotin 2015, Feugère *et al.* 2017]. Cependant, les PMC semblent mieux adaptées au contrôle de variations mélodiques rapides de type yodel. Notre pratique nous laisse tout de même penser que les mouvements d'écriture de la tablette graphique ont un pouvoir expressif plus important que les mouvements tactiles des PMC. De plus, nous pensons que la maîtrise experte quasi-universelle du maniement du stylo rend les tablettes graphiques plus rapides à prendre en main pour des débutants. Quoi qu'il en soit, chacune de ces d'interfaces offre différentes possibilités de mouvements, et le choix d'utilisation devrait dépendre des intentions musicales des interprètes. Pour ce qui est du contrôle précis et

simultané de deux voix de synthèse, les retours tactiles et kinesthésiques qu'offrent certains PMC constituent un avantage par rapport aux tablettes graphiques, pour lesquelles la vision joue un rôle très important [Perrotin & D'alessandro 2016a].

Cependant, rien ne dit pour l'instant que les PMC soient mieux adaptées au contrôle expressif de la parole : la polyphonie est une pratique exclusivement réservée au chant, la rapidité et la précision de contrôle de la hauteur jouent une importance moindre pour la parole dans le cas des langues non-tonales. De plus, la tablette graphique semble mieux adaptée à une utilisation qui ne soit pas réservée à des musiciens, en raison de l'expertise très répandue des gestes d'écriture. L'utilisation de gestes d'écriture pour modifier les courbes de hauteur de signaux obtenus par synthèse HMM-TTS expressive afin d'en améliorer l'expressivité a été évaluée de façon tout à fait positive par des études perceptives. La comparaison de productions expressives naturelles, HMM-TTS et chironomiques a permis de montrer que le contrôle chironomique de la hauteur permet d'augmenter le taux de reconnaissance des types d'expressivité pour lesquels la hauteur joue un rôle important. Ces modifications permettent en outre d'améliorer la qualité globale perçue de la synthèse.

Le développement de Vokinesis a nécessité l'amélioration de l'algorithme RT-PSOLA [Le Beux *et al.* 2010]. Le problème des bruits tonaux indésirables lors de l'allongement de signaux non-voisés a été résolu par l'utilisation de fenêtre d'analyse de tailles consécutives aléatoires. Nous y avons également ajouté une méthode de ré-échantillonnage temps-réel permettant de créer des effets de changement de taille du conduit vocal. Nous avons baptisé l'adaptation de cet algorithme VRT-PSOLA (Vokinesis RT-PSOLA). Par ailleurs, nous avons mis en place une méthode permettant de simuler les variations de fréquence centrale du formant glottique lors de variations de tension vocale, en agissant sur l'amplitude du premier harmonique de synthèse. Enfin, nous avons adapté le filtre de pente spectrale présenté dans [Doval *et al.* 2006] pour permettre l'augmentation de l'effort vocal, en plus de sa diminution. Ces techniques de traitement de signal assemblées forment la méthode VoPTiQ (Voice Pitch, Time and Quality modification). Cette méthode pourrait être améliorée en remplaçant VRT-PSOLA par le vocoder WORLD [Morise *et al.* 2016]. En effet, ce vocoder permettrait de conserver les effets de modification de hauteur, de durée et de taille du conduit vocal, mais permettrait en plus d'avoir un contrôle sur la quantité de souffle dans le signal de synthèse, et offrirait donc une modalité de contrôle expressif supplémentaire.

Vokinesis est le fruit de toutes ces réflexions. C'est un logiciel robuste qui a été développé pour être utilisé dans des situations de concert, mais également pour des applications de recherche. Sa modularité ouvre de nombreuses perspectives d'applications musicales, scientifiques, thérapeutiques ou encore pédagogiques. Dans les sections suivantes, nous présentons quelques pistes de réflexion concernant nos perspectives d'applications.

## 7.2 Perspectives d'applications

Nous avons vu dans le Chapitre 4 que le geste manuel imite très bien le geste vocal. Les questions suivantes se posent alors : à quel point le geste vocal est-il en mesure d'imiter le geste manuel ? L'externalisation vocale en facilite-elle l'internalisation ? Si tel était le cas, Vokinesis pourrait être utilisé à des fins pédagogiques ou thérapeutiques.

### 7.2.1 Apprentissage des langues tonales

Cette section présente les prémices d'un projet de collaboration avec Xiao Xiao [1]. Tout comme 60% à 70% des langues mondiales, le chinois mandarin est de nature tonale : les courbes de variation de hauteur ont un rôle sémantique [Yip 2002]. Cette langue contient quatre tons principaux, présentés Figure 7.1, et un ton neutre. Chaque ton est associé à une syllabe. La signification d'une syllabe est donc définie par ses phonèmes d'un part, et par son ton d'autre part. Par exemple, la syllabe « *ma* » peut prendre 5 significations différentes selon le ton qui lui est associé : mère (1), engourdissement (2), cheval (3), réprimander (4), indique une question (neutre).

Les tons du chinois mandarin sont souvent reconnus comme l'aspect le plus difficile à apprendre pour des locuteurs non-natifs [Kiriloff 1969, Shen 1989]. Des études ont montré que des locuteurs dont l'anglais est la langue maternelle sont moins sensibles que les chinois aux différences tonales [White 1981]. Pour la production, les courbes d'intonation de l'anglais semblent interférer avec la prononciation des tons chinois [Chiang 1979]. Pour surmonter ces difficultés, les enseignants recommandent une attention particulière sur les tons dès le début de l'apprentissage [Chiang 1979, Kiriloff 1969]. Les méthodes habituelles présentent d'abord l'allure sonore de chaque ton, et préconisent des exercices d'identification et de prononciation réguliers [Wang *et al.* 1999], ainsi qu'une écoute fréquente de locuteurs natifs [Orton 2011]. Cependant, l'acquisition d'une maitrise basique prend généralement plusieurs mois, et une prononciation parfaite reste souvent hors de portée, même après plusieurs années de pratique [Chen 1993]. Des recherches plus récentes se sont tournées vers l'utilisation de systèmes d'analyse de la voix comme outil d'évaluation pour la prononciation des tons [Chan 2003]. Une étude avec des étudiants débutant l'apprentissage du chinois a mis en évidence une amélioration significative de la prononciation des tons après l'utilisation d'un retour visuel de leurs propres courbes de hauteur comparées avec celles de locuteurs natifs [Chun *et al.* 2012]. D'autres ont cherché comment des gestes physiques pourraient faciliter l'apprentissage des langues en générale et plus particulièrement des tons chinois [Roth 2001, Orton 2011].

S'il s'avère que l'externalisation des gestes vocaux en facilite l'internalisation, Vokinesis pourrait être un outil puissant pour l'apprentissage des langues tonales. Cette idée a vu le jour à la suite d'une expérimentation informelle que nous avons menée à la conférence NIME (New Interfaces for Musical Expression) en mai 2017, lors d'une démonstration de Vokinesis. Xiao, qui travaille principalement sur l'ap-

1. https://tangible.media.mit.edu/person/xiao-xiao/

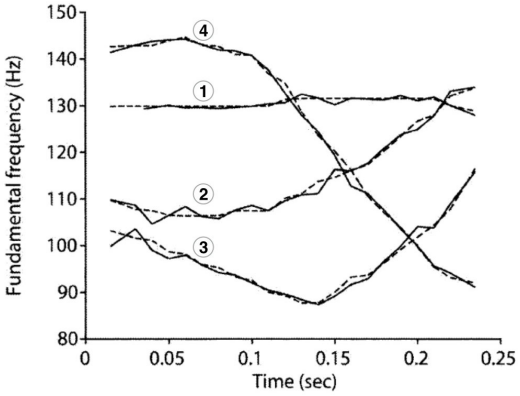


FIGURE 7.1 – *Variations de hauteur pour les 4 tons (non neutres) du chinois mandarin. Image issue* [Krishnan *et al.* 2004].

prentissage de la musique par imitation et par imprégnation corporelle du son [Xiao *et al.* 2013, Xiao & Ishii 2016, Xiao *et al.* 2016], a tout de suite été inspirée par notre système. La FIGURE 7.2 la montre en train de rechercher comment dessiner les tons du mandarin sur la tablette graphique en utilisant Vokinesis. Pour tester cette idée, Xiao et Beici Liang (une autre locutrice native du chinois, du *Centre for Digital Music, Queen Mary University of London*) ont effectué cette tâche de recherche préliminaire des formes gestuelles des différents tons du chinois mandarin, avant de les retranscrire sur une feuille de papier. Leurs tracés sont représentés dans la FIGURE 7.3. Il est intéressant de voir que leurs tracés tonaux ne sont pas identiques, les plus grandes différences étant observables entre les tons 1 et 3.

Nous avons ensuite utilisé ces tracés tonaux pour prononcer une série de phrases chinoises simples avec Vokinesis, en apposant les feuilles de papier sur la tablette graphique. Pour ce faire, nos deux locutrices nous ont alors enregistré neuf phrases chinoises courtes (de deux à trois syllabes), présentées dans le Tableau 7.1. À chaque syllabe est associé un ton, indiqué par son numéro. Nous avons assigné le contrôle du rythme en mode *Tap* à l'état de toucher du stylet sur le tablette : les durées des noyaux vocaliques sont contrôlées par les durées de contact du stylet avec la tablette, et les durées des liaisons par les durées de non-contact. L'effort vocal a été assigné à la position du stylet sur l'axe $y$ de la tablette, nous évitant ainsi de couper le son lorsqu'il n'est plus en contact. La prononciation d'une syllabe (en durée) et d'un ton (en hauteur) s'effectuait donc avec des mouvements d'un seul et même membre. Prenons l'exemple de la phrase « *wo ai ni* », dont l'étiquetage phonétique et les FCP sont présentés FIGURE 7.4. Poser le stylet au début du tracé tonal 3 déclenche la première syllabe, suivre ce tracé en prononce le ton. Un fois la fin du ton 3 atteinte, il faut passer au ton de la seconde syllabe, qui correspond au ton 4 dans notre exemple. Pour ce faire, il faut d'abord relever le stylet, ce qui déclenche la liaison rythmique suivante (la plosive glottique entre le premier noyau et le second). Il faut ensuite reposer le stylet au début du tracé de ce ton, ce qui déclenche le noyau suivant. Le ton est ensuite prononcé en suivant son tracé, et lorsque sa fin est

Figure 7.2 – *Xiao Xiao recherchant les courbes de hauteur des différents tons du chinois mandarin* (Conférence NIME 2017 à Copenhague, Danemark).

atteinte, il faudra prononcer le groupe rythmique suivant en effectuant les mêmes opérations. Ainsi, chaque groupe rythmique est prononcé par un mouvement de *posé - tracé - relevé* du stylet.

Xiao et Liang ont testé leurs tracés tonaux respectifs avec le stylet et ont ainsi pu vérifier que les deux jeux de tracés permettaient bien de prononcer des tons chinois à l'allure authentique. Pour ma part, bien que je n'aie aucune expérience préalable en chinois, nos deux locutrices ont pu juger mes productions comme très convaincantes, que je suive les tracés tonaux de Xiao ou de Liang. Au bout d'une trentaine de minutes de traçage tonal, j'étais en mesure d'imiter les tons à la voix, et nos locutrices étaient très convaincues de mes productions. Cependant, cette rapidité d'apprentissage est sans doute biaisée par mes connaissances dans le domaine de la parole et par ma pratique musicale.

Afin d'améliorer le contrôle temporel, nous pourrions également imaginer un mode *Fader* dont la valeur du potentiomètre correspondrait à la position du stylet sur le tracé tonal prononcé. Cela permettrait de décomposer les phonèmes et les tons tout en gardant le contrôle instantané de la durée, fournissant ainsi une information précise sur le lien entre hauteur et articulation au sein d'une syllabe.

Notre expérimentation informelle a démontré l'efficacité du suivi de tracés tonaux pour la prononciation de tons chinois. Elle a alors soulevé plusieurs questions : Quelle est l'étendue de variation de hauteur pour la prononciation correcte de chaque ton chez des locuteurs natifs ? À quel point l'apprentissage de la prononciation des tons peut il être accéléré par des tâches de tracé tonal ? Quel niveau de précision la pratique de tracés tonaux permettrait-elle d'atteindre, et en combien de temps ? Quelles tâches parallèles pourraient faciliter cet apprentissage ?

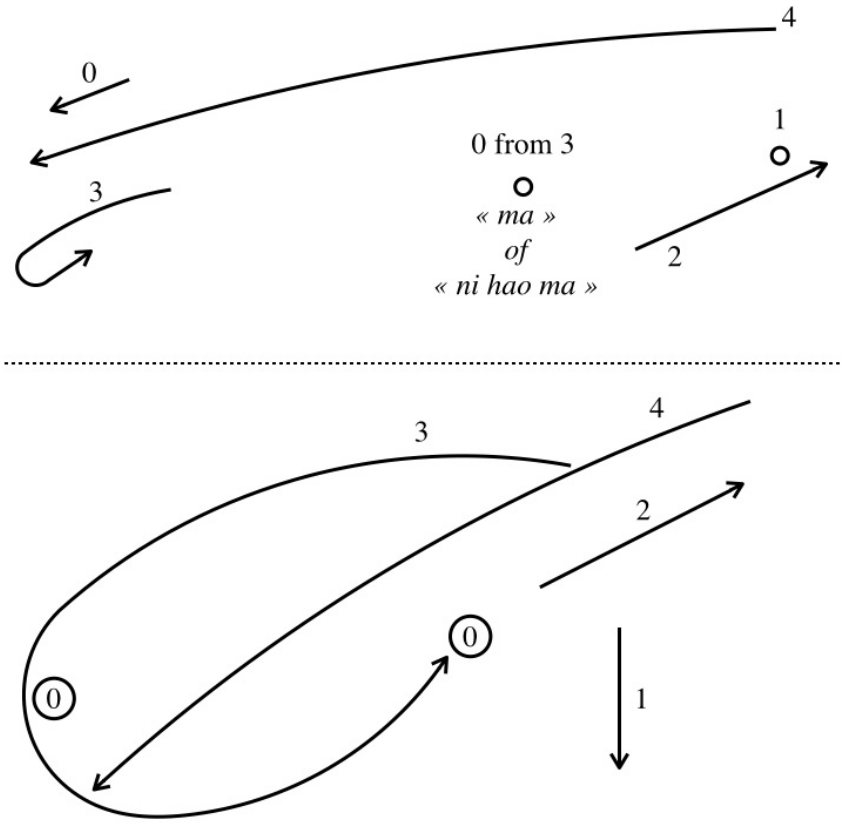


FIGURE 7.3 – *Tracés tonaux dessinés par Xiao (en haut) et Liang (en bas)* (Conférence NIME 2017 à Copenhague, Danemark).

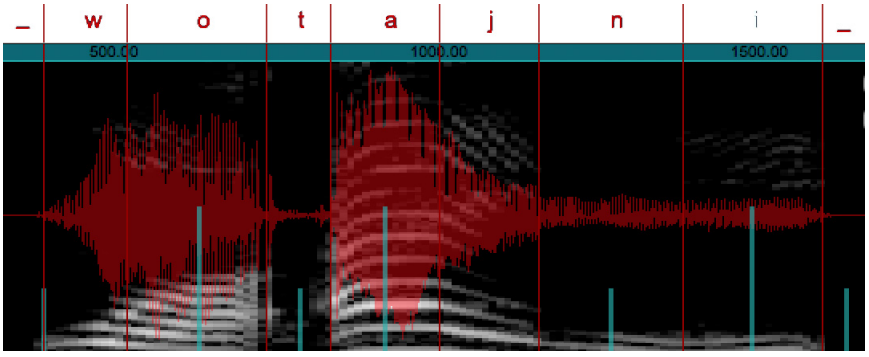


FIGURE 7.4 – *Phonèmes et FCP pour la phrase « wo ai ni ». Le système n'ayant pas été conçu pour le chinois, la plosive glottale entre le* [o] *et le* [a] *a été étiquetée comme une plosive dentale, ce qui ne change rien au contrôle.*

Tableau 7.1 – *Phrases chinoises enregistrées par nos deux locutrices. Chaque syllabe est numérotée par son tonème correspondant. Les phrases qui contiennent deux numérotations correspondent à la numérotation théorique suivie de la numérotation pratique. Les transcriptions phonétiques et les traductions françaises sont données à droite de chaque phrase.*

| **Phrase chinoise** | **Transcription phonétique** | **Traduction française** |
|---|---|---|
| **4** **1** **4**<br>Bu zhi dao | [budʒydaw] | *Je ne sais pas* |
| **1** **4**<br>chi fan | [ʃyfan] | *Manger* |
| **4** **3**<br>dian nao | [djɛnnaw] | *Ordinateur* |
| **1** **2** **3**<br>fei chang hao | [fejtʃãŋhaw] | *Très bien* |
| **3** **3** **1**<br>**2** **3** **0**<br>ni hao ma | [nihawma] | *Comment ça va ?* |
| **3** **3**<br>**2** **3**<br>ni hao | [nihaw] | *Bonjour* |
| **3** **4** **3**<br>wo ai ni | [waʔajni] | *Je t'aime* |
| **4** **0**<br>xie xie | [sjesje] | *Merci* |
| **4** **4**<br>zai jian | [tsajsjɛn] | *Au-revoir* |

Nos futures recherches tenteront de répondre à ces questions en rassemblant tout d'abord un plus grand nombre de tracés et de prononciations tonaux produits par des locuteurs natifs. Ces données devraient nous permettre de mettre en place des exercices d'apprentissage, qui pourront alors être évalués en comparant la rapidité d'apprentissage de prononciation et de reconnaissance des tons pour des groupes de sujets débutants utilisant Vokinesis d'une part, et des méthodes conventionnelles d'autre part.

### 7.2.2 Entraînement à la compréhension d'accents complexes, ou même de phonèmes d'autres langues

Les travaux de [Wanat *et al.* 2017] ont permis de montrer l'efficacité d'une tâche de battement du rythme accentuel pour l'apprentissage de la compréhension de l'accent de Glasgow. Une tâche de frappe du rythme d'un texte prononcé par un locuteur gaswégien proposée à une groupe de sujets chinois débutants en anglais leur aurait permis d'apprendre à détecter les fortes réductions des syllabes faibles et ainsi d'améliorer significativement leur compréhension des mots-outils.

Nous pensons que des expériences similaires pourraient être menées avec Vokinesis, en utilisant les différents modes de contrôle temporel pour modifier les durées de phrases originales, enregistrées dans un accent complexe, tout en lisant le texte correspondant. En mode *Fader*, notre système pourrait en plus permettre de décomposer chaque syllabe et d'apporter des informations supplémentaires sur la façon dont celles-ci sont prononcées / réduites... Nous pensons donc que cette tâche pourrait apporter une aide à l'apprentissage de la compréhension de tels accents.

### 7.2.3 Enseignement du Chant

Nous vous avons déjà présenté Robert Expert, contre-ténor et professeur de chant lyrique au conservatoire de Bobigny, qui a participé à deux de nos représentations du Chorus Digitalis (sections 6.1.2 et 6.1.3). Dans le cadre de nos répétitions, nous avons eu l'opportunité de lui faire utiliser Vokinesis, en contrôlant sa propre voix. Lors de nos discussions qui ont suivi ces essais, Robert Expert nous a confié que notre système serait sans doute très utile pour l'enseignement du glissando. En effet, il semblerait que cet exercice soit particulièrement difficile à enseigner à certains élèves. Expert pense que si le professeur pouvait illustrer ses attentes en utilisant la voix de son élève, cela permettrait à ce dernier de passer outre ses barrières psychologiques qui seraient souvent à l'origine de ses difficultés. Une utilisation de Vokinesis permettrait donc à un professeur d'enregistrer la voix de son élève (un [a] tenu monotone, par exemple), puis d'illustrer ses attentes en modifiant cet enregistrement. Ainsi, la tâche de l'élève consisterait à imiter sa propre voix plutôt que celle de son professeur.

### 7.2.4 Outil thérapeutique

Vokinesis pourrait offrir de nouvelles techniques d'orthophonie. Il pourrait par exemple aider des patients souffrant de l'aphasie de Broca [Broca 1861] à recouvrer la parole, ou du moins faciliter la communication : la thérapie MIT (Melodic Intonation Therapy), qui est l'une des plus efficaces, semble fonctionner en majeure partie grâce à des tâches de « mélodisation » de l'intonation associées à des frappes rythmique [Schlaug *et al.* 2008] : les patients apprennent alors à utiliser la partie musicale du cerveau (hémisphère droit) pour remplacer la partie de production de la parole lésée (hémisphère gauche) [Schlaug *et al.* 2009]. Peut-être que l'externalisation vocale pourrait produire des effets similaires ? De même, pour des patients souffrant de bégaiement, peut-être que des tâches de lecture simultané au contrôle externe du même texte les aiderait à apprendre à planifier leurs phrases ? Nous n'en sommes qu'au stade des suppositions, mais de nombreux orthophonistes semblaient très intéressés par l'étude des capacités thérapeutiques de notre système.

### 7.2.5 Outil de Recherche

Vokinesis pourrait être un outil puissant pour les recherches en parole expressive. En effet ses capacités de décomposition des différents éléments suprasegmentaux (hauteur, rythme, qualité vocale) peut permettre l'étude de l'effet du contrôle de chacun de ces paramètres sur le rendu expressif de façon indépendante. Nous en avons d'ailleurs fait une première expérience en ne laissant la possibilité de contrôler que la hauteur (section A.3), qui a effectivement permis de comprendre le rôle que joue (ou que ne joue pas) la hauteur pour certains types d'expressivité.

Le contrôle performatif du rythme offre aussi un nouveau paradigme d'étude pour les recherches en prosodie. Par exemple, bien que des taches de frappe rythmique aient été utilisées pour déterminer la position des p-centers [Repp 2005, Villing *et al.* 2011], le mouvement de relâchement n'a pas encore été étudié. Des mesures des variations de ce mouvement lors de tâches de contrôle rythmique pour déterminer si les p-centers sont représentées par un ou deux paramètres temporels serait un point intéressant à explorer.

Par ailleurs, l'externalisation de la production vocale serait sans doute très utile à des applications d'études du cerveau par imagerie cérébrale. En effet, de nombreux paradigmes d'études sur la production de la parole cherchent à annihiler un certain aspect (par exemple le contrôle des phonèmes) pour permettre une meilleure observation d'un autre aspect (par exemple le contrôle du rythme) [MacNeilage 1998]. Vokinesis pourrait permettre d'effectuer de telles séparations.

### 7.2.6 Aller plus loin : contrôle du texte

Vokinesis permet de contrôler et d'improviser le rythme d'un texte préparé. La prochaine étape consisterait à permettre l'improvisation d'un texte, sans avoir à contrôler tous les phonèmes de façon indépendante. Il s'agirait alors de réfléchir à des méthodes de sélection de syllabes. Pour permettre la production de phrases sensées,

nous pouvons imaginer des méthodes de sélection dynamique, où les propositions des syllabes suivantes dépendraient des syllabes précédemment prononcées. Pour un cadre exclusivement musical, nous pouvons imaginer un espace contenant un nombre limité de syllabes, qui permettrait d'effectuer des productions rythmiques improvisées dont le sens n'aurait pas d'importance.

Annexe A

# Amélioration de la synthèse TTS expressive par stylisation chironomique de l'intonation

**Sommaire**



Cette annexe présente les résultats des travaux que nous avons effectués avec [Evrard *et al.* 2015], lors desquels nous avons cherché à vérifier les capacités du contrôle chironomique de la hauteur à améliorer l'expressivité de signaux de parole issus d'un système de synthèse HMM expressive. Cela pourrait s'avérer particulièrement utile à certaines applications, telles que la production de livres audio ou de dialogues pour les jeux vidéo. Nous tenterons ici de répondre aux questions suivantes :

- La stylisation chironomique de l'intonation est-elle en mesure d'améliorer l'expressivité de signaux obtenus par modèle statistique ?
- Les modification chironomiques améliorent-elles (ou dégradent-elles) la qualité globale de la synthèse ?

Nous avons utilisé Vokinesis pour modifier l'intonation et améliorer ainsi l'expressivité de signaux de synthèse obtenus à partir d'un synthétiseur expressif *Text-To-Speech* (TTS) français basé sur un modèle HMM d'un corpus de parole expressive. La description du système TTS sera présentée dans la section suivante. Nous détaillerons ensuite la procédure de modification chironomique de la hauteur des signaux de synthèse TTS, puis nous fournirons une évaluation perceptive des effets de la modification chironomique sur l'expressivité et la qualité générale de la synthèse.

## A.1 LIPS$^3$ : système de synthèse TTS expressive

Le système LIPS$^3$ est un système de synthèse par modèle paramétrique statistique développé lors de la thèse de [Evrard 2015] et spécialement conçu pour la

synthèse TTS expressive du français : il a été entraîné sur un corpus de parole expressive enregistré par une actrice professionnelle parisienne. Le *Speech Signal Synthesis Toolkit* [Tokuda *et al.* 2012] a été utilisé pour l'extraction de paramètres acoustiques adaptés au français, selon [Le Maguer *et al.* 2013]. Pour la re-synthèse de ces paramètres, LIPS$^3$ fait usage du vocoder STRAIGHT [Kawahara *et al.* 1999]. Les modèles acoustiques de ces paramètres ont été entraînés avec la plateforme HTS [Tokuda *et al.* 2013], avec une adaptation particulière pour la synthèse expressive.

Le corpus était composé d'un groupe de phrases neutres et de six groupes de phrase expressives moins volumineux. Le groupe neutre comportait 1402 phrases et les groupes expressifs étaient composés des 160 premières phrases du groupe neutre. Les types d'expressivité ont été sélectionnés afin d'assurer une grande différence dans leurs production acoustiques (variations de hauteur, valeur moyenne de la fréquence fondamentale, phonation soufflée ou non, sourire...) Ainsi, 5 types d'expressivité vocale ont été sélectionnés à partir du corpus GEMEP (GEneva Multimodal Emotion Portrayals) [Bänziger *et al.* 2012] (colère, peur, joie, tristesse et surprise), et la sensualité y a été ajoutée pour avoir une modalité comportant du souffle [Léon 1993]. 5 phrases ont été retirées du corpus afin d'être utilisées pour les évaluations. Le système a donc été entrainé sur 1397 phrases neutres et 155 phrases pour chaque modalité expressive.

## A.2 Amélioration chironomique de la synthèse expressive

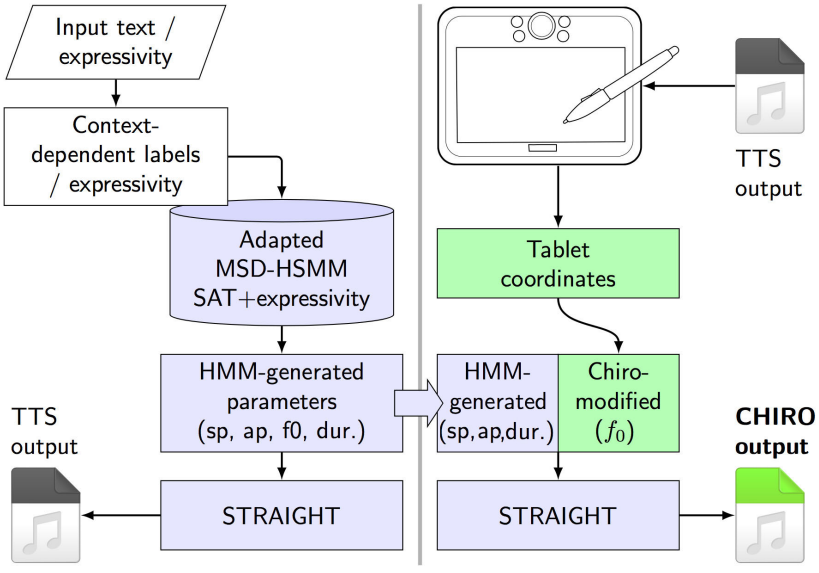


FIGURE A.1 – *Production chironomique de parole expressive : synthèse de parole expressive (à gauche), modification par Vokinesis (à droite).*

Les 5 phrases d'évaluation ont été synthétisées par le système LIPS$^3$ pour les six

types d'expressivité considérés. Dans la FIGURE A.1, à gauche, ces signaux correspondent à la sortie *TTS output*. Ils ont ensuite été utilisés comme signaux originaux dans Vokinesis (en haut à droite de la figure). Trois musiciens experts du contrôle chironomique de la voix ont été sélectionnés pour modifier ces phrases de synthèse afin d'en améliorer l'expressivité. Par exemple, lorsqu'il leur était demandé d'améliorer l'expressivité de la peur ou de la surprise, les signaux originaux qu'ils devaient modifier possédaient déjà les expressions respectives de peur et de joie, selon les paramètres générés par le système de synthèse. Ils pouvaient uniquement modifier la hauteur sur l'axe horizontal de la tablette. Pour chacune des phrases synthétisées par LIPS$^3$ (les 5 phrases d'évaluation pour chacun des 6 types d'expressivité), les interprètes devaient fournir les deux transformations dont ils étaient le plus satisfaits. Nous avons ensuite sélectionné la transformation que nous jugions la mieux réussie, afin de garder un jeu de 30 phrases chironomiques par interprète. Les données chironomiques de fréquence fondamentale ont ensuite été réutilisées dans le vocoder STRAIGHT afin d'obtenir des signaux TTS et chironomiques qui aient été synthétisés par le même vocoder. Pour les tests perceptifs qui seront présentés dans la section suivante, nous disposions donc de 30 phrases TTS, 30 phrases chironomiques (CHIRO), et 30 phrases naturelles (NAT, les phrases d'évaluation). La FIGURE A.1 montre la procédure mise en place pour l'obtention des signaux modifiés par chironomie.

## A.3 Évaluation de l'apport du contrôle chironomique

Deux tests perceptifs ont été mis en place : un test de reconnaissance du type d'expressivité, lors duquel les sujets devaient écouter un stimuli et indiquer de quel type d'expressivité il s'agissait (colère, peur, joie, tristesse, sensualité et surprise), et un test d'évaluation de la qualité, lors duquel les sujets devaient noter la qualité globale du signal entendu. 21 sujets ont participé à ces deux tests, en commençant par le test de reconnaissance. Les 5 mêmes phrases pour chacun des 6 types d'expressivité ont été utilisées, produites par l'actrice professionnelle (NAT), le système LIPS$^3$ (TTS) et par modification chironomique avec Vokinesis (CHIRO).

### A.3.1 Reconnaissance de l'expressivité

Les phrases expressives ont été présentées aux sujets dans un ordre aléatoire. Il leur était demandé de sélectionner le type d'expressivité qu'ils pensaient reconnaitre parmi les 6 types considérés. Les résultats ont été exprimés en tant que score de reconnaissance binaire, et stockés dans une matrice de contingence. Une régression a été utilisée pour l'analyse de l'influence relative des facteurs suivants : Expressivité cible (Expr : 6 niveaux), le sytème de production (Prod : 3 niveaux) et la phrase (Sent : 5 niveaux). L'influence individuelle des sujets a été modélisée par un facteur aléatoire. La librairie `lme40` de R a été utilisée [Bates *et al.* 2014, Baayen 2008]. Les résultats de l'analyse sont présentés dans le TABLEAU A.1.

TABLEAU A.1 – *Analyse de la déviance de sortie de la régression logistique appliquée au résultat du test de reconnaissance (voir le texte pour la description des labels). La significativité (p) est évaluée avec une distribution $\chi^2$, en fonction du degré de liberté (df) des facteurs et de leur interaction.*

| **Facteur** | $\chi^2$ | *df* | *p* |
|---|---|---|---|
| Expr | 69.5 | 5 | <0.0001 |
| Prod | 164.8 | 2 | <0.0001 |
| Sent | 34.3 | 4 | <0.0001 |
| Expr : Prod | 111.2 | 10 | <0.0001 |
| Expr : Sent | 64.2 | 20 | <0.0001 |
| Prod : Sent | 19.2 | 8 | <0.05 |

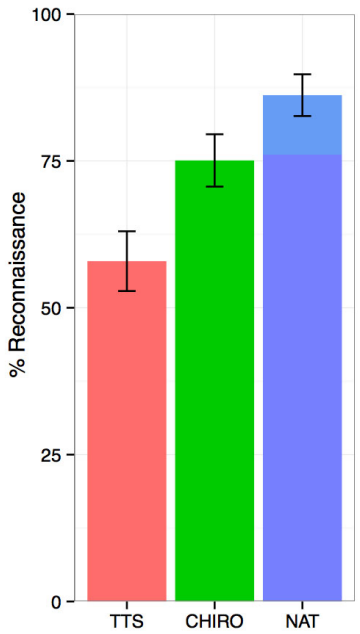


FIGURE A.2 – *Moyenne des résultats de reconnaissance pour l'interaction entre les conditions expressives et les systèmes de production.*

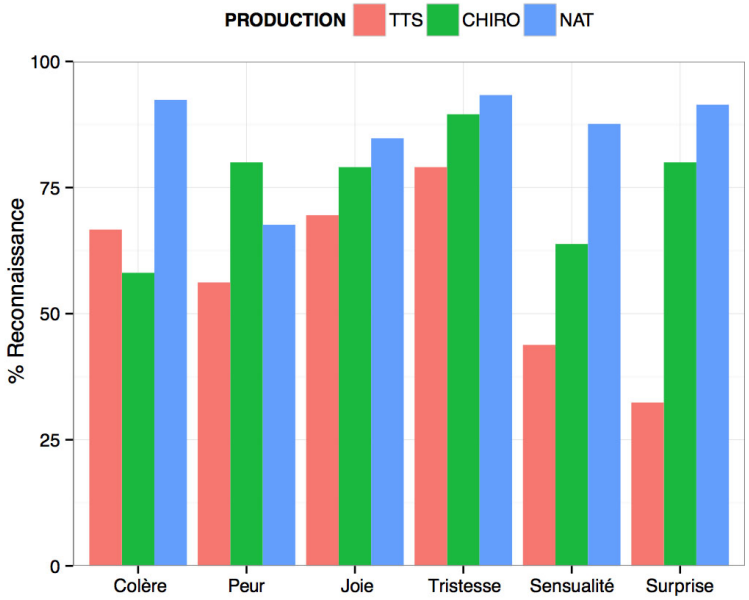


FIGURE A.3 – *Pourcentage de reconnaissance du type d'expressivité pour les phrases naturelles (NAT), les phrases du système Text-to-Speech (TTS), et les phrases transformées par chironomie (CHIRO).*

Tous les facteurs, et leurs interactions, ont un effet significatif sur le score de reconnaissance. Les différents systèmes de production (NAT, TTS et CHIRO) expliquent la majeure partie de la déviance, suivi de la condition expressive présentée et de l'interaction entre ces deux facteurs. Les types d'expressivité des phrases produites par le système TTS ont été reconnus à 58%, alors ce score dépasse les 75% pour le système CHIRO, et les 86% pour le naturel (NAT) (voir la FIGURE A.2).

La FIGURE A.3 montre que la reconnaissance dépend également du type d'expressivité. La plupart du temps, les performances chironomiques améliorent les scores de reconnaissance. Sauf pour le cas de la colère, pour lequel le score de reconnaissance a été abaissé de 9%, les modifications chironomiques de la hauteur ont aidé les sujets à reconnaitre les types d'expressivité. Elle est même meilleure que le naturel dans le cas de la peur. Le cas de la surprise constitue la plus grande amélioration du contrôle chironomique : le score de reconnaissance est supérieur de 48% à celui du système TTS.

Une analyse de la matrice de contingence a permis de montrer que les sujets effectuaient peu d'erreurs systématiques de reconnaissance. Une étude de classification a montré que les confusions importantes entre types d'expressivité n'étaient présentes que dans les phrases produites par le système TTS. Ceci peut être expliqué par la proximité des paramètres acoustiques de ces signaux, en termes de fréquence fondamentale moyenne notamment. Les sujets ont eu tendance à confondre la sensualité avec la tristesse, toutes deux produites avec une fréquence fondamentale assez basse. par ailleurs, le souffle présent dans les phrases sensuelles naturelles n'était pas très

bien reproduit par le système TTS. La peur et la surprise, qui possèdent une fréquence fondamentale élevée, étaient respectivement confondues avec la colère et la peur, mais pas avec la joie, malgré sa haute fréquence fondamentale. Ceci peut être expliqué par le fait que le système TTS ait très bien assimilé les variations acoustiques liées au sourire (augmentation des fréquences des formants [Tartter 1980]).

### A.3.2 Évaluation de la qualité

Il est fréquent que les traitement de signaux dégradent la qualité perçue de la parole. Afin d'évaluer l'effet de la modification chironomique appliquée aux signaux de sortie du système TTS sur la qualité globale, nous avons demandé aux sujets de juger la qualité perçue des mêmes 5 phrases pour les 6 types d'expressivité et les 3 systèmes de production, par une note pouvant aller de 1 à 5. Le TABLEAU A.2 présente les résultats d'une analyse ANOVA ayant comme facteurs les types d'expressivité (Expr : six niveaux), le système de production (Prod : trois niveaux) et la phrase (Sent : cinq niveaux).

Tous les facteurs ont un effet significatif sur les scores moyens obtenus (ou MOS pour *Mean Opinion Score*). Cependant, le système de production a la plus grande puissance explicative ($\eta_p^2 = 0.60$), suivi par le type d'expressivité ($\eta_p^2 = 0.09$). L'effet induit par le système de production est présenté FIGURE A.5. On peut y voir que les scores obtenus pour les phrases naturelles sont meilleures que celles obtenues pour la modification chironomique, elles-mêmes meilleures que pour les phrases TTS. La significativité de ces différents scores a été validée par un test post-hoc de Tukey. L'interaction entre le type d'expressivité et le système de production montre que l'amélioration de la qualité des phrases TTS par le contrôle chironomique est toujours observée. La FIGURE A.5 montre que les expressivités synthétiques comportant de hautes valeurs de fréquence fondamentale (colère, peur, joie, surprise) ont été jugées moins bonnes que la tristesse et la sensualité, caractérisées par des fréquences fondamentales plus basses.

TABLEAU A.2 – *L'analyse de la variance (ANOVA) expliquée pour chacun des facteurs (voir le texte pour la description des labels) et leurs interactions sur le test MOS. Les résultats incluent le test F pour les facteurs et le degré de liberté (df) de l'erreur, la valeur-p associée et la taille d'effet ($\eta_p^2$)*

| Class | *df* | erreur *df* | *F* | *p* | $\eta_p^2$ |
|---|---|---|---|---|---|
| Expre | 5 | 1840 | 36.3 | <0.001 | 0.09 |
| Prod | 2 | 1840 | 1405.4 | <0.001 | 0.60 |
| Sent | 4 | 1840 | 7.8 | <0.001 | 0.02 |
| Expr :Prod | 10 | 1840 | 7.9 | <0.001 | 0.04 |
| Expr :Sent | 20 | 1840 | 1.6 | <0.05 | 0.02 |
| Prod :Sent | 8 | 1840 | 6.4 | <0.001 | 0.03 |

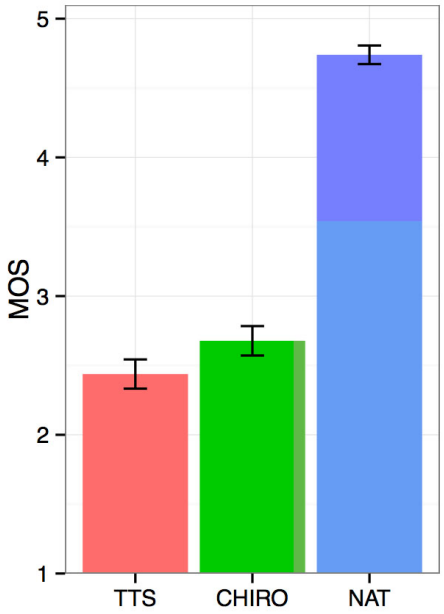


FIGURE A.4 – *Moyenne des MOS pour l'interaction entre les conditions expressives et les systèmes de production.*

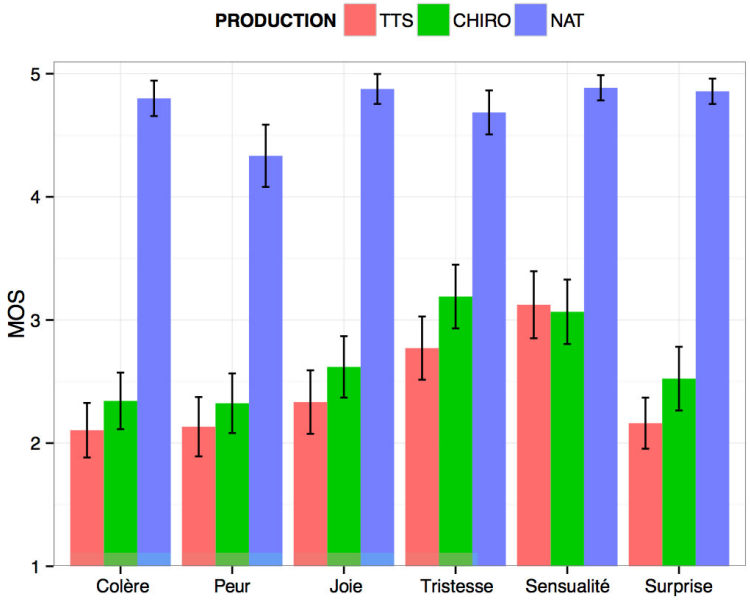


FIGURE A.5 – *MOS détaillé pour l'interaction entre les conditions expressives et les systèmes de production.*

## A.4 Discussion

Cette étude nous permet principalement de conclure que la modification chironomique de la hauteur améliore significativement l'expressivité et, à moindre mesure, la qualité globale de phrases issues d'un système TTS de parole expressive. Le contrôle chironomique semble donc prometteur pour l'amélioration du naturel, de la qualité et de l'expressivité de la synthèse de la parole. L'étude des résultats a permis de mettre en évidence les forces et les faiblesses du système TTS utilisé et de Vokinesis.

Le système TTS a obtenu les plus grandes différences de taux de reconnaissance (voir FIGURE A.3), avec la surprise et la sensualité inférieures à 50%, alors que la tristesse dépasse les 75%. Ces faibles résultats pourraient être expliqués d'une part par l'incapacité du système TTS à capturer les caractéristiques du souffle pour la voix sensuelle, et d'autre part par la mauvaise modélisation de l'intonation pour la surprise. En effet, pour ces deux types d'expressivité, l'apport de la chironomie est contrasté. La forte amélioration du score de reconnaissance pour la surprise indique qu'une bonne modélisation de la courbe de hauteur est très importante pour cette catégorie. Par contre, la faible amélioration du score de reconnaissance pour la sensualité indique que la courbe de hauteur ne joue pas un rôle aussi important que pour la surprise, et qu'une meilleure modélisation des paramètres de qualité de voix serait nécessaire. Ce cas peut également être observé pour la colère, pour laquelle le taux de reconnaissance est dégradé par la chironomie. Au contraire, les performances de la chironomie pour la peur sont encore meilleures que celles de la voix naturelle. Ceci peut sans doute être expliqué par le fait que l'exagération de ce sentiment soit facilement effectuée par des tremblements de la main dans le contrôle chironomique, impliquant des variations de hauteur rapides qui créent un tremblement de la voix.

L'utilisation de Vokinesis et d'un système de synthèse paramétrique offre un nouveau paradigme d'étude pour des applications de recherche dans le domaine de la prosodie et de l'analyse de la voix expressive. L'étude des rôles respectifs des différents paramètres prosodiques sur le rendu expressif d'une phrase est depuis longtemps considérée et débattue (voir par exemple [Ladd *et al.* 1985, Greenberg *et al.* 2006, Goudbeek & Scherer 2010, Bänziger *et al.* 2012, de Moraes & Rilliard 2014]), mais a souvent été limitée par la complexité de la modification indépendante des différentes composantes de la prosodie et de la qualité vocale. Il est aujourd'hui possible d'assigner, en plus de la hauteur, l'un des paramètres de durée ou de qualité vocale présents dans le système TTS à l'un des paramètres de contrôle de la tablette ou d'une autre interface connectée à Vokinesis, afin d'en permettre la manipulation. Par exemple, les mauvais résultats concernant la colère auraient sans doute pu être améliorés si les interprètes avaient eu la possibilité de modifier d'autres paramètres tels que la force de voix [Liénard & Barras 2013], ou encore d'avoir un contrôle sur le rythme syllabique pour produire des rythmes staccato [Kehrein 2002].

En ce qui concerne la qualité générale, le test MOS résumé dans la FIGURE A.5 montre que les manipulations chironomiques apportent une légère amélioration. Cela peut être expliqué par deux éléments. Tout d'abord, le même vocoder a été

utilisé pour les signaux TTS et chironomiques : les données chironomiques ont été appliquées au vecteur paramétrique utilisé en entrée du vocoder du système TTS afin de s'affranchir de toute perte de qualité liée aux algorithmes de traitement de signal de l'un des systèmes. Ensuite, les courbes de fréquence fondamentale du système HMM-TTS possèdent des variations microprosodiques importantes et peu naturelles, alors que les mouvements continus de la main fournissent des courbes de fréquence fondamentale plus lisses, qui seront alors perçues comme plus naturelles.

Pour conclure, les interfaces chironomiques permettent d'améliorer la qualité générale et l'expressivité de signaux de parole obtenus par des systèmes HMM-TTS. La modification chironomique est un outil polyvalent qui peut être utilisé pour des application de recherche (expériences de modification prosodique), mais également pour des applications industrielles (amélioration de l'expressivité pour des livres audio, des doublages de jeux vidéo...)

Il serait à présent intéressant de renouveler l'expérience en offrant la possibilité aux interprètes de contrôler d'autres paramètres tels que le rythme syllabique, la force de voix, la tension vocale, le souffle...

# Bibliographie


[Ardaillon *et al.* 2015] Luc Ardaillon, Gilles Degottex and Axel Roebel. *A multi-layer F0 model for singing voice synthesis using a B-spline representation with intuitive controls.* In Interspeech 2015, 2015. (Cité page 138.)

[Astrinaki *et al.* 2012] Maria Astrinaki, Nicolas d'Alessandro, Benjamin Picart, Thomas Drugman and Thierry Dutoit. *Reactive and continuous control of HMM-based speech synthesis.* In Spoken language technology workshop (SLT), 2012 IEEE, pages 252–257. IEEE, 2012. (Cité pages 2, 9, 11, 78 et 154.)

[Atal & Hanauer 1971] Bishnu S Atal and Suzanne L Hanauer. *Speech analysis and synthesis by linear prediction of the speech wave.* The journal of the acoustical society of America, vol. 50, no. 2B, pages 637–655, 1971. (Cité page 16.)

[Baayen 2008] R Harald Baayen. Analyzing linguistic data : A practical introduction to statistics using R. Cambridge University Press, 2008. (Cité page 167.)

[Balazs *et al.* 2011] Peter Balazs, Monika Dörfler, Florent Jaillet, Nicki Holighaus and G Velasco. *Theory, implementation and applications of nonstationary Gabor frames.* Journal of computational and applied mathematics, vol. 236, no. 6, pages 1481–1496, 2011. (Cité page 17.)

[Bänziger *et al.* 2012] Tanja Bänziger, Marcello Mortillaro and Klaus R Scherer. *Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception.* Emotion, vol. 12, no. 5, page 1161, 2012. (Cité pages 166 et 172.)

[Barbosa & Bailly 1994] Plínio Barbosa and Gérard Bailly. *Characterisation of rhythmic patterns for text-to-speech synthesis.* Speech Communication, vol. 15, no. 1-2, pages 127–137, 1994. (Cité pages 55 et 153.)

[Barbosa *et al.* 2005] Plínio A Barbosa, Pablo Arantes, Alexsandro R Meireles and Jussara M Vieira. *Abstractness in speech-metronome synchronisation : P-centres as cyclic attractors.* In Interspeech, pages 1441–1444, 2005. (Cité page 48.)

[Bartkova & Sorin 1987] Katarina Bartkova and Christel Sorin. *A model of segmental duration for speech synthesis in French.* Speech communication, vol. 6, no. 3, pages 245–260, 1987. (Cité page 48.)

[Bates *et al.* 2014] Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker*et al. lme4 : Linear mixed-effects models using Eigen and S4.* R package version, vol. 1, no. 7, pages 1–23, 2014. (Cité page 167.)

[Beaudouin-Lafon 2004] Michel Beaudouin-Lafon. *Designing interaction, not interfaces.* In Proceedings of the working conference on Advanced visual interfaces, pages 15–22. ACM, 2004. (Cité page 56.)

[Beaudouin-Lafon 2016] Michel Beaudouin-Lafon. *Mieux penser les interfaces informatiques*. 2016. (Cité page 98.)

[Berthoz 1997] Alain Berthoz. Sens du mouvement (le). Odile Jacob, 1997. (Cité pages 63, 64 et 65.)

[Boersma *et al.* 2002] Paul Boersma*et al.* *Praat, a system for doing phonetics by computer*. Glot international, vol. 5, no. 9/10, pages 341–345, 2002. (Cité page 129.)

[Bonada 2000] Jordi Bonada. *Automatic technique in frequency domain for near-lossless time-scale modification of audio*. In International Computer Music Conference, 2000. (Cité page 17.)

[Broca 1861] Paul Broca. *Remarques sur le siège de la faculté du langage articulé, suivies d'une observation d'aphémie (perte de la parole)*. Bulletin et Mémoires de la Société anatomique de Paris, vol. 6, pages 330–357, 1861. (Cité page 162.)

[Browman & Goldstein 1990a] Catherine P Browman and Louis Goldstein. *Gestural specification using dynamically-defined articulatory structures*. DOCUMENT RESUME ED 331 100 CS 507 425, page 95, 1990. (Cité pages 49, 50 et 154.)

[Browman & Goldstein 1990b] Catherine P Browman and Louis Goldstein. *Tiers in articulatory phonology, with some implications for casual speech*. Papers in laboratory phonology I : Between the grammar and physics of speech, pages 341–376, 1990. (Cité pages 50, 51 et 154.)

[Browman & Goldstein 1992] Catherine P Browman and Louis Goldstein. *Articulatory phonology : An overview*. Phonetica, vol. 49, no. 3-4, pages 155–180, 1992. (Cité pages 50, 52, 59, 65 et 154.)

[Canault 2007] Mélanie Canault. *L'émergence du contrôle articulatoire au stade du babillage. Une étude acoustique et cinématique*. PhD thesis, Université Marc Bloch-Strasbourg II, 2007. (Cité page 53.)

[Chan 2003] Marjorie KM Chan. *The digital age and speech technology for Chinese language teaching and learning*. Journal-Chinese Language Teachers Association, vol. 38, no. 2, pages 49–86, 2003. (Cité page 156.)

[Chen 1993] Quinghai Chen. *Toward a Sequential Approach for Tonal Error Analysis*. In Deseret Language and Linguistic Society Symposium, volume 19, page 8, 1993. (Cité page 156.)

[Chiang 1979] Th Chiang. *Some interferences of English intonation with Chinese tones*. IRAL : International Review of Applied Linguistics in Language Teaching, vol. 17, no. 3, page 245, 1979. (Cité page 156.)

[Chun *et al.* 2012] Dorothy M Chun, Yan Jiang and Natalia Ávila. *Visualization of tone for learning Mandarin Chinese*. In Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference, pages 77–89, 2012. (Cité page 156.)

[Cook 1991] Perry Cook. *Identification of control parameters in an articulatory vocal tract model, with applications to the synthesis of singing.* PhD thesis, Stanford University, 1991. (Cité page 5.)

[Cook 1993] Perry R Cook. *SPASM, a real-time vocal tract physical model controller ; and singer, the companion software synthesis system.* Computer Music Journal, vol. 17, no. 1, pages 30–44, 1993. (Cité pages 2 et 6.)

[Cook 2005] Perry R Cook. *Real-time performance controllers for synthesized singing.* In Proceedings of the 2005 conference on New interfaces for musical expression, pages 236–237. National University of Singapore, 2005. (Cité page 6.)

[D'Alessandro & Dutoit 2007] Nicolas D'Alessandro and Thierry Dutoit. *Handsketch bi-manual controller : Investigation on expressive control issues of an augmented tablet.* In Proceedings of the 7th international conference on New interfaces for musical expression, pages 78–81. ACM, 2007. (Cité pages 2, 9, 11, 78 et 154.)

[D'Alessandro & Dutoit 2009] Nicolas D'Alessandro and Thierry Dutoit. *Advanced techniques for vertical tablet playing : an overview of two years of practicing the HandSketch.* In NIME, pages 173–174, 2009. (Cité pages 9 et 11.)

[d'Alessandro *et al.* 2005] Christophe d'Alessandro, Nicolas D'Alessandro, S Le Beux, Juraj Simko, Feride Çetin and Hannes Pirker. *The speech conductor : gestural control of speech synthesis.* In eINTERFACE'05-Summer Workshop on Multimodal Interfaces, 2005. (Cité page 9.)

[D'Alessandro *et al.* 2006a] Nicolas D'Alessandro, Christophe d'Alessandro, S Le Beux and Boris Doval. *Real-time CALM synthesizer new approaches in hands-controlled voice synthesis.* In Proceedings of the 2006 conference on New interfaces for musical expression, pages 266–271. IRCAM ?Centre Pompidou, 2006. (Cité page 9.)

[d'Alessandro *et al.* 2006b] Nicolas d'Alessandro, Boris Doval, S Le Beux, P Woodruff and Y Fabre. *Ramcess : Realtime and accurate musical control of expression in singing synthesis.* In eINTERFACE'06-SIMILAR NoE Summer Workshop on Multimodal Interfaces, 2006. (Cité page 9.)

[d'Alessandro *et al.* 2011] Christophe d'Alessandro, Albert Rilliard and Sylvain Le Beux. *Chironomic stylization of intonation.* The Journal of the Acoustical Society of America, vol. 129, no. 3, pages 1594–1604, 2011. (Cité pages 70, 72, 78 et 154.)

[d'Alessandro *et al.* 2014] Christophe d'Alessandro, Lionel Feugere, Sylvain Le Beux, Olivier Perrotin and Albert Rilliard. *Drawing melodies : Evaluation of chironomic singing synthesis.* The Journal of the Acoustical Society of America, vol. 135, no. 6, pages 3601–3612, 2014. (Cité pages 79, 91 et 154.)

[d'Alessandro 2009] Nicolas d'Alessandro. *Realtime and Accurate Musical Control of Expression in Voice Synthesis.* PhD thesis, 2009. (Cité page 9.)

[de Moraes & Rilliard 2014] João Antônio de Moraes and Albert Rilliard. *Illocution, attitudes and prosody.* Spoken Corpora and Linguistic Studies, vol. 61, page 233, 2014. (Cité page 172.)

[Dörfler 2011] Monika Dörfler. *Quilted Gabor frames–A new concept for adaptive time-frequency representation.* Advances in Applied Mathematics, vol. 47, no. 4, pages 668–687, 2011. (Cité page 17.)

[Doval & d'Alessandro 1997] Boris Doval and Christophe d'Alessandro. *Spectral correlates of glottal waveform models : an analytic study.* In IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 2, pages 1295–1298, 1997. (Cité page 19.)

[Doval *et al.* 2003] Boris Doval, Christophe d'Alessandro and Nathalie Henrich. *The voice source as a causal/anticausal linear filter.* In ISCA Tutorial and Research Workshop on Voice Quality : Functions, Analysis and Synthesis, 2003. (Cité page 36.)

[Doval *et al.* 2006] Boris Doval, Christophe d'Alessandro and Nathalie Henrich. *The spectrum of glottal flow models.* Acta acustica united with acustica, vol. 92, no. 6, pages 1026–1046, 2006. (Cité pages 19, 35 et 155.)

[Dudley *et al.* 1939] Homer Dudley, RR Riesz and SSA Watkins. *A synthetic speaker.* Journal of the Franklin Institute, vol. 227, no. 6, pages 739–764, 1939. (Cité pages 4 et 5.)

[Dudley 1939] Homer Dudley. *Remaking speech.* The Journal of the Acoustical Society of America, vol. 11, no. 2, pages 169–177, 1939. (Cité pages 2, 15 et 17.)

[Dutoit *et al.* 1996] Thierry Dutoit, Vincent Pagel, Nicolas Pierret, François Bataille and Olivier Van der Vrecken. *The MBROLA project : Towards a set of high quality speech synthesizers free of use for non commercial purposes.* In Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, volume 3, pages 1393–1396. IEEE, 1996. (Cité page 19.)

[Duxbury *et al.* 2002] Chris Duxbury, Mike Davies and Mark B Sandler. *Improved time-scaling of musical audio using phase locking at transients.* In Audio Engineering Society Convention 112. Audio Engineering Society, 2002. (Cité page 17.)

[Evangelista *et al.* 2012] Gianpaolo Evangelista, Monika Dörfler and Ewa Matusiak. *Phase vocoders with arbitrary frequency band selection.* In Proceedings of the 9th Sound and Music Computing Conference (SMC'12), Copenhagen, 2012. (Cité page 17.)

[Evrard *et al.* 2015] Marc Evrard, Samuel Delalez, Christophe d'Alessandro and Albert Rilliard. *Comparison of chironomic stylization versus statistical modeling of prosody for expressive speech synthesis.* In Sixteenth Annual Conference of the International Speech Communication Association, 2015. (Cité page 165.)

[Evrard 2015] Marc Evrard. *Synthèse de parole expressive à partir du texte : Des phonostyles au contrôle gestuel pour la synthèse paramétrique statistique.* PhD thesis, Paris 11, 2015. (Cité pages 19 et 165.)

[Fant 1970] Gunnar Fant. Acoustic theory of speech production. Mouton, 1970. (Cité pages 14 et 15.)

[Fels & Hinton 1993] S Sidney Fels and Geoffrey E Hinton. *Glove-Talk : a neural network interface between a data-glove and a speech synthesizer.* IEEE transactions on Neural Networks, vol. 4, no. 1, pages 2–8, 1993. (Cité pages 6 et 154.)

[Fels & Hinton 1998] S Sidney Fels and Geoffrey E Hinton. *Glove-Talk II : a neural-network interface which maps gestures to parallel formant speech synthesizer controls.* IEEE transactions on neural networks, vol. 9, no. 1, pages 205–212, 1998. (Cité pages 2, 6, 64 et 154.)

[Feugère *et al.* 2017] Lionel Feugère, Christophe d'Alessandro, Boris Doval and Olivier Perrotin. *Cantor Digitalis : chironomic parametric synthesis of singing.* EURASIP Journal on Audio, Speech, and Music, 2017. (Cité pages 2, 10, 11, 36, 78, 81, 121 et 154.)

[Feugere 2013] Lionel Feugere. *Synthèse par règles de la voix chantée contrôlée par le geste et applications musicales.* PhD thesis, Université Pierre et Marie Curie-Paris VI, 2013. (Cité page 10.)

[Fiebrink & Cook 2010] Rebecca Fiebrink and Perry R Cook. *The Wekinator : a system for real-time, interactive machine learning in music.* In Proceedings of The Eleventh International Society for Music Information Retrieval Conference (ISMIR 2010)(Utrecht), 2010. (Cité page 63.)

[Flanagan & Golden 1966] James L Flanagan and RM Golden. *Phase vocoder.* Bell Labs Technical Journal, vol. 45, no. 9, pages 1493–1509, 1966. (Cité page 16.)

[Goldman 2011] Jean-Philippe Goldman. *EasyAlign : an automatic phonetic alignment tool under Praat.* In Interspeech, 2011. (Cité page 129.)

[Goudbeek & Scherer 2010] Martijn Goudbeek and Klaus Scherer. *Beyond arousal : Valence and potency/control cues in the vocal expression of emotion.* The Journal of the Acoustical Society of America, vol. 128, no. 3, pages 1322–1336, 2010. (Cité page 172.)

[Greenberg *et al.* 2006] Yoko Greenberg, M Tsuzaki, K Kato and Y Sagisaka. *A trial of communicative prosody generation based on control characteristic of one word utterance observed in real conversational speech.* In Proc. Speech Prosody, volume 2006, pages 37–40, 2006. (Cité page 172.)

[Hamon *et al.* 1989] Christian Hamon, E Mouline and Francis Charpentier. *A diphone synthesis system based on time-domain prosodic modifications of speech.* In Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on, pages 238–241. IEEE, 1989. (Cité page 14.)

[Henrich 2001] Nathalie Henrich. *Etude de la source glottique en voix parlée et chantée : modélisation et estimation, mesures acoustiques et électroglottographiques, perception.* PhD thesis, 2001. (Cité pages 15 et 19.)

[Hunt *et al.* 2003] Andy Hunt, Marcelo M Wanderley and Matthew Paradis. *The importance of parameter mapping in electronic instrument design.* Journal of New Music Research, vol. 32, no. 4, pages 429–440, 2003. (Cité page 107.)

[IMA 1983] International MIDI Association IMA. *MIDI musical instrument digital interface specification 1.0.* Los Angeles, 1983. (Cité page 85.)

[Imai 1983] Satoshi Imai. *Cepstral analysis synthesis on the mel frequency scale.* In Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'83., volume 8, pages 93–96. IEEE, 1983. (Cité page 16.)

[Itakura 1970] Fumitada Itakura. *A statistical method for estimation of speech spectral density and formant frequency.* IEICE Trans., vol. 53, no. 1, pages 35–42, 1970. (Cité page 16.)

[Itakura 1975] Fumitada Itakura. *Line spectrum representation of linear predictor coefficients of speech signals.* The Journal of the Acoustical Society of America, vol. 57, no. S1, pages S35–S35, 1975. (Cité page 16.)

[Jaillet & Torrésani 2007] Florent Jaillet and Bruno Torrésani. *Time-frequency jigsaw puzzle : Adaptive multiwindow and multilayered Gabor expansions.* International Journal of Wavelets, Multiresolution and Information Processing, vol. 5, no. 02, pages 293–315, 2007. (Cité page 17.)

[Katamba 1989] Francis Katamba. An introduction to phonology, volume 48. Longman London, 1989. (Cité page 46.)

[Kawahara *et al.* 1999] Hideki Kawahara, Ikuyo Masuda-Katsuse and Alain De Cheveigne. *Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction : Possible role of a repetitive structure in sounds.* Speech communication, vol. 27, no. 3, pages 187–207, 1999. (Cité pages 16 et 166.)

[Kehrein 2002] Roland Kehrein. Prosodie und emotionen, volume 231. Walter de Gruyter, 2002. (Cité pages 40, 49 et 172.)

[Kenmochi & Ohshita 2007] Hideki Kenmochi and Hayato Ohshita. *VOCALOID-commercial singing synthesizer based on sample concatenation.* In Interspeech, volume 2007, pages 4009–4010, 2007. (Cité page 8.)

[Kessous 2004a] Loic Kessous. *Contrôles gestuels bi-manuels de processus sonores.* 2004. (Cité pages 8 et 11.)

[Kessous 2004b] Loic Kessous. *Gestural control of singing voice, a musical instrument.* Proceedings of Sound and Music Computing, 2004. (Cité pages 8, 11, 78 et 154.)

[Kiriloff 1969] Constantine Kiriloff. *On the auditory perception of tones in Mandarin.* Phonetica, vol. 20, no. 2-4, pages 63–67, 1969. (Cité page 156.)

[Krishnan *et al.* 2004] Ananthanarayan Krishnan, Yisheng Xu, Jackson T Gandour and Peter A Cariani. *Human frequency-following response : representation of pitch contours in Chinese tones.* Hearing research, vol. 189, no. 1, pages 1–12, 2004. (Cité page 157.)

[Kuwabara 1996] Hisao Kuwabara. *Acoustic properties of phonemes in continuous speech for different speaking rate.* In Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, volume 4, pages 2435–2438. IEEE, 1996. (Cité page 48.)

[Ladd *et al.* 1985] D Robert Ladd, Kim EA Silverman, Frank Tolkmitt, Günther Bergmann and Klaus R Scherer. *Evidence for the independent function of intonation contour type, voice quality, and F 0 range in signaling speaker affect.* The Journal of the Acoustical Society of America, vol. 78, no. 2, pages 435–444, 1985. (Cité page 172.)

[Laroche & Dolson 1999] Jean Laroche and Mark Dolson. *New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects.* In Applications of Signal Processing to Audio and Acoustics, 1999 IEEE Workshop on, pages 91–94. IEEE, 1999. (Cité page 17.)

[Le Beux *et al.* 2007] Sylvain Le Beux, Albert Rilliard and Christophe d'Alessandro. *Calliphony : a real-time intonation controller for expressive speech synthesis.* In SSW, pages 345–350, 2007. (Cité pages 2, 10, 41, 78 et 154.)

[Le Beux *et al.* 2010] Sylvain Le Beux, Boris Doval and Christophe d'Alessandro. *Issues and solutions related to real-time TD-PSOLA implementation.* In Audio Engineering Society Convention 128. Audio Engineering Society, 2010. (Cité pages 14, 18, 26, 28, 37 et 155.)

[Le Beux 2009] Sylvain Le Beux. *Contrôle gestuel de la prosodie et de la qualité vocale.* PhD thesis, Université Paris Sud - Paris XI, 2009. (Cité pages 9, 10 et 41.)

[Le Maguer *et al.* 2013] Sébastien Le Maguer, Nelly Barbot, Olivier Boeffard*et al.* *Evaluation of contextual descriptors for HMM-based speech synthesis in French.* In SSW, pages 153–158, 2013. (Cité page 166.)

[Léon 1993] Pierre R Léon. Précis de phonostylistique : parole et expressivité. Nathan, 1993. (Cité page 166.)

[Levine & Smith III 1998] Scott N Levine and Julius O Smith III. *A sines+ transients+ noise audio representation for data compression and time/pitch scale modifications.* In Audio Engineering Society Convention 105. Audio Engineering Society, 1998. (Cité page 18.)

[Levitt 1991] Andrea G Levitt. *Reiterant speech as a test of non-native speakers' mastery of the timing of French.* The Journal of the Acoustical Society of America, vol. 90, no. 6, pages 3008–3018, 1991. (Cité page 71.)

[Liénard & Barras 2013] Jean-Sylvain Liénard and Claude Barras. *Fine-grain voice strength estimation from vowel spectral cues.* In INTERSPEECH, pages 128–132, 2013. (Cité page 172.)

[Lindblom & Studdert-Kennedy 1967] Björn EF Lindblom and Michael Studdert-Kennedy. *On the role of formant transitions in vowel recognition.* The Journal of the Acoustical society of America, vol. 42, no. 4, pages 830–843, 1967. (Cité pages 58 et 61.)

[Liuni & Röbel 2013] Marco Liuni and Axel Röbel. *Phase vocoder and beyond.* Musica, Tecnologia, vol. 7, pages 73–120, 2013. (Cité page 16.)

[Liuni *et al.* 2011a] Marco Liuni, Peter Balazs and Axel Röbel. *Sound analysis and synthesis adaptive in time and two frequency bands.* arXiv preprint arXiv :1109.6651, 2011. (Cité page 17.)

[Liuni *et al.* 2011b] Marco Liuni, Axel Röbel, Marco Romito and Xavier Rodet. *A reduced multiple Gabor frame for local time adaptation of the spectrogram.* arXiv preprint arXiv :1109.6313, 2011. (Cité page 17.)

[MacNeilage 1998] Peter F MacNeilage. *The frame/content theory of evolution of speech production.* Behavioral and brain sciences, vol. 21, no. 04, pages 499–511, 1998. (Cité pages 44, 53, 59, 65, 153 et 162.)

[Mairano 2011] Paolo Mairano. *Rhythm typology : acoustic and perceptive studies.* PhD thesis, Università di Torino, 2011. (Cité page 48.)

[McAulay & Quatieri 1986] Robert McAulay and Thomas Quatieri. *Speech analysis/synthesis based on a sinusoidal representation.* IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 34, no. 4, pages 744–754, 1986. (Cité page 17.)

[Morise *et al.* 2016] Masanori Morise, Fumiya Yokomori and Kenji Ozawa. *WORLD : A vocoder-based high-quality speech synthesis system for real-time applications.* IEICE TRANSACTIONS on Information and Systems, vol. 99, no. 7, pages 1877–1884, 2016. (Cité pages 16, 19, 37 et 155.)

[Moulines & Charpentier 1990] Eric Moulines and Francis Charpentier. *Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones.* Speech communication, vol. 9, no. 5-6, pages 453–467, 1990. (Cité pages 14, 18, 20 et 37.)

[Moulines & Laroche 1995] Eric Moulines and Jean Laroche. *Non-parametric techniques for pitch-scale and time-scale modification of speech.* Speech communication, vol. 16, no. 2, pages 175–205, 1995. (Cité pages 14, 18, 20, 21, 25, 28 et 37.)

[Ortega *et al.* 2014] Laura Ortega, Emmanuel Guzman-Martinez, Marcia Grabowecky and Satoru Suzuki. *Audition dominates vision in duration perception irrespective of salience, attention, and temporal discriminability.* Attention, Perception, & Psychophysics, vol. 76, no. 5, pages 1485–1502, 2014. (Cité page 56.)

[Orton 2011] Jane Orton. *Educating Chinese language teachers–Some fundamentals.* Teaching and learning Chinese in global contexts : CFL worldwide, pages 151–164, 2011. (Cité page 156.)

[Perrotin & d'Alessandro 2013] Olivier Perrotin and Christophe d'Alessandro. *Adaptive mapping for improved pitch accuracy on touch user interfaces.* In NIME, pages 186–189, 2013. (Cité pages 80 et 126.)

[Perrotin & D'alessandro 2016a] Olivier Perrotin and Christophe D'alessandro. *Seeing, listening, drawing : interferences between sensorimotor modalities in the use of a tablet musical interface.* ACM Transactions on Applied Perception (TAP), vol. 14, no. 2, page 10, 2016. (Cité pages 56, 95 et 155.)

[Perrotin & d'Alessandro 2016b] Olivier Perrotin and Christophe d'Alessandro. *Vocal effort modification for singing synthesis.* Spectrum, vol. 100, page 50, 2016. (Cité page 19.)

[Perrotin 2015] Olivier Perrotin. *Chanter avec les mains : interfaces chironomiques pour les instruments de musique numériques.* PhD thesis, Université Paris Sud-Paris XI, 2015. (Cité pages 3, 9, 80, 121, 139 et 154.)

[Pompino-Marschall 1989] Bernd Pompino-Marschall. *On the psychoacoustic nature of the P-center phenomenon.* Journal of phonetics, 1989. (Cité page 48.)

[Pöppel 1994] Ernst Pöppel. *TENAPORAL MECHANISMS IN PERCEPTION.* Selectionism and the Brain, vol. 37, page 185, 1994. (Cité page 45.)

[Puckette *et al.* 1998] Miller S Puckette, Miller S Puckette Ucsd, Theodore Apel*et al. Real-time audio analysis tools for Pd and MSP.* 1998. (Cité page 120.)

[Quatieri & McAulay 1986] T Quatieri and Rl McAulay. *Speech transformations based on a sinusoidal representation.* IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 34, no. 6, pages 1449–1464, 1986. (Cité page 18.)

[Ramstein 1991] Christophe Ramstein. *Analyse, représentation et traitement du geste instrumental : application aux instruments à clavier.* PhD thesis, Institut National Polytechnique de Grenoble-INPG, 1991. (Cité page 60.)

[Repp 2005] Bruno H Repp. *Sensorimotor synchronization : a review of the tapping literature.* Psychonomic bulletin & review, vol. 12, no. 6, pages 969–992, 2005. (Cité page 162.)

[Rispal-Padel *et al.* 1982] L Rispal-Padel, F Cicirata and C Pons. *Cerebellar nuclear topography of simple and synergistic movements in the alert baboon (Papio papio).* Experimental Brain Research, vol. 47, no. 3, pages 365–380, 1982. (Cité page 63.)

[Röbel 2003] Axel Röbel. *A new approach to transient processing in the phase vocoder.* In 6th International Conference on Digital Audio Effects (DAFx), pages 344–349, 2003. (Cité page 17.)

[Roth 2001] Wolff-Michael Roth. *Gestures : Their role in teaching and learning.* Review of educational research, vol. 71, no. 3, pages 365–392, 2001. (Cité page 156.)

[Rovan *et al.* 1997] Joseph Butch Rovan, Marcelo M Wanderley, Shlomo Dubnov and Philippe Depalle. *Instrumental gestural mapping strategies as expressi-*

*vity determinants in computer music performance.* In Proceedings of Kansei-The Technology of Emotion Workshop, pages 3–4, 1997. (Cité page 107.)

[Rudoy *et al.* 2010] Daniel Rudoy, Prabahan Basu and Patrick J Wolfe. *Superposition frames for adaptive time-frequency analysis and fast reconstruction.* IEEE Transactions on Signal Processing, vol. 58, no. 5, pages 2581–2596, 2010. (Cité page 17.)

[Rye & Holmes 1982] JM Rye and John Nicholas Holmes. *A versatile software parallel-formant speech synthesizer.* Joint Speech Res. Unit, Malvern, UK, Tech. Rep. JSRU-RR-1016, 1982. (Cité page 6.)

[Schafer & Rabiner 1970] Ronald W Schafer and Lawrence R Rabiner. *System for automatic formant analysis of voiced speech.* The Journal of the Acoustical Society of America, vol. 47, no. 2B, pages 634–648, 1970. (Cité page 16.)

[Schlaug *et al.* 2008] Gottfried Schlaug, Sarah Marchina and Andrea Norton. *From singing to speaking : why singing may lead to recovery of expressive language function in patients with Broca's aphasia.* Music perception : An interdisciplinary journal, vol. 25, no. 4, pages 315–323, 2008. (Cité page 162.)

[Schlaug *et al.* 2009] Gottfried Schlaug, Sarah Marchina and Andrea Norton. *Evidence for in white-matter tracts of patients with chronic broca's aphasia undergoing intense intonation-based speech therapy.* Annals of the New York Academy of Sciences, vol. 1169, no. 1, pages 385–394, 2009. (Cité page 162.)

[Scott 1993] Sophie Kerttu Scott. *Perceptual centers in speech - An acoustic analysis.* PhD thesis, University College London (University of London), 1993. (Cité page 48.)

[Serra & Smith 1990] Xavier Serra and Julius Smith. *Spectral modeling synthesis : A sound analysis/synthesis system based on a deterministic plus stochastic decomposition.* Computer Music Journal, vol. 14, no. 4, pages 12–24, 1990. (Cité page 18.)

[Shen 1989] Xiaonan S Shen. *Toward a register approach in teaching Mandarin tones.* Journal of Chinese Language Teachers Association, vol. 24, no. 3, pages 27–47, 1989. (Cité page 156.)

[Stylianou 1996] Yannis Stylianou. *Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification.* http ://www. research. att. com/˜ styliano/, 1996. (Cité page 18.)

[Sundberg & Rossing 1990] Johan Sundberg and Thomas D Rossing. *The science of singing voice.* the Journal of the Acoustical Society of America, vol. 87, no. 1, pages 462–463, 1990. (Cité page 19.)

[Tartter 1980] Vivien C Tartter. *Happy talk : Perceptual and acoustic effects of smiling on speech.* Perception & psychophysics, vol. 27, no. 1, pages 24–27, 1980. (Cité page 170.)

[Titze & Sundberg 1992] Ingo R Titze and Johan Sundberg. *Vocal intensity in speakers and singers.* the Journal of the Acoustical Society of America, vol. 91, no. 5, pages 2936–2946, 1992. (Cité page 19.)

[Tokuda *et al.* 1994] Keiichi Tokuda, Takao Kobayashi, Takashi Masuko and Satoshi Imai. *Mel-generalized cepstral analysis-a unified approach to speech spectral estimation.* In ICSLP, volume 94, pages 18–22, 1994. (Cité page 16.)

[Tokuda *et al.* 2012] K Tokuda, K Oura, A Tamamori, S Sako, H Zen, T Nose, T Takahashi, J Yamagishi and Y Nankaku. *Speech signal processing toolkit (SPTK).* Online], recent version, 2012. (Cité page 166.)

[Tokuda *et al.* 2013] Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi and Keiichiro Oura. *Speech synthesis based on hidden Markov models.* Proceedings of the IEEE, vol. 101, no. 5, pages 1234–1252, 2013. (Cité page 166.)

[Umbert *et al.* 2015] Marti Umbert, Jordi Bonada, Masataka Goto, Tomoyasu Nakano and Johan Sundberg. *Expression control in singing voice synthesis : features, approaches, evaluation, and challenges.* IEEE Signal Processing Magazine, vol. 32, no. 6, pages 55–73, 2015. (Cité page 19.)

[Villing *et al.* 2011] Rudi C Villing, Bruno H Repp, Tomas E Ward and Joseph M Timoney. *Measuring perceptual centers using the phase correction response.* Attention, Perception, & Psychophysics, vol. 73, no. 5, pages 1614–1629, 2011. (Cité page 162.)

[Von Kempelen 1791] Wolfgang Von Kempelen. Mechanismus der menschlichen sprache. Degen, 1791. (Cité page 3.)

[Wagner 2008] P Wagner. *The rhythm of language and speech : Constraining factors, models, metrics and applications.* PhD thesis, Habilitationsschrift, University of Bonn, 2008. (Cité pages 44, 45, 47, 48, 66, 71, 74, 75 et 154.)

[Wanat *et al.* 2017] Ewa Wanat, Rachel Smith, Jane Stuart-Smith and Caroline Palmer. *The Role of Tapping in Improving Connected Speech Comprehension of a Non-Native Variety of English.* 2017. (Cité page 161.)

[Wanderley & Depalle 2004] Marcelo M Wanderley and Philippe Depalle. *Gestural control of sound synthesis.* Proceedings of the IEEE, vol. 92, no. 4, pages 632–644, 2004. (Cité pages 2 et 40.)

[Wanderley *et al.* 2000] Marcelo M Wanderley, Jean-Philippe Viollet, Fabrice Isart and Xavier Rodet. *On the Choice of Transducer Technologies for Specific Musical Functions.* In International Computer Music Conference, 2000. (Cité pages 8 et 154.)

[Wang *et al.* 1999] Yue Wang, Michelle M Spence, Allard Jongman and Joan A Sereno. *Training American listeners to perceive Mandarin tones.* The Journal of the Acoustical Society of America, vol. 106, no. 6, pages 3649–3658, 1999. (Cité page 156.)

[Wells 1965] John C Wells. *The phonological status of syllabic consonants in English RP.* Phonetica, vol. 13, no. 1-2, pages 110–113, 1965. (Cité page 45.)

[White 1981] Carolyn M White. *Tonal perception errors and interference from English intonation.* Journal of Chinese Language Teachers Association, vol. 16, no. 2, pages 27–56, 1981. (Cité page 156.)

[Wise *et al.* 2007] Timothy Wise*et al.* *Yodel species : a typology of falsetto effects in popular music vocal styles.* Radical Musicology, vol. 2, no. 2007, page 57, 2007. (Cité page 83.)

[Wright *et al.* 1997] Matthew Wright, David Wessel and Adrian Freed. *New Musical Control Structures from Standard Gestural Controllers.* In International Computer Music Conference, 1997. (Cité page 8.)

[Xiao & Ishii 2016] Xiao Xiao and Hiroshi Ishii. *Inspect, Embody, Invent : A Design Framework for Music Learning and Beyond.* In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pages 5397–5408. ACM, 2016. (Cité page 157.)

[Xiao *et al.* 2013] Xiao Xiao, Paula Aguilera, Jonathan Williams and Hiroshi Ishii. *MirrorFugue iii : conjuring the recorded pianist.* In CHI Extended Abstracts, pages 2891–2892, 2013. (Cité page 157.)

[Xiao *et al.* 2016] Xiao Xiao, Pablo Puentes, Edith Ackermann and Hiroshi Ishii. *Andantino : Teaching Children Piano with Projected Animated Characters.* In Proceedings of the The 15th International Conference on Interaction Design and Children, pages 37–45. ACM, 2016. (Cité page 157.)

[Yip 2002] Moira Yip. Tone. Cambridge University Press, 2002. (Cité page 156.)

**Titre :** Vokinesis : instrument de contrôle suprasegmental de la synthèse vocale

**Mots clés :** synthèse vocale, interactions Humain-Machine, informatique musicale, prosodie, traitement du signal vocal

Ce travail s'inscrit dans le domaine du contrôle performatif de la synthèse vocale, et plus particulièrement de la modification temps-réel de signaux de voix préenregistrés. Dans un contexte où de tels systèmes n'étaient en mesure de modifier que des paramètres de hauteur, de durée et de qualité vocale, nos travaux sont centrés sur la question de la modification performative du rythme de la voix. Une grande partie de ce travail de thèse a été consacrée au développement de Vokinesis, un logiciel de modification performative de signaux de voix préenregistrés. Il a été développé selon ces objectifs: permettre le contrôle du rythme de la voix, avoir un système modulaire, utilisable en situation de concert ainsi que pour des applications de recherche. Son développement a nécessité une réflexion sur la nature du rythme vocal et sur la façon dont il doit être contrôlé. Il est alors apparu que l'unité rythmique inter-linguistique de base pour la production du rythme vocal est de l'ordre de la syllabe, mais que les règles de syllabification sont trop variables d'un langage à l'autre pour permettre de définir un motif rythmique inter-linguistique invariant. Nous avons alors pu montrer que le séquencement précis et expressif du rythme vocal nécessite le contrôle de deux phases, qui assemblées forment un groupe rythmique: le *noyau* et la *liaison* rythmiques. Nous avons mis en place plusieurs méthodes de contrôle rythmique que nous avons testées avec différentes interfaces de contrôle. Une évaluation objective a permis de valider l'une de nos méthodes du point de vue de la précision du contrôle rythmique. De nouvelles stratégies de contrôle de la hauteur et de paramètres de qualité vocale avec une tablette graphique ont été mises en place. Une réflexion sur la pertinence de cette interface au regard de l'essor des nouvelles interfaces musicales continues nous a laissé penser que la tablette est la mieux adaptée au contrôle expressif de l'intonation et de la mélodie monophonique, mais que les PMC (Polyphonic Multidimensional Controllers) sont mieux adaptés au contrôle polyphonique. Le développement de Vokinesis a également nécessité la mise en place de la méthode de traitement de signal VoPTiQ (Voice Pitch, Time and Quality modification), combinant une adaptation de l'algorithme RT-PSOLA et des techniques particulières de filtrage pour les modulations de qualité vocale. L'utilisation musicale de Vokinesis a été évaluée avec succès dans le cadre de représentations publiques du Chorus Digitalis, pour du chant de type variété ou musique contemporaine. L'utilisation dans un cadre de musique électronique a également été explorée par l'interfaçage du logiciel de création musicale Ableton Live. Les perspectives d'application sont multiples: études scientifiques (recherches en prosodie, en parole expressive, en neurosciences...), productions sonores et musicales, pédagogie des langues, thérapies vocales.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

**Title :** Vokinesis : an instrument for supra-segmental control of voice synthesis

**Keywords :** voice synthesis, human-computer interactions, sound and music computing, prosody, vocal signal processing

This work belongs to the field of performative control of voice synthesis, and more particularly of real-time modification of pre-recorded voice signals. In a context where such systems were only capable of modifying parameters such as pitch, duration and voice quality, our work focuses on the question of performative modification of voice rhythm. One significant part of this thesis has been devoted to the development of Vokinesis, a program for performative modification of pre-recorded voice. It has been developed under these goals: to allow for voice rhythm control, to obtain a modular system, usable in public performances situations as well as for research applications. To achieve this development, a reflexion about the nature of voice rhythm and how it should be controlled has been carried out. It appeared that the basic inter-linguistic rhythmic unit is syllable-sized, but that syllabification rules are too language-dependant to provide an invariant inter-linguistic rhythmic pattern. We showed that accurate and expressive sequencing of voice rhythm is performed by controlling the timing of two phases, which together form a rhythmic group: the rhythmic *nucleus* and the rhythmic *link*. We developed several rhythm control methods, tested with several control interfaces. An objective evaluation showed that one of our methods allows for very accurate control of rhythm. New strategies for voice pitch and quality control with a graphic tablet have been established. A reflexion about the relevance of graphic tablets for pitch control, regarding the rise of new continuous musical interfaces, has left us think that they best fit expressive monophonic intonation and melody control, but that PMC (Polyphonic Multidimensional controllers) are better for polyphonic control. The development of Vokinesis also required the implementation of the VoPTiQ (Voice Pitch, Time and Quality modification) signal processing method, which combines an adaptation of the RT-PSOLA algorithm and some specific filtering techniques for voice quality modulations. The use of Vokinesis as a musical instrument has been successfully evaluated in public representations of the Chorus Digitalis ensemble, for various singing styles (from pop to contemporary music). Its use for electronic music has also been exploredbyinterfacingthe digital audio workstation Ableton Live with Vokinesis. Application perspectives are diverse: scientific studies (research in prosody, expressive speech, neurosciences...), sound and music production, language learning and teaching, speech therapies.

**Université Paris-Saclay**
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France