Synthèse par règles de la voix chantée contrôlée par le geste et applications musicales

Lionel Feugère

THÈSE DE DOCTORAT DE
L'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité :
Acoustique

Présentée par :

Lionel FEUGÈRE

Pour obtenir le grade de
DOCTEUR DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Synthèse par règles de la voix chantée contrôlée par le geste et applications musicales

Soutenue le jeudi 26 septembre 2013

devant le jury composé de :

Laurent Girin	Professeur Grenoble-INP / GIPSA-lab	Rapporteur
Marcello M. Wanderley	Professeur MacGill University / CIRMMT	Rapporteur
Boris Doval	Maître de conférence UPMC / d'Alembert	Examinateur
Nathalie Henrich	Chargée de recherche CNRS / GIPSA-lab	Examinatrice
Jean-Luc Zarader	Professeur UPMC / ISIR	Examinateur
Christophe d'Alessandro	Directeur de Recherche CNRS / LIMSI	Directeur de thèse

Groupe Audio et Acoustique
LIMSI-CNRS
B.P. 133
91403 Orsay Cedex, France

ED SMAER
Campus Jussieu - BC 270
4 place Jussieu
75252 Paris cedex 05, France

Cette thèse est dédiée aux divinités suivantes :
Gnowê, maître de la tirolienne et de la clave ;
Kimchi, maître du légume fermenté ;
G-switch, maître de la patate douce.

Remerciements

Mes premiers remerciements s'adressent à Christophe d'Alessandro qui a accepté d'encadrer mon doctorat et a su diriger intelligemment mes recherches doctorales pendant ces 3 ans et 9 mois. Grâce à son aide, j'ai pu aboutir à un projet de recherche cohérent, articulant différentes facettes du contrôle gestuel de la synthèse de voix chantée que sont le traitement du signal audio, la facture d'instruments numériques et l'étude de leurs usages. J'ai apprécié l'orientation musicale qu'ont pris ces recherches, en partie grâce à l'ouverture d'esprit de Christophe et son intérêt pour les questions scientifiques appliquées à la musique. Ce ne fût pas simple, mais la pluridisciplinarité de l'approche a rendu le travail passionnant.

Merci aux rapporteurs de cette thèse, Laurent Girin et Marcello M. Wanderley, et aux examinateurs, Boris Doval et Nathalie Henrich, d'avoir accepté de faire partie du jury et pour leur lecture très attentive, leurs commentaires éclairés ainsi que leur regard critique.

Mon doctorat a commencé en exploitant quelques patchs Max *magiques* de Sylvain Le Beux, qui venait de soutenir sa thèse dans l'équipe. Son travail m'a permis de partir sur de solides bases et le décryptage de ses patchs a été un bon exercice pour apprendre à maîtriser le langage Max. Je tiens à le remercier pour tout cela. Neuf mois plus tard, son retour d'un an dans l'équipe a été fort intéressant pour moi.

Je tiens également à remercier Boris Doval qui avait travaillé avec Christophe et Sylvain sur la modélisation de la source glottique, et qui a continué à s'intéresser avec enthousiasme au projet depuis son nouveau laboratoire à l'Institut Jean le Rond d'Alembert. Je le remercie spécialement d'avoir mis *les mains dans le cambouis* pour m'aider à améliorer le moteur de synthèse vocale. J'ai aussi beaucoup apprécié son investissement en tant que *Pandit Boris Khan* pour avoir fait vivre avec talent cette voix de synthèse dans des styles d'Inde du Nord, et pour les quelques tentatives théâtrales communes pour expliquer notre travail au grand public.

Les prouesses d'Albert Rilliard en statistique ont été d'une grande importance pour chercher la signification statistique dans les gestes de nos musiciens-cobayes. Pendant la dernière année de mon doctorat, il m'a aussi été très agréable de travailler avec Olivier Perrotin (et de vagabonder avec lui dans le pays du Kimchi), qui a déjà et va continuer à faire évoluer les recherches avec brio. Merci aussi à Guillaume Mahenc, stagiaire de Master 1 qui a fait littéralement trembler la voix de synthèse.

Le développement des instruments de voix chantée s'est fait avec un constant aller-retour avec les musiciens du Chorus Digitalis, chorale de voix synthétiques contrôlées à l'aide de tablettes graphiques. Elle nous a aussi permis d'étudier l'usage de ces tout nouveaux instruments. Je tiens donc à remercier tous les participants, à savoir Annelies B., Boris D., Christophe D., Emmanuelle F., Hélène M., Marc E., Olivier P. et Sylvain L., pour leur faculté à chanter si bien avec leurs petites mains, leur disponibilité, les compositions et arrangements pour nos divers concerts.

Merci à Charles Gondres du Laboratoire de Mécanique et d'Acoustique de Marseille pour sa maîtrise dans la récupération des données des tablettes graphiques dans Max. Sa réactivité et son investissement dans le débogage et l'ajout de fonctionnalités m'ont beaucoup aidé à réaliser une partie de mon travail sur le plan technique.

Ces années ont été joyeusement entretenues par les collègues de l'équipe AAA et du LIMSI, qui en ont fait un agréable séjour sur les hauteurs du plateau d'Orsay. Sans se risquer à sortir de l'équipe, commençons tout d'abord par remercier les frères et soeurs de thèse : Nicolas S. le jeûneur, Marc R. le papa, David S. le surfeur, Gaëtan la LDP, Tifanie Tiphany, David Fumée-Toutounet-Tête Pensante, Paul le hackeur d'affiche, Marc E. le bassiste imperturbable, Olivier Kimchi-Klaus, David Clarinette-Coach de soutenance, Matthieu Président, Tran Sécurité-Sociale ; les compagnons de plus court passage : Maître Sylvain, Johan G-switch, DJ Evi, Vincent Trombone, Cédric Twix, Davidé Primodraft, Simon Pshh-Pshh-Pshh, Marc F, Guillaume le perturbateur, Aurore, Renaud Sabir-fretless, Nicolas A. multilab, Frédéric, Thanh ; les permanents : Christophe, Albert, Brian, Nathalie et Laurent. Dédicace à Driss qui est toujours au LIMSI pour ouvrir les portes les jours les plus improbables quand on oublie ses clés à Paris. Dédicace au LIMSI qui regorge de beaux spécimens dans la catégorie *musiciens* et qui m'ont bien stimulé ces années-là. Sans oublier le RER B, bien sûr, compagnon quotidien pour le meilleur et le pire.

En dehors des murs du laboratoire, mais toujours sur le campus d'Orsay, dédicace à Federico, ami et profesor de español à Paris Sud, *carrément et cerclement* atypique et plein d'humanité, et qui nous a quitté brusquement.

Merci beaucoup aux relecteurs de ce document qui ont fait disparaître un maximum de fautes d'or t'agrafes, à savoir Maman, Smaa, Séb, David-Clarinette, YéYé et le jury de thèse ... Il y avait du boulot !

Enfin, un grand merci à mes parents pour m'avoir permis d'en arriver là, aux 2 frèrosbiquets, mamie Poulet, Asmaa Ananas (KBBG) et tous mes proches qui sauront se reconnaître. Une pensée à papi Jacques et mamie Hélène.

Table des matières

Notations et expressions

Notation ou expression	**Signification**
A_i avec $i \in 1, 2, 3, 4, 5$	Amplitude du filtre formantique i
α_m	Coefficient d'asymétrie de l'ODGD
B_i avec $i \in \{1, 2, 3, 4, 5\}$	Bande-passante du filtre formantique i
Consonne tenue	Phase médiane de l'articulation d'une séquence VCV
CV	Consonne-Voyelle
F_e	Fréquence d'échantillonnage
F_i avec $i \in \{1, 2, 3, 4, 5\}$	Fréquence centrale du filtre formantique i
F_0	Fréquence fondamentale de vibration des plis vocaux
ODG	Onde de Débit Glottique
ODGD	Onde de débit glottique dérivée
O_q	Quotient ouvert de l'ODGD
Phase d'articulation	Position des articulateurs entre deux phonèmes
Plis vocaux	Cordes vocales
T_0	Période fondamentale de vibration des plis vocaux
VE	Paramètre d'effort vocal
VC	Voyelle-Consonne

Liste des fichiers audio-visuels

Les fichiers audios et vidéos listés ci-desous sont disponibles à l'adresse suivante :
http://groupeaa.limsi.fr/membres:feugere:these
ou sur simple demande au groupe *Audio et Acoustique* du LIMSI-CNRS.

Chapitre 2 : Cantor Digitalis

1. Analyse synthèse d'un /a/ (partie 2.3.4)
   - Voix naturelle
   - Voix synthétique (sans perturbation automatique)
2. Atténuation des amplitudes des filtres formantiques avec F0 (partie 2.4.1)
   - Glissando sans atténuation
   - Glissando avec atténuation
3. Dépendance de F1 avec l'effort vocal (partie 2.4.2)
   - Crescendo sans dépendance
   - Crescendo avec dépendnace
4. Dépendance des fréquences F1,F2,F3,F4,F5 des filtres formantiques avec F0 (partie 2.4.3)
   - Glissando sans dépendance
   - Glissando avec dépendance
5. Types de voix 1 (partie 2.5.1)
   - Soprano
   - Alto
   - Ténor
   - Basse
   - Bébé
   - Enfant
6. Types de voix 2 (partie 2.5.2)
   - Soprano bruité
   - Alto bruité
   - Fauve
   - Monstre
7. Types de voix 3 (partie 2.7.3)
   - Chant diphonique

Chapitre 3 : Digitartic

8. Quelques syllabes (partie 3.3.5)
   – apa,ata,aka,ava,aza,a3a,awa,aya,aja,ama,ana/
9. Degré et vitesse d'articulation (partie 3.3.6)
   – Différents degrés d'articulation avec /aja/
   – Vitesse de contrôle articulatoire avec /awa/
10. Démonstration générale
   – Expressivité de l'articulation (partie 3.5)

Annexe D : L'ensemble Chorus Digitalis

11. Classique européen
   – Vidéo du choral « Alta Trinita Beata » de Bach @PS3 workshop, Vancouver (partie D.3.2)
12. Raga d'Inde du Nord (partie D.3.3)
   – Journée Sciences et Musique 2012, Rennes (Audio)
   – Journées Art Science 2012, Printemps de la Culture, Orsay (Vidéo)
13. Contemporain
   – Vidéo de la polyphonie contemporaine «Valse » (Bruno Lecossois) @Journées Art Science 2012, Printemps de la Culture, Orsay (partie D.3.4)
14. Autre
   – Vidéo intégrale du concert du Printemps de la Culture 2012, moins les morceaux « Ocean » et « North Star » n'ayant pas l'autorisation des auteurs.

Introduction

Contexte et problématique

Cette thèse se situe dans le cadre du contrôle gestuel de la synthèse vocale, et plus particulièrement celui de l'articulation dans des contextes musicaux. La problématique associée est la modélisation des mouvements articulatoires de la production de syllabes et de leur contrôle. Ainsi, il s'agit d'étudier l'externalisation de mouvements internes de l'appareil vocal en gestes manuels. Il faut alors déterminer quels sont les paramètres de haut-niveau qui permettent de le contrôler, en trouvant un équilibre entre simplicité et richesse de contrôle. En d'autres termes, il faut examiner où peut se situer la frontière entre modèle de contrôle et modèle de production.

La difficulté réside aussi dans la différence intrinsèque entre les gestes vocaux et manuels. Les gestes vocaux sont issus du déplacement et de l'interaction de nombreux organes et de l'air issu des poumons, alors que les gestes manuels proviennent des deux bras et des articulations des poignets jusqu'aux dernières phalanges. Cependant, tous les deux ont en commun certains types d'actions, comme le fait de pouvoir viser des cibles très précises, serrer, ou encore obstruer. On a donc un changement de paradigme de contrôle entre la production vocale naturelle et des instruments de voix de synthèse qu'il convient de ne pas choisir au hasard. Cette externalisation des gestes articulatoires doit permettre l'analyse par la synthèse de la production vocale.

La voix humaine, qu'elle soit parlée ou chantée, est le résultat d'une utilisation experte de l'appareil vocal. Cet instrument qu'est la voix présente la particularité d'être contrôlé par des organes internes et d'être maîtrisé, en ce qui concerne la parole, par la quasi totalité des Hommes. La compréhension de son fonctionnement a été entreprise notamment par la construction de machines parlantes, au moins à partir du 18$^{\text{ème}}$ siècle.

La compréhension d'un phénomène s'établit quand on parvient à prévoir ses comportements. Pour y parvenir, il faut modéliser le phénomène pour comprendre l'interaction de ses composants et les réponses du phénomène à divers stimulus. La synthèse vocale permet ainsi de simuler le fonctionnement de l'appareil vocal. Différentes techniques de modélisation existent, et présentent chacune une facette de la production vocale.

Tout le monde a une certaine connaissance de la production vocale, l'utilisant de façon régulière et de manière experte. Mais pour émettre un son de parole, seul le sens de la parole est pensé. Inconsciemment, notre cerveau actionne les muscles nécessaires au résultat sonore désiré. Ces enchaînements de gestes internes sont la conséquence d'un apprentissage empirique pendant notre petite enfance.

C'est pour mieux comprendre le fonctionnement de la voix que des scientifiques ont cherché à modéliser la production vocale, par des disciplines diverses comme la phonétique, la linguistique, l'acoustique ou le traitement du signal. Dans le cas de modèles mathématiques, la synthèse a permis de valider certains modèles, attestant de leur bonne prédiction de certains

aspects de la production vocale.

Il est intéressant de noter que tout au long de l'histoire de la synthèse vocale, le contrôle temps-réel est souvent présent comme cible à atteindre. On peut avancer deux raisons sans vouloir être exhaustif. La première est la volonté de pouvoir parler avec la voix de quelqu'un d'autre. En effet, la voix est par définition personnelle, elle nous identifie, et on veut pouvoir briser cette barrière par la synthèse. De plus, son caractère interactif nous pousse à pouvoir faire de même en synthèse. Une deuxième raison est la complexité de sa modélisation. Retirer une partie du phénomène à modéliser et le remplacer par un contrôle manuel permet de ne se préoccuper que de la production vocale et non de l'automatisation de son contrôle. Cependant, on verra dans cette thèse que la séparation du modèle de production et de contrôle n'est pas trivial, puisqu'il forme à l'origine un tout. La façon de relier les gestes manuels et le modèle de production amène à la conception d'un autre modèle à proprement parler.

Ainsi, la machine mécanique de Kempelen au 18$^{\text{ème}}$ siècle [DT50] pouvait se contrôler manuellement. Le VODER [DRW39], avec l'apparition de l'électricité, était destiné au jeu à l'aide d'un clavier. Puis, avec notamment la musique électronique, les modèles de production vocale ou ses dérivés ont été et sont toujours largement utilisés, de la pédale de guitare *Wah-Wah*, qui donne l'impression d'une articulation vocale, au Vocoder, permettant d'utiliser la voix du musicien pour filtrer en temps réel un autre instrument. A mesure que la technologie offrait de nouvelles interfaces, des modèles plus sophistiqués ont fait leur apparition, comme les gants haptiques ou les tablettes graphiques. Des conférences scientifiques destinées à ces nouvelles problématiques ont vu le jour, comme *NIME* (New Interfaces for Musical Expression) depuis 2001 ou encore plus spécifiquement le *Workshop P3S* (Performative Speech and Singing Synthesis) en 2011.

L'aspect musical fait intervenir des contraintes particulières sur les instruments de synthèse vocale. En particulier, pour jouer à plusieurs sur le même tempo, les attaques des syllabes doivent pouvoir être produites sans latence perceptible. Or la position de l'attaque de certaines syllabes, comme /la/, ne coïncide pas avec le début de la syllabe, mais avec la voyelle /a/. Cela implique que pour contrôler la position temporelle d'une telle syllabe sur un tempo extérieur, la phase d'articulation doit être continument contrôlée afin de synchroniser la voyelle avec la position temporelle désirée. D'autre part, l'aspect expressif de l'articulation est prépondérant pour exprimer des nuances musicales. Un maximum de paramètres permettant une richesse articulatoire doit être disponible tout en gardant un jeu possible avec le minimum d'entraînement. Le jeu musical a des formes diverses concernant l'utilisation des phonèmes. On se limite ici à la synthèse de syllabes dépourvues de sens, telles des voyelles et des onomatopées. Tous les phonèmes d'une langue donnée ne sont alors plus nécessaires. On peut se limiter à un sous-ensemble, diminuant les combinaisons articulatoires et simplifiant d'autant l'interface de contrôle de l'instrument.

Il est intéressant par la suite de comparer l'utilisation de ces instruments vocaux à celle de la voix naturelle. D'une part, on peut les évaluer de façon formelle sur des tâches particulières, par des expériences comparant réalisation chironomique (i.e. par des gestes manuels) et réalisation par la voix naturelle. D'autre part, un instrument de musique a pour but d'être pratiqué en situation musicale, et non dans des conditions expérimentales. Utiliser un instrument en concert est l'aboutissement de ce travail, puisque l'instrument doit alors posséder toutes les qualités requises. L'inconvénient est que cela ne permet pas une rigoureuse démonstration scientifique.

Plan du manuscrit

Le premier chapitre a pour but d'inscrire notre recherche dans le cadre de connaissances existantes et de travaux similaires. Notre objet de recherche, l'appareil vocal, est approché selon différents points de vue à savoir phonétique, phonémique et musical. La synthèse vocale est abordée à travers ses différentes méthodes et nous décrirons quelques instruments de synthèse vocale. Ensuite, est traitée brièvement la nomenclature des gestes instrumentaux. Nous expliciterons à la fin de ce chapitre les travaux de recherche passés du LIMSI grâce auxquels cette thèse a pu débuter, à savoir le contrôle gestuel de la prosodie et de la qualité vocale.

Puis viendront deux chapitres regroupés dans une grande partie intitulée « Instruments de synthèse vocale » qui présente le développement des instruments numériques de cette thèse.

Le chapitre 2 présentera le *Cantor Digitalis*, un instrument de voyelles chantées basé sur la synthèse par formants. Nous exposerons d'abord les réglages apportés au modèle de source glottique RT-CALM puis de la modélisation des résonances du conduit vocal. Des interactions sources-filtres ont été ajoutées afin de prendre en compte les dépendances entre les résonances du conduit vocal, l'effort vocal et la fréquence fondamentale. Les voix sont personnalisables pour reproduire la diversité des timbres de voix, notamment par la grandeur de leur conduit vocal, par la qualité vocale moyenne, afin d'aboutir à différents types vocaux existant tels que basse, ténor, alto, soprano ou à de nouvelles voix. Enfin, nous traiterons de l'interface réalisée, basée sur une tablette graphique, et de la correspondance des paramètres de l'interface au modèle de production.

Le deuxième instrument de voix numérique réalisé, le *Digitartic*, est décrit dans le chapitre 3. Il est la continuation du *Cantor Digitalis* et permet le contrôle de l'articulation de manière précise pour pouvoir jouer en rythme musicalement. Un intérêt spécifique sera porté à la construction du modèle de contrôle, bien plus complexe que pour le contrôle des voyelles chantées, de part le nombre de paramètres à contrôler. Deux tablettes graphiques, augmentées d'un calque superposé et munies de stylets, sont utilisées pour le jeu sur l'articulation et l'intonation musicale.

Les trois chapitres suivants sont regroupés dans la deuxième grande partie de ce manuscrit, intitulée « Jeux individuels et collectifs : analyse et évaluation ». Elle permet principalement l'évaluation de notre travail, sur des tâches musicales d'imitation et d'analyse en jeu libre.

Dans le chapitre 4, seront décrits les gestes des deux instruments *Cantor Digitalis* et *Digitartic* suivant différentes nomenclatures. On comparera et analysera ensuite les stratégies des musiciens pour reproduire gestuellement certaines tâches musicales avec le *Cantor Digitalis*, comme le vibrato ou le portamento.

Dans le chapitre 5, nous présenterons une expérience visant à évaluer la tablette graphique augmentée, comme interface de contrôle de la fréquence fondamentale pour des tâches d'imitation d'intervalles et de mélodies. Nous l'étudierons intrinsèquement en terme de justesse et de précision et en la comparant aux performances de la voix naturelle pour des tâches similaires. L'influence du retour audio du synthétiseur sur la qualité de l'imitation sera examinée en demandant aux participants de l'expérience de reproduire les intervalles et les mélodies à l'aide d'une troisième modalité, consistant en la tablette graphique augmentée seule sans synthétiseur vocal.

Une évaluation moins complète que la précédente sera décrite dans le chapitre 5, visant cette fois-ci la justesse relative (inter-musiciens) dans un ensemble de 4 *Cantor Digitalis*.

Nous finirons ce manuscrit en résumant les principaux apports de cette thèse et en développant quelques perspectives futures de recherche.

Une série d'annexes donne d'avantage d'information sur les points techniques du manuscrit ou des applications de notre travail. Les points techniques sont les équations mathématiques régissant le modèle de source glottique CALM, l'interface logicielle des instruments numériques, et quelques idées pour diminuer la latence des instruments fonctionnant sous Max/MSP. Deux applications sont présentées. La première est le *Chorus Digitalis*, notre ensemble de voix de synthèse avec lequel nous jouons régulièrement en concert. La deuxième est une application pédagogique permettant de construire et déconstruire des voix synthétiques, pour l'enseignement de la phonétique ou l'acoustique de l'appareil vocal. Enfin, la liste des publications associées à cette thèse est donnée dans la dernière annexe.

Chapitre 1

La synthèse vocale pour le jeu musical : cadre d'étude et état de l'art

Sommaire

1.1 Introduction

Nous pouvons échanger de l'information avec notre environnement à travers différents supports mécaniques ou électromagnétiques, et cela grâce à la vue, l'ouïe, le toucher, l'odorat ou le goût. Un signal sonore peut être émis à l'aide d'un objet extérieur, ou directement par notre corps. Le meilleur exemple est l'appareil vocal qui permet, à l'aide de ses organes, de produire des sons d'une grande richesse, destinés principalement à la communication.

Nous commencerons par décrire l'appareil vocal comme générateur sonore, puis comme producteur de parole et enfin comme instrument de musique. Nous traiterons ensuite de l'état de l'art en synthèse vocale, en accentuant la présentation sur les instruments de synthèse de voix chantée ou *performative*. Enfin, les travaux du LIMSI à partir desquels cette thèse a pu débuter seront exposés.

1.2 L'appareil vocal, émetteur sonore

L'appareil vocal est un ensemble d'organe permettant l'émission de sons, et qu'on peut décrire suivant plusieurs approches complémentaires : anatomique (quelles sont ces organes ?), comme un système composé d'une source sonore et d'un filtre, ou en analysant le signal acoustique à sa sortie.

1.2.1 Description anatomique de l'appareil vocal

Les organes de l'appareil vocal ne sont pas destinés seulement à la production de la voix. Ils accomplissent d'autres fonctions dont les principales sont la respiration (poumons), l'alimentation (conduit vocal), l'olfaction (cavité nasale), la dégustation (langue), et la communication faciale (lèvres, mâchoires).

En ce qui concerne la production de la voix, ces organes peuvent être présentés suivant trois composantes comme il suit [CT89].

a) La soufflerie

A l'aide du diaphragme, la cage thoracique se déforme et permet de faire entrer ou sortir de l'air des poumons. L'air se propage des poumons par la trachée jusqu'aux orifices respiratoires que sont les lèvres et les narines, en passant par le conduit vocal, dans un sens ou dans l'autre, selon que la personne inspire ou expire. En général, l'appareil vocal est utilisé lors de la phase d'expiration.

b) Le larynx

Le larynx est un organe situé au niveau du cou. Son squelette est constitué de cartilages mis en mouvement à l'aide de muscles et reliés entre eux par des ligaments.

Les *plis vocaux*, qui comprennent ces ligaments, une muqueuse et les muscles thyro-aryténoïdiens, permettent d'obstruer complètement ou partiellement le flux d'air provenant des poumons.

On appelle *glotte* l'espace formé entre les plis vocaux.

c) Le conduit vocal

Le conduit vocal est composé de la langue, ensemble de 17 muscles lui permettant de se mouvoir avec précision dans la bouche, de la mâchoire pouvant se déplacer suivant trois plans, des lèvres, des dents, et du voile du palais ou luette qui, selon son degré d'ouverture, permet à l'air provenant des poumons d'entrer dans les fosses nasales.

Les organes du conduit vocal et le larynx sont nommés *articulateurs*, car ils permettent d'articuler les sons vocaux.

1.2.2 Description source-filtre de l'appareil vocal

Selon la description donnée par Fant [Fan60], l'appareil vocal peut être décrit par un système source/filtre. La source glottique (i.e. le son formé au niveau de la glotte) transforme la pression constante des poumons en une série d'impulsions par l'alternance des ouvertures et fermetures des plis vocaux, qu'on appelle *onde de débit glottique* (ODG). Elle est vue comme la réponse d'un filtre de réponse impulsionnelle $g[n]$ (où n représente une variable temporelle discrète) excité par un train d'impulsions de période T_0. Le conduit vocal, composé des articulateurs en mouvement, est quant à lui modélisé par un filtrage variable dans le temps $v[n]$ qu'on suppose constant à l'échelle de la période fondamentale. Enfin, le rayonnement de l'onde de débit par les lèvres et les narines est vu comme un filtrage dérivateur $l[n]$. Ainsi, le signal $s[n]$ obtenu à l'extérieur du conduit vocal est la convolution du signal de source $e[n]$ et des réponses impulsionnelles du conduit vocal $v[n]$ et du rayonnement aux lèvres $l[n]$, dans le cas d'un son voisé non bruité :

$$s[n] = e[n] * v[n] * l[n] \tag{1.1}$$

à l'instant temporel n échantillonné. La réponse impulsionnelle de la source $e[n]$ peut s'exprimer comme la convolution de la réponse de la source glottique à une excitation périodique de Dirac :

$$\begin{aligned} e[n] &= \sum_{p\in\mathbb{N}} \delta[n - pT_0] * g[n] \\ &= \sum_{k\in\mathbb{Z}} \sum_{p\in\mathbb{Z}} \delta[n - pT_0 - k] \cdot g[k] \\ &= \sum_{p\in\mathbb{N}} g[n - pT_0] \end{aligned} \tag{1.2}$$

Dans le domaine fréquentiel, on obtient alors un produit des réponses en fréquence de la source, du filtre du conduit vocal et du rayonnement aux lèvres :

$$S[w] = E[w] \cdot V[w] \cdot L[w] \tag{1.3}$$

à la pulsation w.

Une description schématique du fonctionnement du modèle source-filtre est donnée à la figure 1.1, où le train d'impulsions périodiques peut être remplacé ou mis en parallèle par une excitation de type bruit, afin de modéliser le bruit de souffle vocal.

Ce modèle suppose une indépendance entre la source glottique et le conduit vocal, ce qui n'est vrai qu'en première approximation. Le couplage provoque notamment des changements dans les fréquences centrales et bandes-passantes des résonances du conduit vocal, au moins dans le cas de voyelles [Car81].

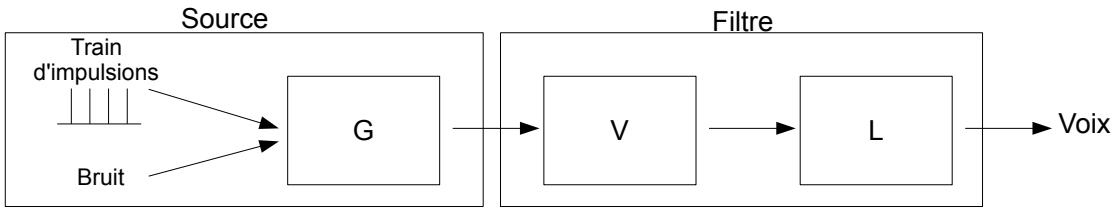


FIGURE 1.1 – *Représentation schématique du fonctionnement du modèle source-filtre*

1.2.3 Description acoustique de l'appareil vocal

A partir de l'analyse acoustique de la voix, on peut en déduire les valeurs types et l'étendue d'un certain nombre de paramètres temporels et spectraux caractérisant le fonctionnement de la source glottique et/ou du conduit vocal. Ils sont définis en Annexe A. On donne ici quelques exemples de ce qui peut être analysé acoustiquement selon d'Alessandro [d'A06].

a) L'effort vocal

L'action de parler plus fort, l'effort vocal ou force vocale, n'engendre pas seulement une augmentation du niveau de pression sonore. Il en résulte également la réduction de la pente spectrale du signal acoustique de la voix dans les hautes fréquences.

b) La dimension tendue / relâchée

Plus la pression de la partie postérieure des plis vocaux sera élevée et plus la voix sera qualifiée de tendue ou pressée. Inversement, lorsque les plis vocaux sont relâchés, correspondant à une vibration non contrainte des cordes vocales, alors la qualité de voix est dite relâchée, ou détendue. Pour une voix totalement relâchée, la forme de l'onde de débit glottique est quasiment sinusoïdale. Une voix relâchée verra la fréquence de son maximum spectral s'abaisser et au contraire une voix tendue la verra s'élever [DdH06].

c) Les apériodicités

La partie bruitée du signal vocal a essentiellement deux origines : le bruit structurel et le bruit additif. Le bruit structurel correspond aux variations aléatoires de période à période dans la source glottique : le système biomécanique formé par le larynx n'est pas une source parfaitement périodique. Les deux aspects du bruit structurel sont le jitter et le shimmer, respectivement les variations d'une période à l'autre de la fréquence fondamentale (grandeur sans unité, définie en pourcentage de la fréquence fondamentale) et de l'amplitude (pourcentage de déviation de l'amplitude d'une période fondamentale à l'autre). Une voix avec un jitter et un shimmer prononcés sera qualifiée de rugueuse ou âpre [Hes59].

Le bruit additif est lié à la friction créée au niveau des cordes vocales lors du passage de l'air provenant des poumons. Ce bruit additif est particulièrement présent pour deux principaux types de qualité vocale : la voix chuchotée, lorsque les cordes vocales sont ouvertes (pas de vibration) et la voix soufflée, pour laquelle les cordes vocales vibrent mais ne se ferment pas complètement.

d) Le phonétogramme

Le phonétogramme représente le diagramme formé par la fréquence fondamentale et le niveau de pression acoustique que peut réaliser un locuteur ou un chanteur donné. Il exprime les limites non atteignables ou au-delà desquelles la production vocale s'interrompt brusquement [HdDC05].

e) La notion de registre ou mécanisme laryngé

La source glottique peut parcourir une grande étendue de fréquences en utilisant plusieurs mécanismes de vibration appelés mécanismes ou registres laryngés. On observe quatre mécanismes de vibration des plis vocaux : M0 (friture ou voix craquée), M1 (modal ou voix de poitrine), M2 (voix de fausset ou de tête), M3 (voix de sifflet) [RHC09].

f) Les formants

Un formant à la fréquence F_i (i indexant le formant) est la manifestation d'une ou plusieurs résonances du conduit vocal aux fréquences R_k suffisamment proches (k indexant la résonance). En analysant le spectre de sons voisés, nous pouvons en déduire une estimation des fréquence, amplitude et bande-passante de ce formant. Typiquement, on trouve environ 5-6 formants dans l'intervalle fréquentielle $[0 - 5000]$ Hz.

1.3 L'appareil vocal, producteur de phonèmes

La capacité à émettre des sons riches dans leur diversité a été nécessaire pour pouvoir développer un système codé permettant de véhiculer une information complexe. On se place dans cette section dans une description moins large de l'appareil vocal, centré sur la production d'unités sonores et leur articulation.

1.3.1 Cadre phonologique et phonétique

La variété de sons réalisables avec notre appareil vocal est infinie. L'étude de leur production est la phonétique. La phonologie, quant à elle, étudie les sons qui forment le sens dans une langue donnée. Les sons utilisés par la langue sont en nombre fini [Tro39].

On parle ainsi d'opposition phonologique pour toute opposition phonique qui peut dans une langue donnée différencier des significations. Un phonème est défini comme la plus petite unité qui permet une opposition phonologique. Par exemple en français, l'opposition [p]-[b] est phonologiquement distinctive (les mots *pas* et *bas* ont des sens différents) alors qu'il n'existe pas d'opposition phonologique entre le /r/ fricatif [ʁ] (« parisien ») et le /r/ roulé [r], car ne créant pas de différence de sens, par exemple entre rat prononcé [ʁa] et rat prononcé [ra]. On appelle alors [ʁ] et [r] des réalisations du même phonème /r/, appelées variantes [Tro39], ou allophones [Lav94]. L'utilisation d'un allophone ou d'un autre dépend généralement de l'accent du locuteur. On note qu'on utilise les notations entre crochets [...] pour parler des sons issus de l'alphabet phonétique international (API) et de barres /.../ pour citer des phonèmes. L'API est donné à la figure 1.2.

Enfin, alors que deux mêmes variantes d'un phonème seront identiques du point de vue phonétique (et donc phonologique), elles ne le seront très probablement pas du point de vue acoustique, d'autant plus si ces deux variantes sont produites par des personnes différentes (impliquant au moins une anatomie différente, sauf peut être dans le cas de vrais jumeaux) [Lav94].

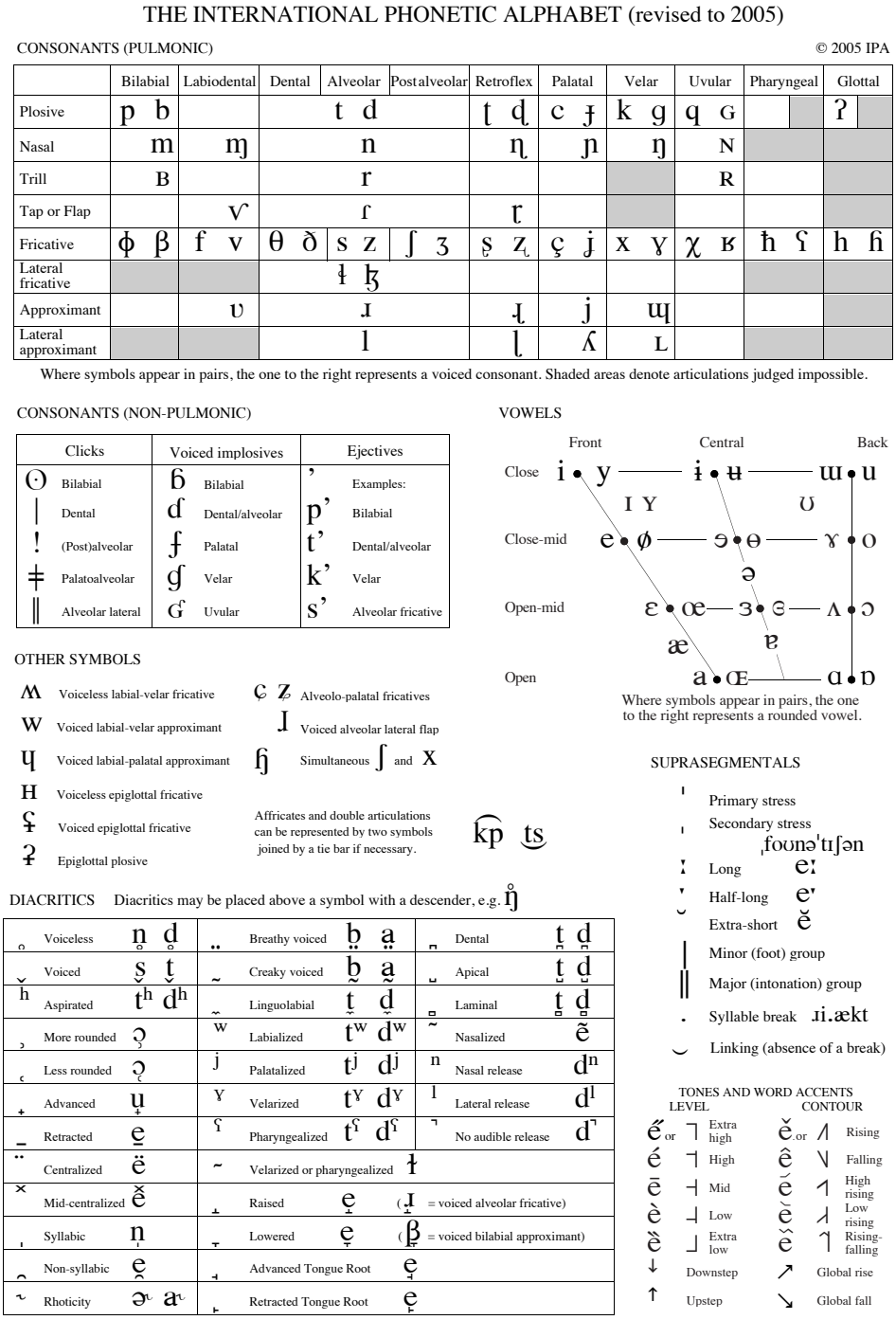


FIGURE 1.2 – *L'alphabet phonétique international (IPA) mis à jour en 2005 [IPA05].*

1.3.2 Articulation, voyelles et consonnes

L'articulation est le déplacement des organes articulateurs pour passer de la configuration d'un phonème à un autre. Le mécanisme d'articulation a une structure caractérisée par des contraintes neurologiques, anatomiques et mécaniques.

La coarticulation est quant à elle l'influence d'une configuration articulatoire sur les suivantes ou les précédentes. Daniloff et Hammarberg [DR73] définissent la coarticulation comme « l'influence d'un segment de parole sur un autre ou encore l'influence d'un contexte phonétique sur un segment de parole donnée ».

Parmi les phonèmes, les voyelles sont caractérisées par la vibration des plis vocaux et une configuration des articulateurs relativement stable pendant leur production. Les voyelles du français sont décrites suivant l'ouverture de la mâchoire (on parle de voyelle ouverte / mi-ouverte / fermée), la position de la langue suivant l'axe antéro-postérieur (voyelle antérieure / centrale / postérieure), l'utilisation du conduit nasal (voyelle nasale / orale), et la forme des lèvres (voyelle arrondie / non arrondie). D'un point de vue acoustique, les voyelles orales sont bien distribuées suivant les fréquences des deux premiers formants, chacun des axes correspondant approximativement et respectivement à l'ouverture de la mâchoire et à la position de la langue.

On distingue les consonnes des voyelles par la présence d'une ou plusieurs constrictions totales (mais temporaires) ou partielles du flux d'air par les organes articulateurs. On peut les catégoriser suivant le lieu d'articulation déterminé par la localisation de la constriction principale, par le mode d'articulation et la présence ou non de voisement.

Les lieux d'articulation s'étendent des lèvres au larynx (figure 1.3). Les obstructions se produisent entre deux articulateurs comme les deux lèvres ou une certaine partie du palais avec une autre de la langue. On parle de double articulation quand les deux constrictions majeures sont de degré équivalent et d'articulation secondaire quand les deux constrictions sont de degré distinct [Lav94]. Par exemple, le son [w] de l'onomatopée *wouah* peut être associé à une double articulation labiale et vélaire. La figure 1.4 donne la position de la langue pour les lieux d'articulation alvéolaire et uvulaire.

Quant aux modes d'articulation, différents traits phonétiques permettent de les catégoriser. L'un d'entre eux est le degré d'obstruction du lieu d'articulation, permettant de dégager trois grandes catégories [Lav94] :

- Si l'obstruction dans le conduit oral est totale, on parle d'occlusives telles que [p,b,-t,d,k,g,m,n]. Une occlusive est suivi d'un relâchement de cette obstruction créant une explosion, avant la transition vers le phonème suivant.
- Si l'obstruction est partielle mais suffisante pour produire une composante apériodique forte, on parle de fricatives telles que [f,v,s,z,ʃ,ʒ,ʁ]. Contrairement aux occlusives, elles peuvent être stables dans le temps, le flux d'air n'étant pas interrompu.
- Si l'obstruction est assez faible pour ne pas produire de forte composante apériodique, on parle de sonnantes telles que les approximantes centrales (ou semi-voyelles) [ɥ,w,j] ou les approximantes latérales (ou liquides) telle que [l]. Une représentation dans le plan sagittal de [l] est proche de l'occlusive alvéolaire [t]. La différence réside dans la forme de la langue, qui laisse passer l'air de part et d'autre de son centre, en contact avec l'alvéole, d'où sa qualification de latérale.

Pour que les lieux et modes d'articulation permettent de distinguer toutes les consonnes du français, il nous faut rajouter les modes d'articulation suivants : oral / nasal qui fait par exemple la distinction entre le [m] (nasal) et le [b] (oral) ; sourd / voisé (phonation) qui par exemple, distinguent [p] et [b].

Il est phonétiquement possible d'avoir une continuité entre certains lieux d'articulation

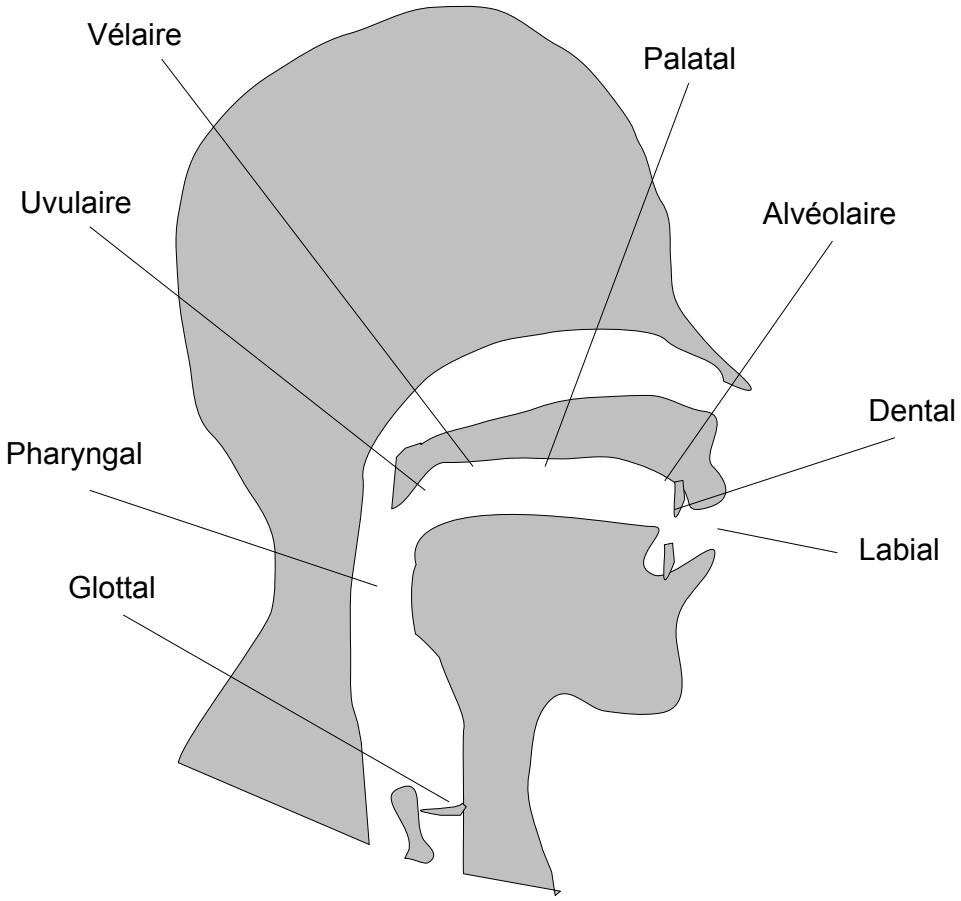
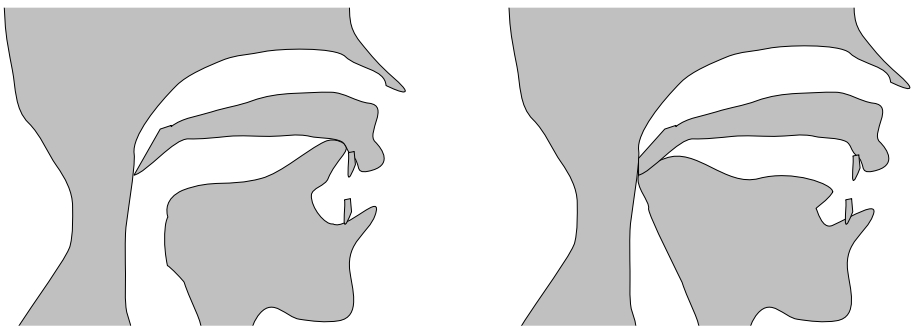


FIGURE 1.3 – *Les lieux d'articulation*



(a) *Occlusive alvéolaire* (b) *Occlusive uvulaire*

FIGURE 1.4 – *Plan sagittal du conduit vocal pour deux occlusives orales*

(déplacement de la langue sur le palais) ou entre certains modes d'articulation (degré d'ouverture de la cavité nasale). Cependant, c'est là où la différence entre phonétique et phonologie intervient. Notre perception étant catégorielle, on aura tendance à percevoir le trait distinguant une consonne d'une autre comme soit présent, soit non présent, mais non d'une manière continue même si c'est phonétiquement le cas. On placera alors la consonne dans une des catégories phonologiques perçues.

1.3.3 Organisation temporelle de l'articulation

Laver identifie 3 phases de l'articulation d'une consonne [Lav94] (p. 133-134) :
- la phase de déclenchement (*onset*) où les articulateurs se déplacent vers la position de constriction maximale
- la phase médiane correspondant au maximum de constriction
- la phase de fin (*offset*), où les articulateurs s'éloignent les uns des autres

Ainsi, suivant le type de consonne, on observera lors de la phase médiane un arrêt du flux d'air pour les occlusives, un écoulement d'air turbulent pour les fricatives, ou un écoulement d'air relativement libre pour les approximantes.

La durée de chacune de ces phases va dépendre de la contrainte articulatoire (durée minimale nécessaire pour passer d'une configuration à une autre), et de contrainte de sens pour faire des distinctions phonologiques ou expressives.

En anglais, Laver indique que la vitesse moyenne de l'articulation chez des locuteurs de langue maternelle anglaise est de 4-6 syllabes par seconde [Lav94].

1.4 L'appareil vocal, instrument de musique

Alors que la musique instrumentale peut ne pas être très développée dans toutes les cultures (Aborigènes d'Australie, Veddas du Sri Lanka), la musique vocale est utilisée partout dans le monde [Léo04]. Il sera question ici de décrire brièvement certaines utilisations de la voix dans un contexte musical à travers le monde.

1.4.1 Quelques techniques musicales à travers le monde

L'appareil vocal peut se suffire à lui-même pour produire de la musique, mais il peut également se coupler à des éléments extérieurs, ou être utilisé comme moyen ou objet d'imitation. Léothaud [Léo04] distingue 6 types de relation entre voix et instruments de musique non vocaux :
- la *voix déguisée*, par exemple en atténuant les hautes-fréquences dans le cas du port d'un masque.
- la *voix mélangée à l'instrument de musique*, qui peut être un résonateur (par exemple la bouche de la personne en face de soi dans les *katajjaq* ou un tronc creusé dans le *didgeridoo* des aborigènes d'Australie) ou un excitateur (par exemple le chant dans une flûte).
- la *voix associée à l'instrument* pour créer une voix composite (alternance rapide de la voix et du sifflet hindewou chez les Pygmés Aka et Mbenzele).
- l'*imitation vocale des instruments de musique*, pour les remplacer, pour se rapprocher de l'instrument accompagnant ou pour favoriser l'apprentissage d'un instrument.
- les *instruments parleurs*. Par exemple, dans les pays d'Afrique où la langue possède des tons, il est possible de reconstituer des phrases avec des tambours à hauteurs mélo-

diques différentes, pour échanger entre villages éloignés ou glisser des messages (souvent moqueurs pour amuser l'audience) au milieu d'un rythme de danse.
– les imitations des animaux, pour des raisons ludiques, signalétiques (e.g. chasse), ou magiques.

Concernant l'usage de l'appareil vocal seul, Léothaud retrace les différentes techniques vocales utilisées à des fins musicales et répertoriées par les ethnomusicologues [Léo04].

Parmi les appels, cris et clameurs, on peut citer les Kecak à Bali et Java formés d'onomatopées inspirées des cris du singe et les youyous d'Afrique du nord.

Les mécanismes laryngés peuvent être adaptés à la hauteur mélodique utilisée, liés à des codes sociaux (en inde, le raga interdit la voix de tête pour les hommes), ou peuvent être utilisés pour des effets particuliers comme les yodels tyroliens suisses ou les yodels des pygmés.

Enfin, on peut citer le chant diphonique, qui consiste à produire un son à spectre riche en harmoniques auquel on associe une résonance par la position de la langue, qui va créer la deuxième note perçue.

1.4.2 Gestes vocaux et gestes percussifs

La voix est utilisée pour imiter certains instruments de musique pour des fonctions de remplacement, ludique ou didactique. Dans cette dernière catégorie, répandue dans la musique savante asiatique, c'est la reproduction d'instruments de percussion qui est la plus développée [Léo04].

C'est en Inde (ou plus généralement dans le monde indien) où cette imitation des percussions est la plus riche. Elle est utilisée à la base comme outil pour l'enseignement de la plupart des percussions, comme les tablas, le pakhawaj, le mridangam, ou encore le ghattam. Chacune des frappes (ou association de deux frappes jouées au même instant) d'une percussion donnée est associée à une ou plusieurs onomatopées ou *bol*. Pour une même frappe et un même instrument, le choix de l'onomatopée est régi par des règles syntaxiques guidées par des choix esthétiques, pratiques ou de style musical. De la même façon qu'à une langue sont associés des dialectes, ce langage va présenter des variantes selon le style musical et le groupe social auquel le musicien appartient, qu'on regroupe sous le nom de *gharana* [Kip88].

Pour une gharana, une percussion, et un contexte syntaxique donnés, le bol utilisé va dépendre d'un certain nombre de paramètres qu'on peut rapprocher de ceux de l'articulation.
– Il va dépendre tout d'abord de la partie de la main qui frappe l'instrument et de la zone de frappe, les percussions indiennes étant souvent dotées de plusieurs peaux disposées de manière concentrique et d'une pâte collée en leur centre. L'analogie avec l'articulation peut être une constriction principale entre un articulateur actif et un autre passif.
– Il va dépendre également de l'existence ou non d'une frappe simultanée avec l'autre main. On peut faire l'analogie avec une double articulation, les deux frappes étant de degré similaire. Par exemple en tablas, la frappe fermée de l'index de la main préférée au centre du tablas associée au bol « Tê », et « Ghe » est celle qui est associée à la frappe ouverte de la main secondaire sur le tablas grave. Si les frappes sont effectuées simultanément, alors l'association de ces deux frappes sera considérée comme un bol à par entière, dénommé « Dhe ». Il en est de même avec deux consonnes qui ne se différencient que par une constriction : si une langue possède ces deux consonnes, alors la consonne qui possède une constriction de plus sera perçue comme une consonne à part entière et non comme la superposition des deux consonnes.
– La frappe, et donc le bol associé, vont aussi être caractérisés par la localisation des autres doigts de la main qui empêchent la résonance de certaines harmoniques. L'analogie est

une ou plusieurs constriction(s) secondaire(s).
– Enfin, la frappe sera perçue différemment si elle est fermée (vibration de la peau empêchée) ou ouverte (après la frappe, la partie de la main est retirée pour laisser la vibration libre). L'analogie est le caractère voisé ou non (sonore ou sourd) d'une consonne.

L'analogie peut être poursuivie avec la coarticulation. Outre la frappe simple et double, un bol peut définir deux frappes très rapprochées dans le temps, qui constitue en terme européen un « fla ». En tablas, le bol « Ke » suivi du bol « Re » dans un laps de temps très bref devient le bol « Kre ».

En tablas, une seule des frappes permet d'être modulée en hauteur et d'une manière continue, c'est celle qui correspond au bol « Ghe » : après avoir été frappée, la peau est plus ou moins pressée à l'aide du poignet de la partie de la main qui a frappé. Selon le renommé joueur de tablas et de sitar Nayan Ghosh, le jeu des tablas peut se comparer à de la parole dont l'intonation provient de la modulation de ce « Ghe ».

Cette relation proche entre percussion et parole est renforcée par le fait que cette imitation n'est pas seulement didactique mais aussi acoustique, comme l'ont montré Patel et Iversen [PI03] dans le cas des tablas. Une corrélation existe entre les signaux sonores des bols et de leur frappe associée, quand on considère des descripteurs tels que le barycentre spectral, la fréquence fondamentale, l'enveloppe temporelle. De plus, des sujets non connaisseurs parviennent à associer correctement les bols aux frappes de la percussion.

Ces onomatopées ne sont pas utilisées seulement pendant l'enseignement. Dans les concerts de tablas solo, l'exposé oral d'une composition précède souvent son interprétation. Sa fonction est didactique pour l'auditeur, mais relève aussi un intérêt en tant que performance à part entière. En inde du sud, les « bols » sont appelés « konnakol » et ce terme désigne également l'art de réciter ces syllabes. Ici, l'aspect de performance est complètement assumé, prenant plus de liberté quant à son attachement aux percussions. On trouvera dans un orchestre des musiciens dont la tâche est totalement vouée à ces récitations. Dans la musique populaire, ces onomatopées sont également largement utilisées en dehors de tout aspect didactique. On peut citer aussi les pratiquants de danses savantes indiennes, notamment Kathak, qui dansent sur des compositions fixes de tablas. Tout comme le joueur de tablas qui expose son jeu par une récitation, le danseur Kathak récite la composition des tablas avant de danser la composition, jouée par un tabliste. Ces mouvements de danse se font sur la récitation intérieure de ces bols.

Une autre utilisation de la voix comme percussion est le *human beatbox* qui est l'art d'imiter des instruments de musique, principalement de percussion, avec l'appareil vocal. Elle permet d'accompagner d'autres musiciens en prenant la place de certains instrumentistes, mais sa fonction peut être aussi ludique. Le beatboxer sera à la fois jugé sur la qualité rythmique/mélodique de son morceau, sur sa faculté à imiter tel ou tel instrument qu'il soit acoustique ou électronique, et à les enchevêtrer pour donner l'illusion de plusieurs instruments imités en même temps. Les phonèmes sont utilisés aussi de manière mnémotechnique. Par exemple, les occlusives labiales comme /p/ ou /b/ sont le point de départ articulatoire pour imiter des sons de grosse caisse de batterie alors qu'une fricative palatale comme /ʃ/ permet d'imiter des frottements telle que l'utilisation des balais sur une caisse claire.

On peut ainsi voir la voix non seulement comme un instrument à hauteur continue tel le violon, mais aussi comme un instrument de percussion dès lors qu'on prend en considération l'articulation des phonèmes.

1.5 La voix comme objet de synthèse pour le jeu musical

Maintenant que nous avons décrit l'appareil vocal sous différents points de vue (comme émetteur sonore, producteur de phonèmes et instrument de musique), nous allons passer à sa simulation, correspondant à un objectif de cette thèse.

1.5.1 Méthodes de synthèse vocale

L'objectif de la synthèse vocale comme instrument pour la performance est de produire des sons de parole en contrôlant en temps réel le maximum de paramètres permettant un jeu riche et expressif. La voix est le seul instrument qui fait partie intégrante du corps du musicien. Cela a au moins deux conséquences. D'une part, par de tels systèmes de contrôle gestuel de synthèse vocale, on déplace en quelque sorte la voix dans la catégorie des instruments de musique externes au corps. D'autre part, l'instrument de voix naturelle étant à disposition de chacun « par défaut », tout le monde est expert de cet instrument tant du coté perception que production. Ainsi, les jugements sur la qualité de la synthèse sont très sévères, sans doute beaucoup plus que sur d'autres instruments qu'on essaye de synthétiser, du moins par le grand public. Pour reprendre l'illustration de Cook [Coo11], le jugement du grand public sur les synthétiseurs de voix est comparable à un expert en violon qui parvient à discerner de petites différences entre deux violons.

On peut distinguer deux catégories de méthodes de synthèse vocale : les méthodes par connaissances explicites (« knowledge method ») et les méthodes par connaissances implicites (« ignorant method »). Les premières utilisent des modèles de l'appareil vocal, qu'ils soient de type physique comme la synthèse dite articulatoire, ou de type signal comme la synthèse dite par formants. La deuxième n'utilise pas de modèle de l'appareil vocal mais manipule des segments de voix préenregistrés, c'est la méthode dite par concaténation. D'autres méthodes intermédiaires existent, utilisant des bases de données de voix naturelle dont sont extraits certains paramètres afin de reconstruire le signal à partir de l'élément de la base le plus proche de la cible, comme la synthèse HTS (HMM-To-Speech) utilisant des modèles probabilistes HMM.

Hormis les modèles physiques, l'approche utilisée est phénoménologique, utilisant des notions abstraites issues de l'observation sans référence explicite à l'appareil vocal [WD04].

Une méthode donnée présentant toujours des avantages et des inconvénients, il est important de savoir dans quel but et dans quel contexte on veut produire de la voix de synthèse.

a) La synthèse par formants

La synthèse par formants est basée sur un modèle linéaire source-filtre de production qui utilise des paramètres acoustiques comme entrées du synthétiseur. La source est vue comme le signal à la sortie du larynx ou comme un système physique correspondant au mécanisme de vibration des plis vocaux. Ce signal source est ensuite convolué par des filtres en parallèle ou en cascade dont la fréquence centrale, amplitude et bande-passante correspondent à chacun des formants issus de l'analyse de voix réelle. Sa moindre complexité comparée à un modèle articulatoire permet des applications temps-réel.

En 1922, Stewart inventa l'ancêtre des synthétiseurs à formants, composé d'une source périodique et de deux résonateurs électriques permettant de produire des voyelles, des diphtongues, et quelques mots tels que « mama, anna » [Lié77].

Au début des années 50 apparaissent les premiers véritables synthétiseurs à formants : PAT (Parametric Artificial Talker) fut introduit par Lawrence en 1953, consistant en trois résonateurs électroniques connectés en parallèle [Kla87] ; à peu près au même moment, Gunnar Fant propose le premier synthétiseur à formant en cascade, dénommé OVE I (Orator Verbis Electris). Fant pose les bases de la modélisation source-filtre du système phonatoire et présente notamment un système à base de tuyaux à multiples fréquences de résonance [Fan60].

On peut aussi travailler directement dans le domaine temporel, en utilisant les fonctions d'onde formantique (FOF), qui consistent en une somme temporelle d'impulsions séparées de la période fondamentale. Une FOF est une contribution représentant une période fondamentale d'un signal correspondant à un formant. On obtient la voix de synthèse en sommant ces contributions. Rodet, Potard et Barrière proposent un synthétiseur de voyelles chantées, CHANT, basé sur cette technique [RPB84]. Il est destiné à la composition et ne permet pas un contrôle temps réel du synthétiseur.

b) La synthèse par concaténation

La synthèse par unités concaténées est caractérisée par la concaténation de segments de parole pré-enregistrée, la plupart utilisant le diphone. Pour une synthèse plus naturelle, un algorithme de sélection d'unités de taille non uniforme (diphones, triphones, ...) est employée. Pour ces derniers systèmes, la méthode est spécifique à un locuteur. Si l'on veut une méthode non spécifique, elle nécessite alors des bases de données conséquentes. On peut utiliser en parallèle des méthodes de modification de hauteur, comme PSOLA par exemple, pour agir sur la prosodie [HMC89] [MC90].

Le logiciel commercial Vocaloid [KO07] présente un succès important dans le grand public en utilisant la synthèse par concaténation. Il permet, en achetant la base d'unités d'un chanteur donné, de composer ses morceaux vocaux et de les synthétiser.

c) La synthèse par HMM

On peut aussi utiliser une base de données de parole en employant une méthode de reconstruction de séquence de phonèmes basée sur les HMMs (Hidden Markov Models). Les segments de parole sont étiquetés et on extrait des descripteurs spectraux sur chacun d'eux, puis on entraîne le Modèle de Markov Caché sur cette base. Lors de la phase de synthèse, on reconstruit des séquences de phonèmes à l'aide d'un modèle source-filtre. La forme d'onde est générée en utilisant les valeurs de F_0 et des descripteurs spectraux des séquences les plus proches de la cible à l'aide de calculs probabilistes basés sur les HMMs. Cette méthode de synthèse est appelé HTS [YTM+99] [ZTB09].

d) La synthèse par modèle physique

La synthèse articulatoire repose sur une évolution dynamique des articulateurs sollicités au cours du processus phonatoire. Contrairement à la synthèse par concaténation, aucun segment de parole naturelle n'est stocké, le système s'appuie sur la modélisation explicite du mécanisme humain de production de la parole. Dans ce domaine, on peut s'intéresser au comportement d'un articulateur en particulier ou avoir des visées plus globales de l'appareil vocal pour la synthèse. Elle s'appuie sur trois sous-modèles ([KB08]) :

- génération des mouvements des articulateurs (modèle de contrôle) ;
- conversion des informations de mouvement en une succession continue de géométries du conduit vocal (modèle du conduit vocal) ;

– génération du signal acoustique sur la base de ces informations géométriques (modèle acoustique).

Les plis vocaux peuvent être représentés par un système mécanique oscillant composé de deux masses maintenues chacune à un support fixe par un ressort amorti linéairement et de tension non linéaire, et reliées entre elles par un ressort de raideur linéaire. Le larynx a été modélisé par Titze à l'aide d'un système multi-masses [Tit73] et par Perrier avec un modèle à poutres [Per82]. Bien que plus physique, ces modèles sont complexes et plus difficiles à contrôler que des modèles à deux masses.

Coker présente un modèle du conduit vocal à 7 paramètres : position du corps de la langue, arrondissement et protrusion des lèvres, lieux et degré de constriction de la pointe de la langue, et degré de couplage avec la cavité nasale [Cok68]. Pour calculer l'onde acoustique résultante, il est souvent utilisé des mesures radiographiques pour constituer une base de données de fonctions d'aire qui décrivent l'aire des coupes sagittales à travers le conduit vocal, comme dans le modèle de Maeda [Mae79].

En 1986, Saltzman crée le Task-dynamic Model, modèle articulatoire dynamique reposant sur le geste phonétique (la fermeture labiale en est un exemple) et les contraintes de coordination entre les articulateurs. Il est basé sur l'activation temporelle de chaque geste et la coordination entre les articulateurs contrôlée à chaque instant au cours des gestes [Sal86].

Les modèles récents du conduit vocal et de la transformation acoustique, incluant les modèles de glotte et de source de bruit, offrent des voix de bonne qualité, surtout pour les voyelles statiques et les consonnes comme les fricatives, latérales et nasales, en voix chantée comme dans les travaux de Birkholz [Bir07]. Dans ce dernier exemple, un ensemble de règles transforme les données d'un fichier XML, comprenant la partition musicale, en partition gestuelle, qui à son tour permet de calculer le son rayonné.

1.5.2 Quelques instruments de synthèse vocale pour le jeu musical

On décrit ici les synthétiseurs vocaux pouvant être contrôlés en temps réel pour des applications musicales au sens large. Ne sont donc pas mentionnés les synthétiseurs vocaux à partir de texte ou de partition musicale, ni les instruments dont la première vocation n'est pas d'imiter la voix, comme par exemple le registre *Vox Humana* de l'orgue ou le Theremin.

Il est difficile de trouver des informations sur le contrôle temps réel de la synthèse articulatoire (i.e. par modèle physique). D'une part, les modèles physiques demandent trop de puissance de calcul pour être utilisables en temps réel, et d'autre part ce n'est pas la préoccupation première de ces chercheurs qui se concentrent sur la modélisation physique. Un problème essentiel pour contrôler de tels modèles provient du grand nombre de paramètres. Kroeger insiste sur l'importance de l'amélioration des modèles de contrôle, mais ne semble pas se préoccuper de temps réel, mais seulement de modèles de génération des mouvements des articulateurs [KB08].

a) Machines mécaniques

Avant l'arrivée de l'électricité, en 1780, Kratzenstein fabriqua une machine reproduisant la forme du conduit vocal pour la production de 5 voyelles [Bre83]. Suite à une question ouverte au public posé par l'académie des Sciences de Saint Petersbourg qui était « physiologiquement, qu'est-ce qui fait que les voyelles /a,e,i,o,u/ sonnent différemment ? », il a construit une machine reproduisant ces voyelles. Les formes des résonateurs sont assez différentes de la réalité excepté pour le /a/, laissant penser qu'il a trouvé ces formes par essais

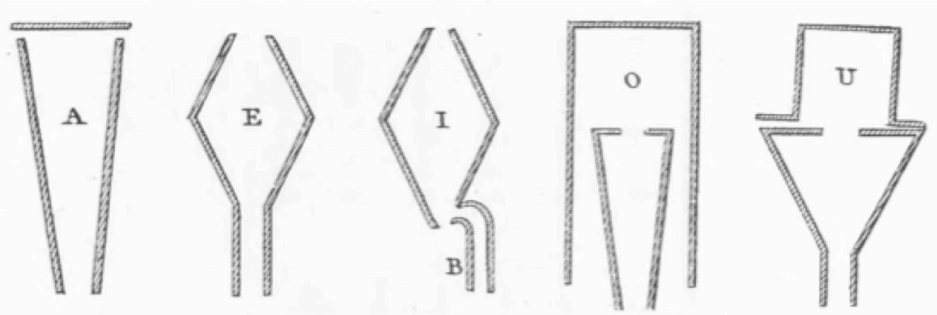


FIGURE 1.5 – *Les formes des tubes de la machine de Kratzenstein pour chacune des voyelles, d'après [DT50]*

et erreurs [Oha11] (voir figure 1.5). L'auteur cite l'abbé Mical, qui aurait aussi construit une machine parlante quelques années auparavant mais sans publier son travail.

En 1791, une autre machine mécanique, celle de Von Kempelen, pouvait émettre une vingtaine de sons différents. Elle était composée d'un soufflet, d'une bouche avec le volume contrôlé par une main pour l'émission des voyelles, de narines et de sifflets actionnés par des leviers à l'autre main pour l'émission des consonnes [DT50] (voir figure 1.6).

b) Le VODER

Une date importante est 1939, avec le VODER par Dudley des laboratoires Bell. C'est une modification du Vocoder, un système ayant pour but de réduire le débit des transmissions téléphoniques en codant l'évolution du spectre du signal suivant des bandes de fréquences fixes. Le VODER permet de contrôler l'activation des résonateurs électriques par des commandes manuelles à l'aide d'un clavier et d'une pédale. Les opérateurs devaient être entrainés pendant au moins un an avant de pouvoir synthétiser des phrases devant le public [DRW39] (voir figure 1.7).

c) Le MUSSE

En 1977, Larsson publie dans le cadre de sa thèse de doctorat un synthétiseur de voyelles à formants réglé pour le chant et contrôlable par un clavier de piano, le *MUsic and Singing Synthesis Equipment* (MUSSE, voir figure 1.8)) [Lar77]. La fréquence fondamentale F_0 de la voix est ajustée sur une gamme tempérée, mais les quantités de vibrato, de bruit glottique, et de variation aléatoire de F_0 peuvent être modifiées par des boutons. La production des consonnes a été entreprise en 1979 par Ponteus, mais peu d'informations sont disponibles, la thèse [Pon79] étant en Suédois (cité dans [Sun06]). En 1984, Zera présente un système basé sur MUSSE auquel il rajoute deux générateurs de bruit et des filtres supplémentaires pour la production des consonnes présentes dans les notes de musique /do, re, mi, fa, sol, la, si/. Les syllables sont contrôlables à l'aide d'un programme de synthèse à partir de texte [ZGS84].

d) Le SPASM

En 1991, Cook crée le système SPASM qui repose sur une modélisation du conduit vocal par un modèle à guide d'onde numérique [Coo91]. Un premier guide d'onde modélise la forme du conduit vocal correspondant à la partie orale et un autre les anti-résonances situées au

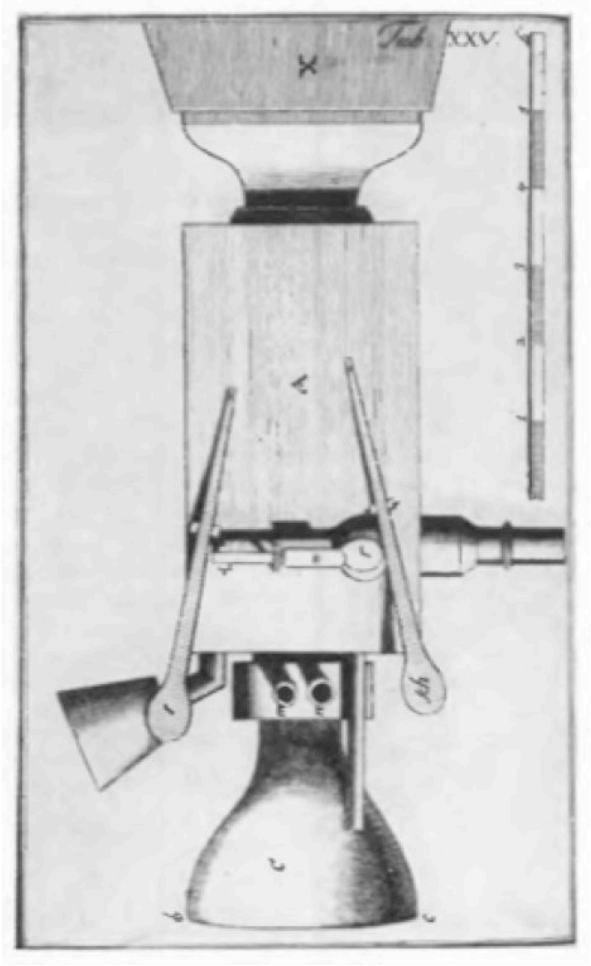


Figure 1.6 – *La machine de Kempelen, d'après [DT50]*

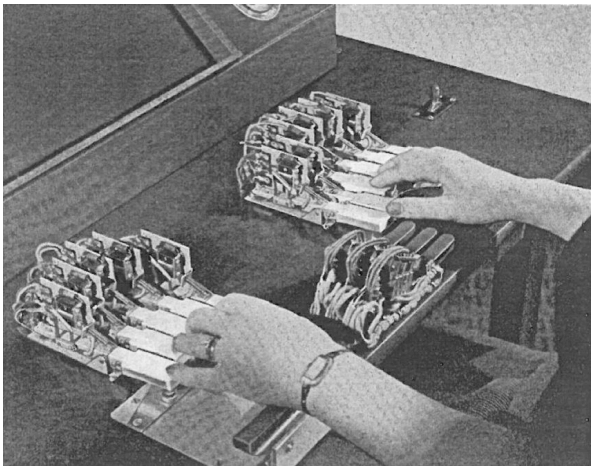


Figure 1.7 – *Le clavier du VODER, d'après [DRW39]*

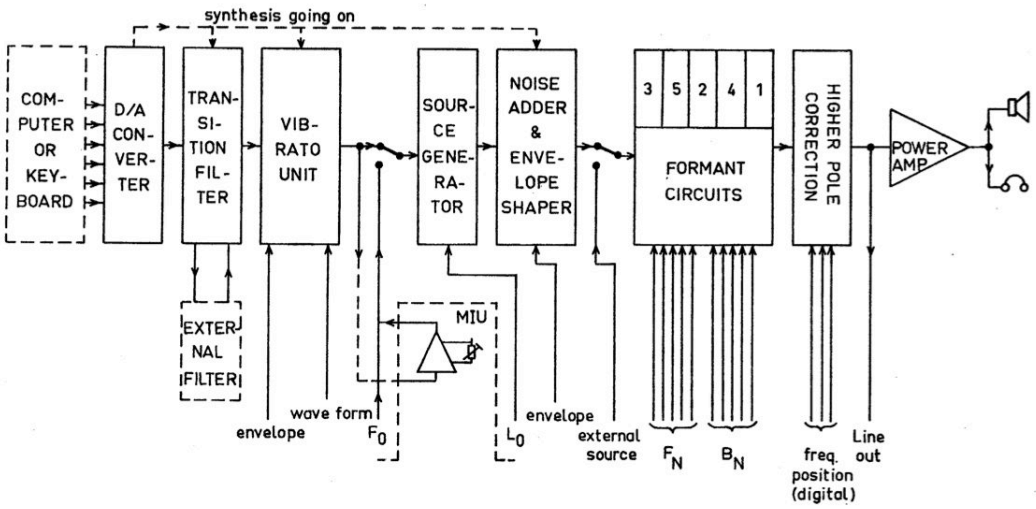


FIGURE 1.8 – *Diagramme de fonctionnement du MUSSE, d'après [Lar77]*

niveau du conduit nasal. La source glottique est quant à elle modélisée par des tables d'onde et une source de bruit modulé. Plus tard, Cook et Leider présentent une interface de contrôle basée sur un accordéon, le SqueezVox [CL00]. Cet instrument à vent est proche de la voix dans la mesure où il fonctionne avec un réservoir d'air. Ainsi, le souffle est contrôlé par le soufflet de l'accordéon, la hauteur tempérée est contrôlée par son clavier, tandis que le contrôle fin et le vibrato sont gérés par une bande tactile linéaire avec l'autre main. Une série de boutons a été ajoutée pour contrôler les voyelles et les consonnes (voir figure 1.9).

e) Glove-Talk II

Fels et Hinton utilisent des réseaux neuronaux pour implémenter une interface adaptative, appelée le Glove-TalkII, qui relie des mouvements de main aux paramètres de contrôle d'un synthétiseur à formants pour permettre à l'utilisateur de parler avec ses mains [FH98] [FH92]. Cette adaptation du mapping se fait automatiquement pendant la phase d'entrainement. Les deux premiers formants des voyelles sont contrôlés par la main gauche, les consonnes par la main droite, exceptées les occlusives qui sont contrôlées avec une pédale. La figure 1.10 donne le schéma de fonctionnement du Glove-TalkII.

f) Contrôle gestuel du programme CHANT

CHANT est un synthétiseur par formant utilisant des Fonctions d'Onde Formantique (FOF) [RPB84] (voir section 1.5.1) publié en 1984. Wanderley *et al.* utilisent une implémentation Max de ce programme écrit par Iovino and Dudas pour en faire un instrument de voix chantée en 2000. Aucune pratique musicale n'est décrite, mais ils sont à notre connaissance les premiers à proposer et tester une tablette graphique pour le contrôle gestuel de synthèse vocale [WVIR00].

En s'appuyant notamment sur les travaux de Vertegaal *et al* [VUK96], ils associent fonctions musicales aux meilleurs types de capteurs, variables du synthétiseur vocal aux fonctions musicales et aux types de geste, et enfin variables de la tablette aux types de capteur. A partir de cette classification, ils proposent d'augmenter la tablette d'un capteur de position et pression, pour aboutir aux correspondances suivantes :

Figure 1.9 – *Le squeezeVox Lisa, d'après [CL00]*

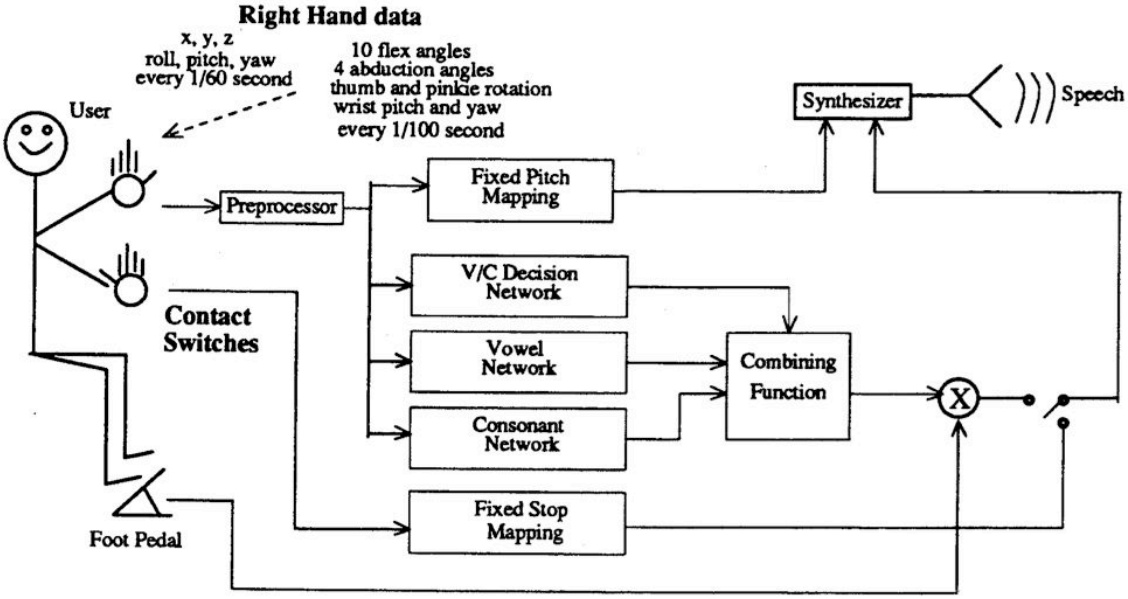


Figure 1.10 – *Diagramme de fonctionnement du Glove-TalkII, d'après [FH98]*

FIGURE 1.11 – *Configuration « Tablette + Joystick » du Voicer, d'après [Kes04]*

- La fréquence fondamentale F_0 de la voix de synthèse avec la position détectée par le capteur additionnel
- La modulation de F_0 (vibrato) avec la pression sur ce même capteur
- Les paramètres des FOFs correspondant aux fréquences centrales des formants F1 et F2 avec la position du stylet sur le plan de la tablette, les autres paramètres des FOFs étant fixés.

g) Le Voicer

Dans sa thèse de doctorat, Kessous présente le Voicer, instrument de synthèse de voyelles chantées, utilisant notamment cinq FOFs en parallèle. Dans la dernière version, le triangle vocalique est contrôlé par un joystick, la source glottique par une tablette graphique (voir figure 1.11). Il propose un mapping intéressant pour F_0 consistant sur la tablette en une spirale où chaque tour complet correspond à un intervalle d'octave [Kes04].

h) Le HandSketch

Le HandSketch est une interface basée sur une tablette graphique, augmentée par des boutons formés de capteurs FSRs [dD07], et tenue verticalement le long du corps [dD09] (voir figure 1.12). Les boutons à la main secondaire permettent de modifier la hauteur mélodique de manière discrète de quelques demi-tons, tandis qu'à la main principale, la mélodie est contrôlée continûment sur une grande étendue fréquentielle, le long d'un arc de cercle correspondant à la courbure du bras. Plusieurs modèles de synthèse sont contrôlables : synthèse par formants et le modèle RT-CLAM [ddLBD06], lecture et modification temps-réel de fichiers, et synthèse par HMM avec le système MAGE [AdD12].

1.6 Gestes instrumentaux

Pour contrôler un instrument de musique, le musicien doit exercer certains gestes qui sont eux-mêmes capturés par l'instrument. Il en sera de même pour notre instrument de voix de synthèse. Ainsi, il est intéressant de réfléchir aux types de gestes à effectuer pour une fonction

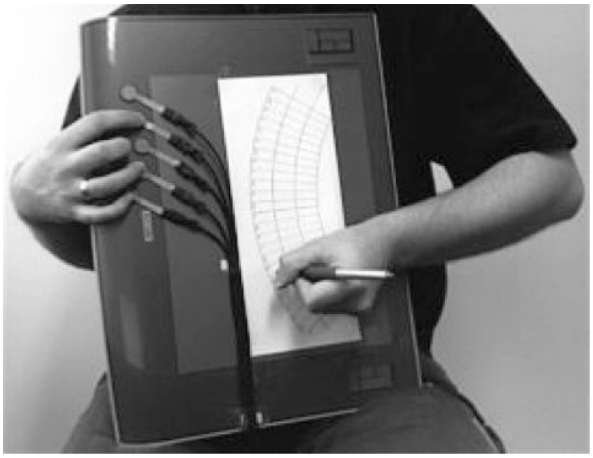


Figure 1.12 – *Le HandSketch : Tablette graphique avec calque de repères et 8 capteurs de pression FSRs, d'après [dD07]*

musicale donnée, et la manière de capter ces gestes afin de les envoyer vers un synthétiseur de sons.

1.6.1 Catégorisation des gestes instrumentaux

À tout instrument sont associés des gestes de contrôle qui permettent de faire le lien entre l'interface de l'instrument et le corps du musicien, avec ou sans intermédiaire. Par exemple, le pianiste devra exercer un mouvement du doigt de haut en bas pour presser une touche du piano et ce geste sera issu d'un mouvement plus global pouvant comprendre d'autres parties du corps, du poignet jusqu'au buste.

L'arrivée des instruments électriques, électroniques et numériques a rompu le lien entre énergie du geste et énergie acoustique [Gen99]. Avec un instrument acoustique, l'énergie acoustique ne peut que provenir de l'énergie gestuelle donnée par le musicien (à l'exception d'instruments acoustiques munis d'autres sources d'énergie comme l'orgue). Les instruments amplifiés se voient fournir une énergie électrique pour décupler l'énergie gestuelle voire la transformer. Quant aux instruments électroniques et numériques, le lien de causalité est rompu ou alors reconstruit virtuellement.

Plusieurs approches d'analyse des gestes musicaux sont possibles. Ramstein [Ram91] en propose trois :

- une approche phénoménologique qui consiste à décrire les gestes suivant leur vitesse, position et fréquence ;
- une approche fonctionnelle qui traite de la fonction des gestes considérés ;
- une approche intrinsèque qui exprime le point de vue du producteur des gestes. Elle ne sera pas traitée dans cette thèse.

On peut classifier les gestes musicaux suivant deux grandes catégories fonctionnelles : les gestes nécessaires à la production du son, et ceux non nécessaires mais présents chez les instrumentistes de haut niveau [WD04].

1.6.2 Analyse fonctionnelle des gestes instrumentaux

Selon Cadoz [Cad88], un geste instrumental est un geste appliqué à un objet matériel avec lequel on interagit physiquement. Le résultat de cette interaction est un phénomène physique dont l'évolution peut être maitrisée par l'utilisateur. Parmi les gestes nécessaires à la production du son, les gestes instrumentaux peuvent être classifiés en trois catégories selon leur fonction [Cad88] :

1. Les gestes d'*excitation*, source d'énergie pour la production sonore, et qui peuvent être :
   - *instantanés* (percussifs ou pincés) ;
   - *continus* quand le geste et le son coexistent.
2. Les gestes de modification, changeant les propriétés de l'instrument sans ajouter d'avantage d'énergie à l'instrument (modification affectant la relation entre le geste d'excitation et le son), et qui peuvent être :
   - *paramétriques* (ou *continus*) entraînant une modification continue de paramètres de l'instrument, comme bouger le doigt sur une corde de violon pour en changer la hauteur ;
   - *structurels* dans le cas d'insertion ou de retrait de partie(s) de l'instrument, par exemple si on applique une sourdine à une trompette.
3. Les gestes de *sélection*, qui permettent de faire un choix parmi des éléments similaires dans un instrument, de manière séquentielle ou parallèle.

1.6.3 Capture des gestes

Marshall *et al.* ont établi un inventaire des capteurs utilisés dans les instruments numériques présentés à la conférence NIME de 2001 à 2008 [MHWL09]. En comptant seulement les capteurs différents par instrument, 26% de ces capteurs étaient des FSRs (Force Sensing Resistor), 21% des accéléromètres, 20% des caméras vidéo, et 19% des bouttons ou des interrupteurs.

Ces différents capteurs sont utilisés pour être mise en correspondance avec les fonctions musicales de l'instrument numérique considéré, pour lesquelles Verteegal *et al.* distinguent trois catégories [VUK96] :

- les fonctions dynamiques absolues, comme la sélection absolue de la hauteur mélodique, de l'amplitude ou du timbre ;
- les donctions dynamiques relatives, comme la modulation d'une hauteur mélodique donnée, d'une amplitude donnée, ou d'un timbre donné ;
- les fonctions statiques, comme la sélection de la tessiture, d'une gamme, ou d'une transposition.

Le but de Marshall *et al.* [MHWL09] était de déterminer un choix optimal de capteur pour le contrôle de la modulation d'une hauteur mélodique. Il a été demandé à 27 pianistes et violonistes de moduler un son par différentes méthodes : en glissant (*sliding*), en appliquant une pression, ou en produisant un mouvement de rotation autour de la pointe du doigt en contact (comme le font les violonistes pour créer un vibrato). Les préférences des utilisateurs et des mesures sur la vitesse de modulation étaient considérés. Les résultats concluent à une préférence de ces sujets pour la méthode de modulation par pression, et aucune influence de l'expérience musicale des sujets n'a été relevé pour le choix de la méthode. Seule la main préférée utilisée pour la tâche de modulation dépendait du passé musical : les violonistes préféraient moduler le son avec leur main secondaire alors que les pianistes préféraient leur main dominante. En ce qui concerne les violonistes, cela correspond à la main utilisée pour produire un vibrato avec leur instrument usuel.

En se basant sur les quatre premières années inventoriées de la conférence, Marshall et Wanderley ont procédé à des expériences évaluant les préférences d'utilisateurs (majoritairement musiciens) lors de tâches musicales simples [MW06]. Les capteurs testés, correspondant aux plus utilisés et représentant toutes les catégories existantes dans leur inventaire, reflétaient les catégories suivantes :

- potentiomètre linéaire ;
- potentiomètre rotatif ;
- capteur de position linéaire ;
- capteur de force ;
- capteur de courbure (*bend*).

Les sujets de l'expérience devaient utiliser un système de synthèse dont ils peuvent contrôler l'émission de chacune des notes d'une séquence à l'aide d'un bouton à la main secondaire, et la fréquence fondamentale (discrète en demi-ton) avec la main principale à l'aide du capteur étudié.

Pour la tâche de déclenchement des notes séquencées, les utilisateurs ont préféré le capteur de position linéaire. Pour celle de la modulation de F_0 (trilles), les utilisateurs sont partagés entre le capteur de position linéaire (ceux-ci utilisent une technique à deux doigts) et le capteur de force (ceux-là utilisent une technique à 1 doigt pour le capteur de position linéaire). Concernant la dernière tâche combinant les deux tâches précédentes (déclenchement et modulation) avec le même capteur, les meilleurs préférences sont obtenues pour les capteurs de position linéaire et le potentiomètre linéaire.

1.7 Outils existants au LIMSI-CNRS au début de la thèse

Les travaux du laboratoire sur la modélisation de la source glottique [DdH03] [DdH06] [LB09] ont donné lieu à un modèle de l'appareil vocal, basé sur la synthèse par formants. La synthèse par formants fonctionne avec un modèle source-filtre, où des filtres modélisent les résonances du conduit vocal et le rayonnement aux lèvres, excitées par le modèle de source glottique RT-CALM [ddLBD06]. Le faible coût de calcul de cette méthode de synthèse permet des applications temps réels, implémentées dans le langage Max/MSP. Plusieurs interfaces de contrôle ont été testées, allant du clavier aux gants haptiques [ddLB+05], en passant par la tablette graphique [LB09].

La représentation simplifiée de ce modèle de production source-filtre est donnée à la figure 1.13.

1.7.1 Le modèle de source glottique RT-CALM

Le modèle CALM (Causal / Anticausal Linear Model) [DdH03] d'Onde de Débit Glottique (ODG) travaille dans le domaine spectral et s'appuie sur une analyse des principaux modèles temporels proposés dans la littérature [HdD01] [Hen01] [DdH06]. L'ODG est souvent représentée par sa dérivée (nommée ODGD), regroupant ainsi le filtre passe-haut résultant du rayonnement aux lèvres. Cette commutativité est rendue possible par les propriétés linéaires d'un tel système. Les propriétés spectrales de la source sont décrites par la forme du spectre de l'ODGD dans les basses fréquences, appelée « formant glottique » et dans les hautes fréquences, par l'inclinaison spectrale du spectre de l'ODGD. Notons que le « formant glottique » ne correspond pas à un formant à proprement parler car il ne résulte pas de la résonance des plis vocaux, mais de leur vibration. Les caractéristiques du formant glottique

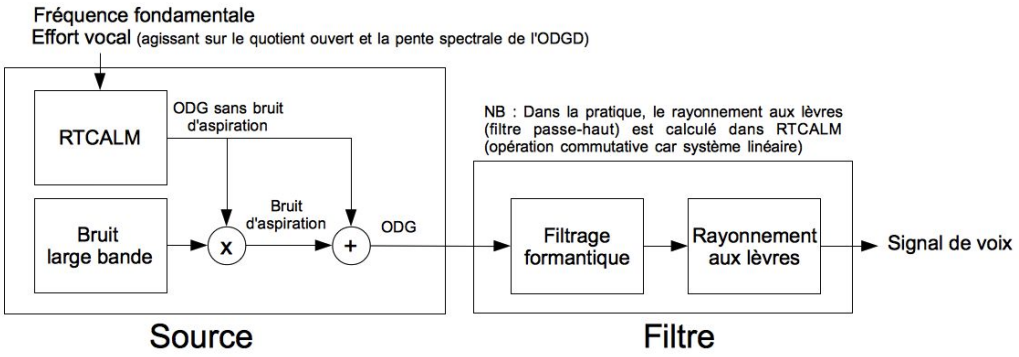


FIGURE 1.13 – *Représentation simplifiée du modèle source-filtre du Cantor Digitalis*

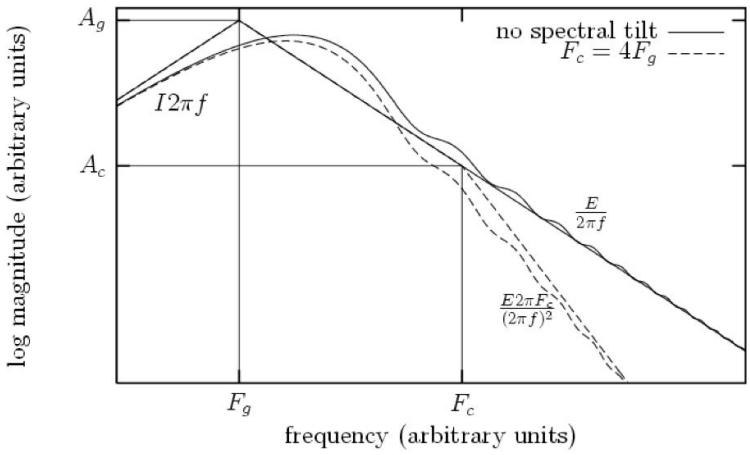


FIGURE 1.14 – *Spectre de l'ODGD avec formant glottique (Fg, Ag) pour deux pentes spectrales (Fc, Ac), d'après Doval et al. [DdH03]*

sont : sa fréquence centrale ; sa largeur de bande ; son amplitude. La figure 1.14 représente le formant glottique et la pente spectrale de l'ODGD.

L'ODG est implémentée numériquement grâce à deux filtres en cascade : un filtre passe-bas résonant du second ordre anti-causal pour le formant glottique, et un filtre passe-bas du premier ordre causal pour l'inclinaison spectrale. Le filtre anti-causal peut poser problème pour une implémentation temps-réel, puisque pour calculer le signal à la sortie du filtre on a besoin de sa valeur dans le futur. Plusieurs solutions sont possibles pour implémenter la partie anti-causale. La première est de se permettre un retard d'une période fondamentale, et ainsi d'avoir accès aux valeurs futures nécessaires au calcul. La deuxième solution est de calculer l'expression analytique de la réponse impulsionnelle du filtre anti-causal stable en le transformant en filtre causal instable, et de s'arrêter à l'instant de fermeture glottique où le filtre diverge. Ce modèle de la source glottique, modifié pour le temps réel avec la première solution d'implémentation, s'appelle RT-CALM (Real-Time Causal Anti-Causal Linear Model) [ddLBD06].

Les paramètres d'entrée du RT-CALM sont la fréquence fondamentale F_0, le taux d'aspiration, les apériodicités (Jitter et Shimmer), et des paramètres caractérisant dans le domaine

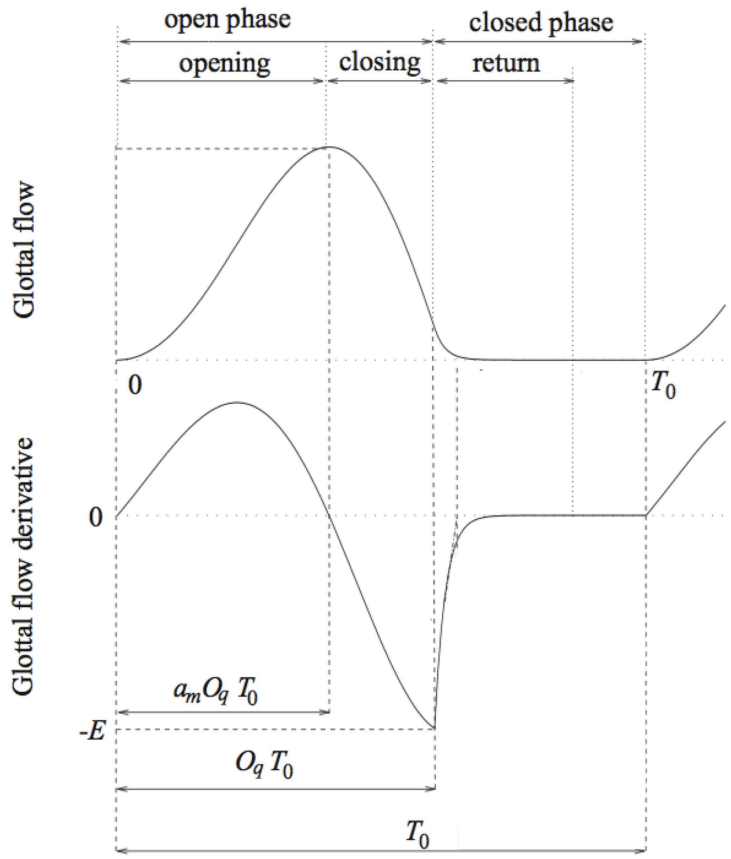


Figure 1.15 – *L'ODG et l'ODGD, et ses paramètres O_q et α_m sur une période fondamentale T_0, d'après Doval et al. [DdH06]*

temporel ou fréquentiel le signal issu de la source (et sa dérivée ODGD), contrôlables par des paramètres de plus haut-niveau tels l'effort vocal et la tension de la voix. L'effort vocal agit à la fois sur l'amplitude du signal et sur la pente spectrale de l'ODGD (voir figure 1.14 pour le spectre de l'ODGD pour deux pentes spectrales). La tension de la voix correspond à la tension des plis vocaux, reliée au quotient ouvert O_q et au coefficient d'asymétrie α_m, coefficients caractérisant la forme temporelle de l'ODG et de l'ODGD (illustrés sur le figure 1.15).

Les fonctions de transfert des deux filtres utilisés pour modéliser l'ODGD, l'expression mathématique de α_m et O_q dans le modèle CALM sont données à l'annexe A.

1.7.2 Modèlisation des résonances du conduit vocal

L'ODG représente seulement la partie « source » de notre modèle « source-filtre » . La partie « filtre » représente ce qui se passe dans le conduit vocal au-dessus de la glotte, autrement dit comment l'ODG est transformée par son passage dans le conduit vocal jusqu'à l'extérieur de la bouche.

La position de chacun des articulateurs définit un volume dont la forme globale correspond à des résonances fréquentielles particulières. L'évolution temporelle des positions des différents articulateurs modifie le signal sonore. On peut donc utiliser des filtres résonnants pour reproduire le comportement des articulateurs, filtres qui seront caractérisés par leur

fréquence centrale, amplitude et bande passante et évolueront dans le temps. Il est utilisé en parallèle quatre filtres résonants du deuxième ordre d'équation aux différences :

$$y[n] = gain \cdot (x[n] - r \cdot x[n-2]) + c_1 \cdot y[n-1] + c_2 \cdot y[n-2] \tag{1.4}$$

où r, c_1, et c_2 sont des paramètres calculés à partir de la fréquence centrale et de la bande-passante du filtre, et y[n] (resp. x[n]) représente la sortie (resp. l'entrée) du filtre à l'instant n.

Le lecteur pourra se reporter aux articles de Holmes [Hol83] et de Klatt [Kla80] pour une discussion entre filtres en parallèle et en cascade pour la synthèse par formants. À chaque voyelle correspond un ensemble de formants qui la caractérise. Ainsi, en agissant directement sur la fréquence, l'amplitude et la bande passante des filtres formantiques, on modélise la position des articulateurs correspondant à une voyelle donnée. On dispose de quelques voyelles de référence définies par la fréquence centrale, largeur de bande-passante et amplitude de chacun des quatre filtres formantiques.

Première partie

INSTRUMENTS DE SYNTHÈSE VOCALE

Chapitre 2

Cantor Digitalis, un instrument de synthèse de voyelles chantées

Sommaire

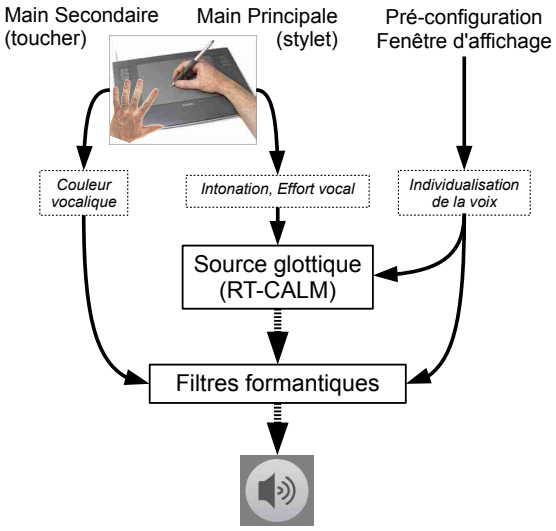


Figure 2.1 – *Représentation schématique du fonctionnement du synthétiseur Cantor Digitalis*

2.1 Introduction

Nous présentons ici le *Cantor Digitalis*, un instrument de synthèse de voyelles chantées, basé sur un modèle source-filtre et la synthèse par formants. L'instrument est configurable pour produire différents types et qualités de voix, et est contrôlable finement à l'aide d'une tablette graphique munie d'un stylet et de capteurs tactiles. Le tout est implémenté dans le langage Max/MSP. La figure 2.1 schématise le fonctionnement du Cantor Digitalis.

On veut contrôler la synthèse de voyelles chantées dans le but de pouvoir jouer principalement sur la hauteur mélodique, l'effort vocal, l'articulation des voyelles, et quelques paramètres de qualité vocale, tout en produisant un rendu aussi naturel que possible. Il convient de réfléchir aux gestes les plus adaptés à ces tâches compte tenu de la spécificité de chacune d'entre elles.

En partant des travaux existant du LIMSI-CNRS, nous utilisons le modèle source-filtre explicité dans la partie 1.7. Celui-ci nous permet un contrôle temps réel des paramètres d'entrée du modèle de source glottique ainsi que des filtres formantiques.

En première approximation, la source glottique peut être considérée comme fonctionnant indépendamment du conduit vocal. Cependant, les interactions source-filtre ne sont pas négligeables. On peut distinguer deux types d'interactions : celles purement mécaniques dues à la production de l'onde acoustique au niveau des plis vocaux qui une fois réfléchie dans le conduit vocal interagit avec la vibration des plis vocaux ; et celles contrôlées consciemment ou non par le chanteur pour optimiser la puissance de sa voix. Ce sont ces dernières qu'on modélisera.

Afin de pouvoir personnaliser la voix de synthèse à contrôler, il importe de pouvoir changer facilement la forme du conduit vocal et les qualités vocales moyennes associées à cette voix. Dans cette direction, on peut aisément tirer avantage du modèle paramétrique de la synthèse par formants pour créer des voix diverses jusqu'aux limites de la normale.

Une des manières de faire vivre une voix de synthèse est son contrôle gestuel, donnant ainsi une certaine variabilité naturelle aux paramètres de haut-niveau comme F_0, l'effort vocal

ou encore la couleur vocalique. Mais les gestes de contrôle ne présentent pas forcément toute la variabilité du geste articulatoire. Ainsi, on peut songer à ajouter un modèle de variabilité qui complètera celle du geste.

On se focalise ici sur la voix chantée. Quelles différences avec la voix parlée ? On peut citer les suivantes :

1. un contrôle fin de la justesse mélodique et du rythme ;
2. la présence éventuelle de vibrato ;
3. une étendue plus importante des fréquences pouvant être parcourues (phonétogramme) ;
4. dans certains styles vocaux, la présence du *formant du chanteur*, c'est-à-dire de l'énergie plus importante dans l'intervalle $2.5 - 3.5kHz$ [Sun01] ;
5. des différences articulatoires pour des mêmes voyelles cibles [Sun69] [JSW04] [HSW11].
6. une possible tenue des sons dans le temps

2.2 Réglage du modèle de source glottique

La figure 2.2 est une capture d'écran du sous-patch Max/MSP modélisant la source glottique. Il est composé de la génération de la source RT-CALM (en orange) décrite dans la partie 1.7.1, d'un bruit modélisant les turbulences au niveau des plis vocaux (en violet), d'un module gérant le contrôle de l'amplitude (en vert) et d'un autre s'occupant du seuil de phonation (en bleu). La source RT-CALM est composée de l'objet Max/MSP *rtcalm* incluant le filtre anti-causal du « formant glottique », et du filtre causal de pente spectrale de l'ODGD calculé à l'extérieur de l'objet *rtcalm* .

Pour éviter des discontinuités de F_0 dans les hautes-fréquences, la source est calculée à une fréquence d'échantillonnage de 8×44100 Hz, puis sous-échantillonnée à $F_e = 44100$ Hz. Un filtre passe-bas de fréquence de coupure à 14000 Hz est ajouté afin de diminuer l'impact du repliement du signal en haute-fréquence sur la qualité du signal.

2.2.1 Effort vocal et pente spectrale

L'élévation de l'intensité de la voix naturelle s'accompagne de changements spectraux, en particulier on observe que la pente spectrale de l'ODGD diminue d'autant plus que l'effort vocal est important. Dans le modèle CALM [DdH03], et comme expliqué dans la partie 1.7 et en annexe A, la pente spectrale de l'ODGD est modélisée par un filtre passe-bas causal. L'équation aux différences entre l'entrée x et la sortie y, et la fonction de transfert $H(z)$ de ce filtre, s'écrivent respectivement :

$$\begin{aligned} y[n] &= b_{TL} \cdot x[n] + (1 - b_{TL}) \cdot y[n-1] \\ H(z) &= \frac{b_{TL}}{1 - (1 - b_{TL}) \cdot z^{-1}} \end{aligned} \tag{2.1}$$

b_{TL} est calculé en fonction de la pente spectrale TL désirée de l'ODGD (définie comme l'amplitude du spectre à 3000 Hz en dB) par :

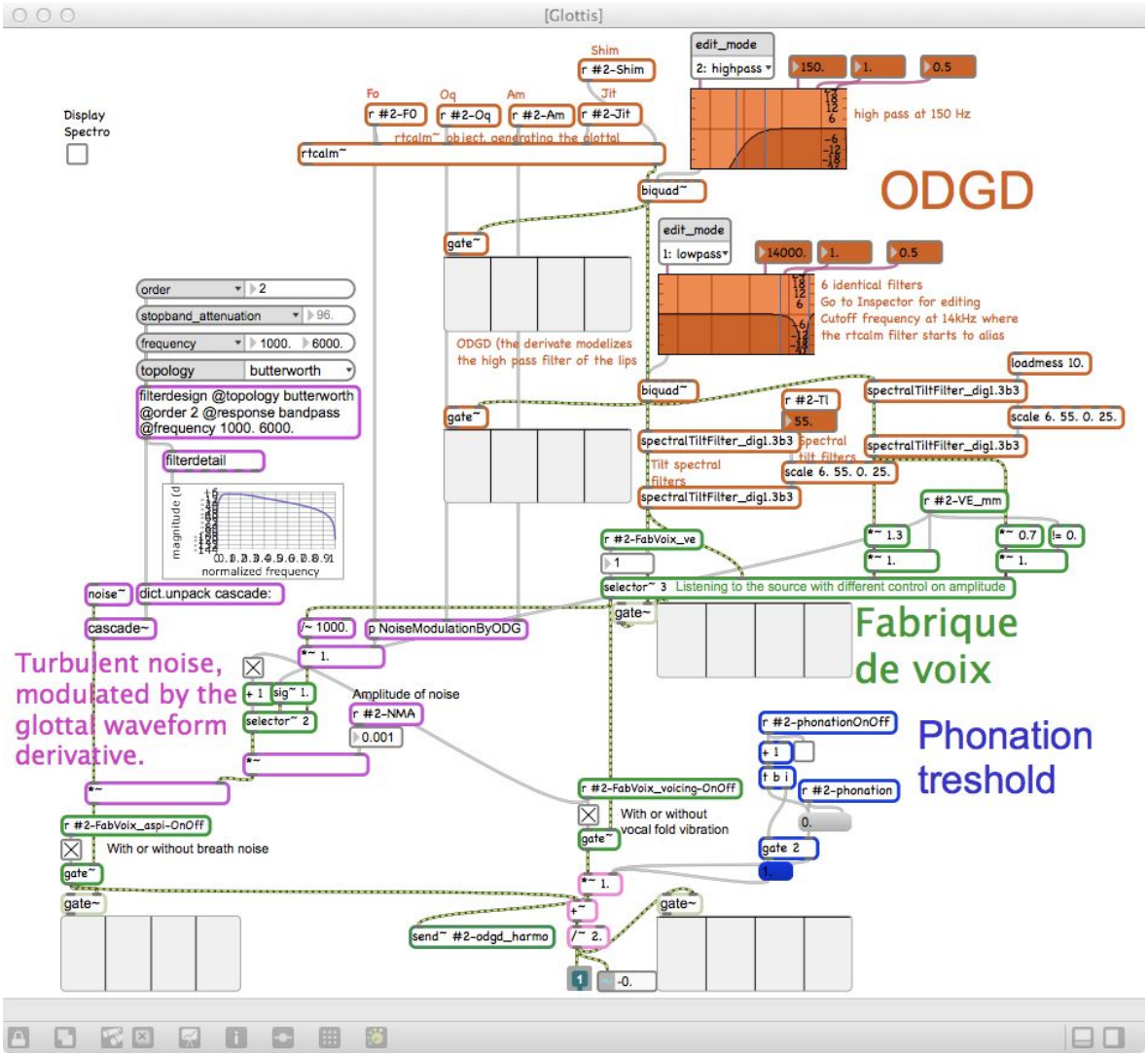


Figure 2.2 – *Capture d'écran du sous-patch Max/MSP du Cantor Digitalis correspondant à la modélisation de la source*

$$\begin{aligned} b_{TL} &= 1 - \nu + \sqrt{\nu^2 - 1} \\ \nu &= 1 - \frac{1}{\eta} \\ \eta &= \frac{10^{TL/10} - 1}{\cos(2\pi \frac{3000}{F_e}) - 1} \end{aligned} \tag{2.2}$$

La réponse en fréquence s'obtient pour $z = e^{2i\pi f/F_e}$ où f est la fréquence d'excitation du filtre. La relation entre pente spectrale de l'ODGD (via b_{TL}) et la fréquence de coupure f_c du filtre est établie par la relation entre les modules de la fonction de transfert du filtre en $f = f_c$ et $f \to 0$:

$$|H(e^{2i\pi fc/F_e})| = \frac{|H(1)|}{\sqrt{2}} \tag{2.3}$$

La réponse en fréquence en $f = 0$ vaut 1 et celle en f_c vaut :

$$|H(e^{2i\pi fc/F_e})| \quad = \quad \frac{b_{TL}}{|1 - (1 - b_{TL}) \cdot e^{2i\pi f_c/F_e}|} \tag{2.4}$$

En élevant au carré (2.3), on obtient :

$$\frac{b_{TL}^2}{2 \cdot (1 - b_{TL})(1 - \cos(2i\pi f_c/F_e)) + b_{TL}^2} \quad = \quad \frac{1}{2} \tag{2.5}$$

soit la fréquence de coupure du filtre en fonction du coefficient b_{TL} :

$$f_c = \frac{F_e}{2 \cdot \pi} \cdot \arccos(1 + \frac{b_{TL}^2}{2 \cdot (b_{TL} - 1)}) \tag{2.6}$$

La correspondance entre le paramètre de haut niveau d'effort vocal et la pente spectrale donnée dans la thèse de d'Alessandro [d'A09] (p. 104) pour l'implémentation temps-réel de CALM ne nous satisfaisait pas complètement. En particulier, en faible effort vocal, la pente spectrale reste trop faible. Pour augmenter cet effet, nous ajoutons un second filtre en série, identique au premier exceptée la correspondance entre pente spectrale et effort vocal.

Pour le premier filtre de pente spectrale, le paramètre d'effort vocal contrôle la gamme de pente spectrale de 6 à 55 dB, comme dans [d'A09]. Pour le second filtre de pente spectrale, nous utilisons la gamme de à 25 dB. Le réglage de ces valeurs limites a été réalisé empiriquement afin d'accentuer la différence de pente spectrale entre effort vocal faible et élevé.

L'implémentation dans Max/MSP est donnée à la figure 2.3 pour les filtres de pente spectrale, et à la figure 2.2 (partie orange, en bas) pour leur intégration dans le modèle de source.

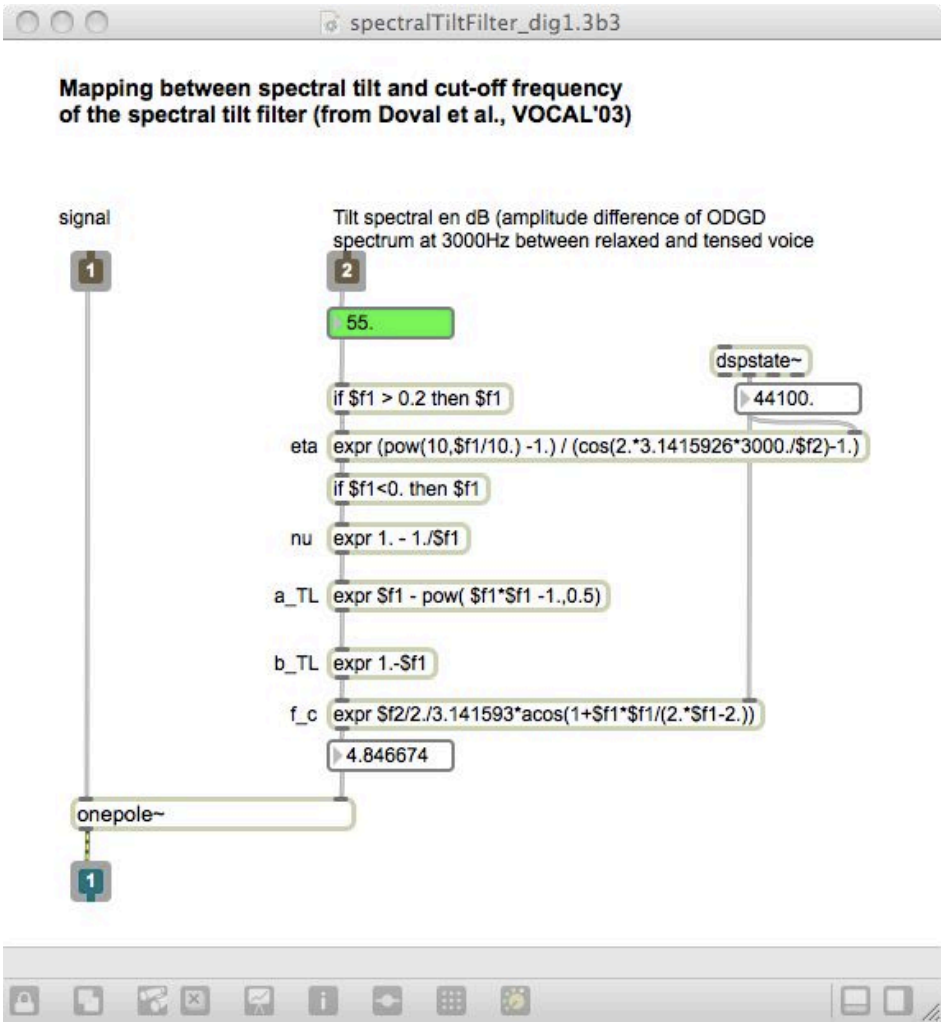


FIGURE 2.3 – *Implémentation en Max/MSP du filtre de pente spectrale*

2.2.2 Seuil de phonation et attaque des voyelles

En voix naturelle, les premiers instants de la production d'une voyelle *piano* se font par l'envoi progressif d'un débit d'air qui, à partir d'un certain seuil met en vibration les plis vocaux. En effet, ceux-là nécessitent une force minimale pour être mis en vibration. Avant cela et en raison de l'écoulement turbulent de l'air issu des poumons à travers les plis vocaux, une source de bruit est créée. Ainsi, on ne peut pas avoir un son voisé avec un niveau sonore aussi faible qu'on le désire : le niveau sonore est borné par une valeur non nulle.

Concernant le modèle RT-CALM, le calcul de l'ODGD débute avec le lancement de l'audio du logiciel Max/MSP et ne s'arrête que lorsque l'audio est éteint. C'est à dire que d'un point de vue temporel, l'onde est constamment en train d'être calculée quel que soit l'effort vocal (même nul). Quand notre voix de synthèse doit commencer à produire un son voisé, alors on applique un facteur d'amplitude non nul à l'ODGD. Par conséquent, on peut remarquer trois principales faiblesses du début de la phonation avec RT-CALM :

- Deux qu'on résout partiellement :
  1. L'absence de seuil de phonation.
     On ajoute un seuil de phonation : l'amplitude de l'ODGD ne devient non nulle qu'à partir d'une certaine valeur de l'effort vocal qu'on fixe à $VE = 0.2$ ($VE \in [01]$). L'opposé, c'est à dire l'arrêt de la phonation, est également soumis à un seuil, mais celui-ci est imposé à une valeur d'effort vocal plus basse, $VE = 0.015$. On reproduit alors l'effet d'Hystérésis du à un potentiel d'énergie à franchir pour atteindre la phonation. Il permet aussi, quand VE est de l'ordre du seuil de phonation, de ne pas alterner d'un état à un autre trop facilement
  2. L'absence de phase transitoire au début de la phonation.
     En voix naturelle, la phase qui correspond à la mise en vibration jusqu'à une certaine stationnarité est complexe. Beaucoup de paramètres de la source, F_0, effort et qualité vocale, doivent évoluer en très peu de temps, et ainsi donner l'attaque caractéristique d'une voyelle. Et ces évolutions sont très variables en fonction de l'intention du chanteur. Dans notre synthétiseur, nous nous contentons de modifier la durée de mise en phonation : 50 ms pour une attaque douce, 1 ms pour une attaque franche, et une interpolation de ces deux valeurs pour des attaques intermédiaires. L'intensité de l'attaque est définie par la vitesse de l'effort vocal VE dès qu'elle devient différente de zéro. Quant à la durée d'arrêt de la phonation, on la fixe à 50 ms car on la considère moins importante à contrôler pour des applications musicales.
- Et une résolue mais non implémentée :
  3. La phase de l'ODGD en début de phonation.
     Contrairement à la réalité, et comme l'ODGD est constamment en train d'être calculée, la valeur de la phase de l'ODGD en début de phonation va dépendre du moment où l'utilisateur choisit de débuter la phonation. Comme la période fondamentale de l'ODG est de l'ordre de quelques millisecondes, l'utilisateur ne peut synchroniser manuellement le début de la phonation avec un zéro de l'onde ODGD. Ainsi, la phase de l'ODGD au début de la phonation sera aléatoire. De plus, la phase n'est pas contrôlable dans l'objet MAX *rtcalm* . Une conséquence va être des attaques différentes (et perceptibles) pour le même contrôle des paramètres de début de phonation. Par exemple, si le début de phonation se fait sur un maximum et avec une attaque franche (1 ms), on entendra un bruit dû à la

troncature du signal, contrairement au cas de début de phonation sur un zéro du signal. Nous n'avons pas encore implémenté cette règle.

L'implémentation dans Max/MSP est représentée à la figure 2.3. Le modèle de seuil est désactivable.

2.2.3 Mécanismes laryngés et tessiture

Le passage du mécanisme laryngé M_1 au mécanisme laryngé M_2 (mécanisme au sens de Roubeau, Henrich et Castellengo [RHC09]) dans la version précédente du synthétiseur de Le Beux [LB09] présentait une importante discontinuité dans le son perçu. Les chanteurs lyriques travaillent cette transition. Et ceux qui maitrisent la voix mixte (tessiture située à cheval sur deux mécanismes) sont capables de produire un timbre de M_1 tout en utilisant le mécanisme M_2, et un timbre de M_2 tout en utilisant le mécanisme M_1 quand ils chantent dans la zone de recouvrement des deux mécanismes. Dans notre synthétiseur, pour éviter de modéliser ces changements de timbre et d'avoir des discontinuités, nous avons fixé le mécanisme pour chacune des tessitures disponibles. Pour la tessiture Sol#1-Sol4, nous y associons le mécanisme M_1 et à partir de la tessiture Sol#2-Sol5, nous y associons le mécanisme M_2. Ainsi, il n'y a pas de transition de mécanisme possible à l'intérieur de chaque tessiture.

Le synthétiseur disposait d'un phonétogramme pour chaque mécanisme, construit à partir de voix réelles [LB09]. Dans notre synthétiseur, la tessiture est plus grande qu'en voix naturelle. Par conséquent, le niveau sonore devient trop faible pour les petits F_0, et trop grand pour les grands F_0, ce qui est incompatible avec l'utilisation de plusieurs tessitures et du même phonétogramme : la tessiture basse devient inaudible comparée à la tessiture la plus haute. La solution idéale serait d'avoir un phonétogramme pour chacune des tessitures, ce que nous n'avons pas entrepris. Nous avons plutôt réduit l'intervalle dans lequel la pente spectrale de l'ODGD évoluait, afin d'obtenir une voix suffisamment forte dans les basses fréquences. Le phonétogramme du mécanisme M_1 est donné à la figure 2.5. Le phonétogramme est désactivable pour obtenir une plus grande liberté dans le niveau sonore selon la fréquence fondamentale (mais on perd du naturel).

2.3 Modélisation des résonances du conduit vocal

Cette partie concerne la modélisation des résonances et anti-résonnances du conduit vocal situées jusqu'à 5500 Hz. La figure 2.6 montre comment cette modélisation a été implémentée dans Max/MSP. La partie « anti-résonance nasale » sera traitée dans le chapitre 3 car on ne fait pas intervenir de voyelles nasales. Les objets Max en vert correspondent à une application pédagogique qui sera traitée dans l'annexe B.

2.3.1 Valeurs des formants des voyelles cibles

Comme expliqué dans la partie 1.7.2, les premières résonances du conduit vocal sont modélisées par des filtres passe-bandes du second ordre, à 2 pôles et 1 zéro. Elles correspondent ici au contenu spectral jusqu'à 4000 Hz.

A partir de valeurs fournies par [GR94], [Lav], de mesures de l'équipe LAM de l'IJLRA, et de leur ajustement « manuel », nous avons abouti aux valeurs des filtres formantiques correspondant à notre voix de ténor de référence. Elles sont données dans le tableau 2.1 (fréquences centrales Fi, bandes-passantes Bi, et amplitudes Ai du formant i, avec $i \in [1, 5]$).

On a choisi des voyelles du français, /i,e,a,o,u/ comprenant les trois voyelles cardinales dans le plan des fréquences centrales F_1 et F_2 des deux premiers formants. Le triangle que

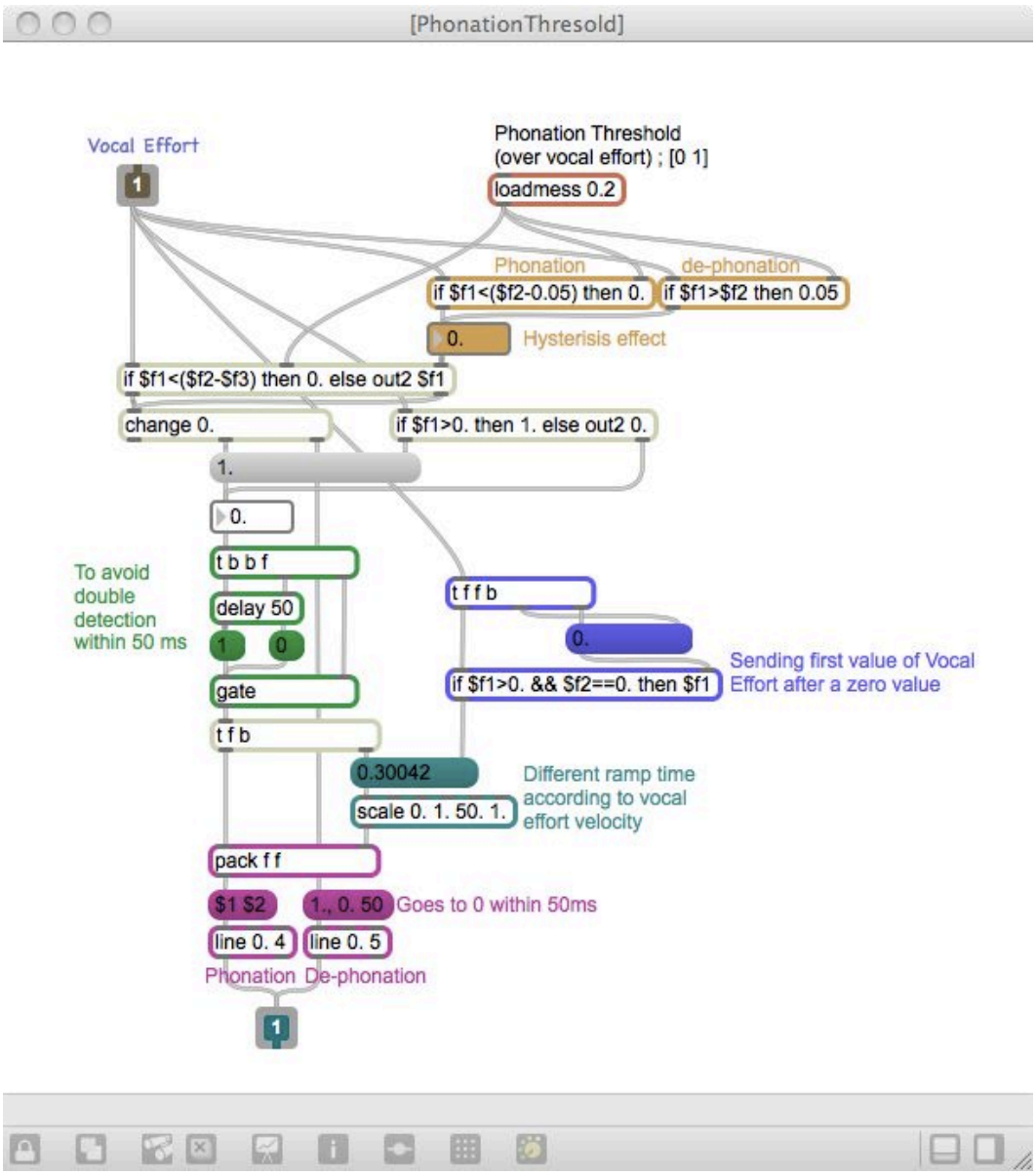


FIGURE 2.4 – *Implémentation en Max/MSP du seuil de phonation*

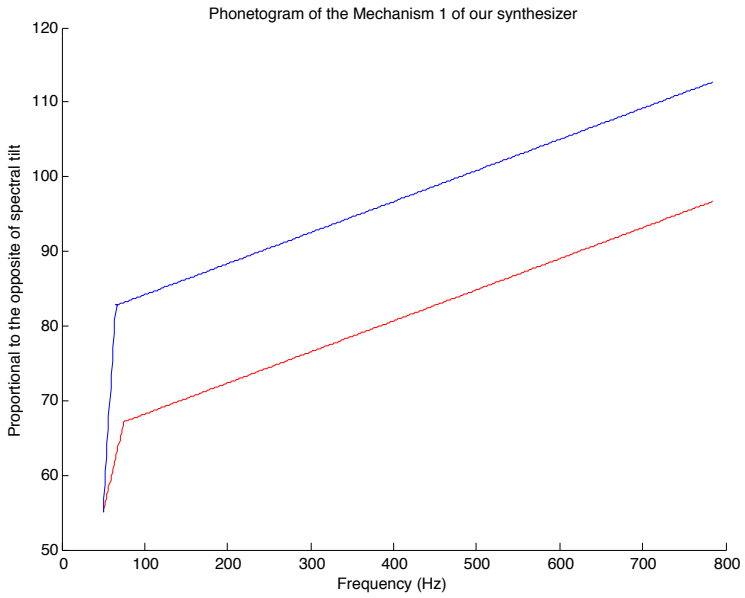


Figure 2.5 – *Phonétogramme du mécanisme M_1*

forment ces trois voyelles dans l'espace $F_1 - F_2$ comprend toutes les voyelles du français. Ainsi, l'interpolation de ces voyelles permet en première approximation de toutes les obtenir, c'est ce que nous expliciterons dans la partie 2.7.3.

L'amplitude du premier formant de la voyelle /a/, a priori le plus sonore parmi tous les formants des différentes voyelles, est pris comme référence dans le calcul de l'amplitude des formants en dB, le portant à 0 dB. On verra dans la partie 2.5 comment extrapoler ces valeurs pour obtenir les formants correspondant à d'autres types de voix.

	Fréquences Fi **(Hz)**	**Bande-passante** Bi **(Hz)**	**Amplitude** Ai **(dB)**
/i/	215. 1900. 2630. 3170. 3600.	10. 18. 20. 30. 20.	-13. -23. -2. 1. -40.
/e/	460. 1550. 2570. 2980. 3600.	10. 15. 20. 30. 20.	-1. -3. -2. -2. -5.
/a/	700. 1200. 2500. 2800. 3600.	13. 13. 40. 60. 40.	0. 0. -5. -7. -30.
/o/	440. 880. 2160. 2860. 3600.	10. 12. 20. 30. 20.	-6. -1. -18. -10. -37.
/u/	290. 750. 2300. 3080. 3600.	10. 10. 20. 30. 20.	-12. -9 -14. -11. -11.

Table 2.1 – *Valeurs des formants de la voix de ténor synthétisée pour plusieurs voyelles du français*

2.3.2 A propos du formant du chanteur

Le style de chant lyrique est notamment caractérisé par la présence du formant du chanteur. Le formant du chanteur est le rapprochement fréquentiel des troisième et cinquième formants autour du quatrième formant, développant ainsi dans une bande fréquentielle du spectre de la voyelle une concentration d'énergie supérieure à celle qu'on trouve dans la voix parlée. Le niveau de cette résonance, sa position et le degré de rapprochement des formants

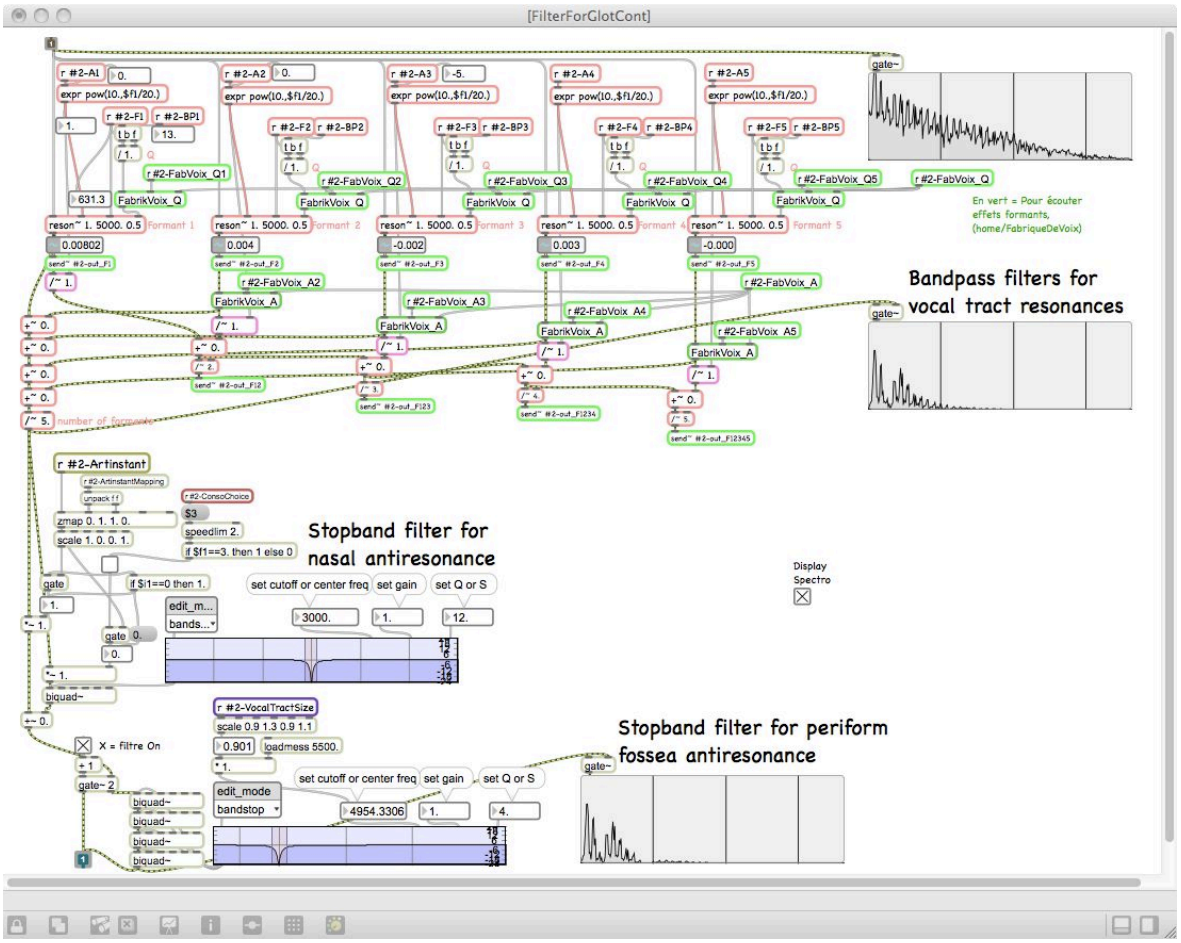


FIGURE 2.6 – *Capture d'écran du sous-patch Max/MSP du Cantor Digitalis correspondant à la modélisation du conduit vocal*

sont dépendants de la voyelle, de l'effort vocal et du type de chanteur [Sun01]. Cette technique de chant, qui donne un timbre particulier, a aussi un rôle pratique en participant au décalage de l'énergie de fréquence élevée vers un domaine plus audible, augmentant ainsi l'énergie perçue de la voix.

Ici, aucun modèle spécifique à cette résonance n'est formulé. Sa présence est reliée aux valeurs formantiques issues de l'analyse et de réglages manuels, puisqu'elle est comprise dans un intervalle fréquentiel correspondant aux formants. Le formant du chanteur peut donc s'entendre et s'observer dans notre voix de synthèse, sans être explicitement modélisé.

Par rapport à la précédente version du synthétiseur qui ne disposait que de quatre formants, nous avons rajouté un cinquième filtre formantique, de même type que les quatre autres.

Le cinquième formant contribue à la perception du formant du chanteur, en ayant une fréquence centrale plus proche du formant 4 qu'en voix parlée. Cet effet est d'autant plus perceptible que l'effort vocal est important car celui-ci fait diminuer la pente spectrale et ainsi fait apparaître plus d'énergie dans les hautes fréquences.

Tout comme le quatrième formant, sa valeur n'est pas rendue dépendante de la configuration vocalique en première approximation. On considère généralement les trois premiers formants pour caractériser les voyelles en français. Ici, on a gardé des valeurs légèrement différentes selon la configuration vocalique pour le formant 4, mais identiques pour le formant 5, étant donné que celui-là est plus difficile à mesurer. Il s'agit plus de rajouter de l'énergie dans la bande de fréquence $3500-4500Hz$ plutôt que de donner un sens phonologique à cette résonance.

2.3.3 Anti-résonance du sinus piriforme

L'hypopharynx (figure 2.7) se situe à proximité du larynx et est composé de trois cavités, le vestibule (cavité laryngeale) et les sinus piriformes bilatéraux. Elles sont responsables d'une résonance à 3-3.5 kHz pour le vestibule [KHT05] et d'une anti-résonance à 4-5 kHz pour les sinus piriformes [DH97].

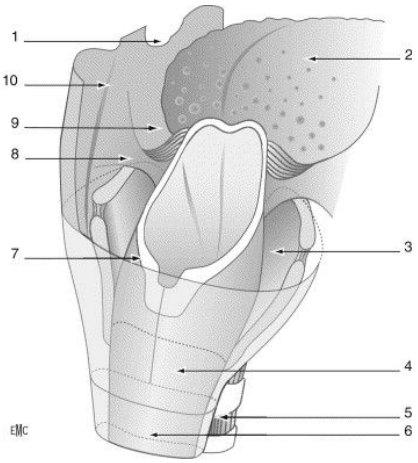
Dans notre modèle, la résonance hypopharyngeale est indirectement prise en compte dans notre base de formants puiqu'elle est comprise dans l'intervalle fréquentiel $[F_0 - 4000]$ Hz correspondant à nos filtres formantiques. L'anti-résonance est quant à elle modélisée par l'ajout d'un filtrage coupe-bande du signal à la sortie des filtres formantiques. On utilise pour cela 4 filtres identiques en série de type coupe-bande *biquad* (2 zéros et 2 pôles). Les coefficients du filtre sont ajustés de façon à avoir un gain valant 1 dans la bande-passante, un facteur de qualité de 2.5 et une fréquence centrale autour de 4500 Hz dépendant du chanteur (voir partie 2.5.1).

La forme des sinus piriformes est relativement invariante avec les voyelles d'un même locuteur [KHT05], mais les fréquences d'anti-résonance peuvent changer de l'ordre de 10 % [HKT+04]. On décide pour simplifier de ne pas faire varier la valeur du filtre modélisant l'anti-résonance pour les différentes configurations vocaliques et pour un même chanteur.

2.3.4 Exemple de comparaison entre une voyelle /a/ réelle et de synthèse

Voir fichiers audios / vidéos 1

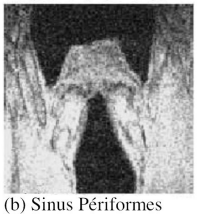
La figure 2.8 permet de comparer un /a/ naturel chanté de ténor issu de la base de données RWC [GHNO03], avec un /a/ synthétisé avec les mêmes caractéristiques.



(a) *Vue postérieure et latérale de l'hypopharynx (d'après Lefebvre et Chevalier [LC04]) 1. Amygdale ; 2. base de la langue ; 3. sinus piriforme ; 4. région rétro-crico-aryténoïdienne ; 5. trachée ; 6. oesophage ; 7. limite supérieure horizontale de l'hypopharynx ; 8. repli pharyngoépiglottique ; 9. fosse sous-amygdalienne ; 10. pilier postérieur*

(b) *Cavité du larynx et fosses piriformes dans l'hypopharynx (d'après Honda et al. [HKT+ 04])*

(a) Cavité du Larynx



(b) Sinus Périformes

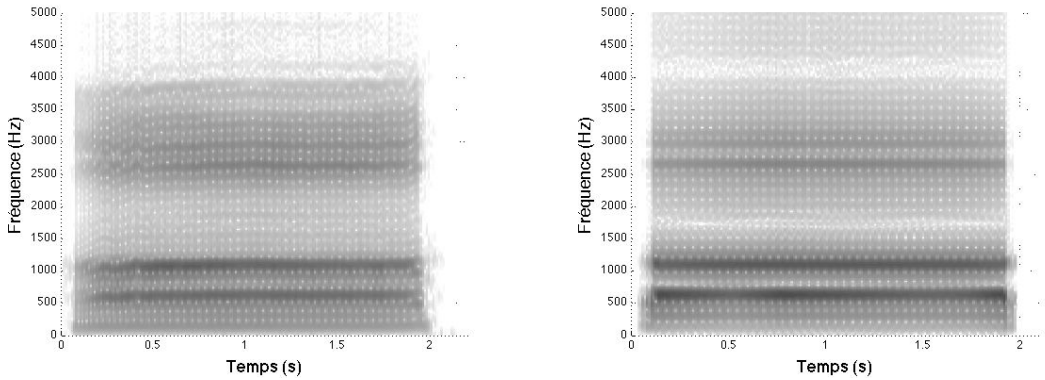
FIGURE 2.7 – *Localisation des sinus piriformes dans l'appareil vocal*

L'allure globale du spectre est proche du /a/ naturel, mais elle amène quelques commentaires. Le /a/ de synthèse présente un spectre plus contrasté, avec des pics de formants d'amplitude un peu plus importante que naturellement, ainsi que moins d'énergie entre les formants. La forme des résonances présente une bande-passante plus grande en voix naturelle sauf en-dessous du troisième formant et au-dessus du deuxième formant où l'énergie s'affaisse plus rapidement avec la voix naturelle. D'autre part, on observe moins d'énergie dans les hautes fréquences avec le /a/ naturel au delà de l'anti-résonance du sinus piriforme. Enfin, l'évolution temporelle diffère, mais la voix de synthèse a été produite avec un effort vocal constant, un geste de contrôle manuel aurait sans doute permis une évolution de l'énergie totale plus proche du cas naturel.

2.4 Dépendances sources-filtres

Le modèle source-filtre de l'appareil vocal postule l'indépendance de la source glottique et du conduit vocal. Or, un certain nombre de dépendances intervient qu'on peut classer suivant leur origine :

– *fonctionnelle* : la source glottique et le conduit vocal ne sont pas isolés l'un de l'autre, une partie de l'onde de débit glottique retourne dans la glotte après réflexions dans le conduit vocal, interférant ainsi avec sa vibration. On a donc un couplage entre la source glottique et le conduit vocal. Dans une certaine mesure, notre modèle source filtre prend en compte ce couplage car toutes les analyses du signal de pression glottique et des résonances sont réalisées sur un appareil vocal entier et non sur la source glottique isolée. L'analyse de la source glottique a été effectuée à partir d'enregistrements électroglottographiques. Les valeurs des formants sont issues de l'analyse acoustique du



(a) */a/ naturel* (b) */a/ synthétisé*

Figure 2.8 – *Comparaison d'un /a/ synthétique et naturel, à 155.8Hz (Ré♯). Voir fichiers audios / vidéos 1*

son rayonné, donc elles prennent également en compte le couplage. On ne cherche alors pas à modéliser cette interaction.

– *liée à l'apprentissage* : afin d'optimiser la puissance perçue de notre voix, nous avons appris au cours de l'enfance à modifier notre conduit vocal en fonction de la fréquence fondamentale de la source glottique ; de la même manière, les chanteurs apprennent des techniques particulières au chant pour optimiser le fonctionnement de leur appareil vocal.

Une dépendance d'une toute autre nature est programmée, visant à diminuer des artefacts de notre modèle, ce que nous allons présenter.

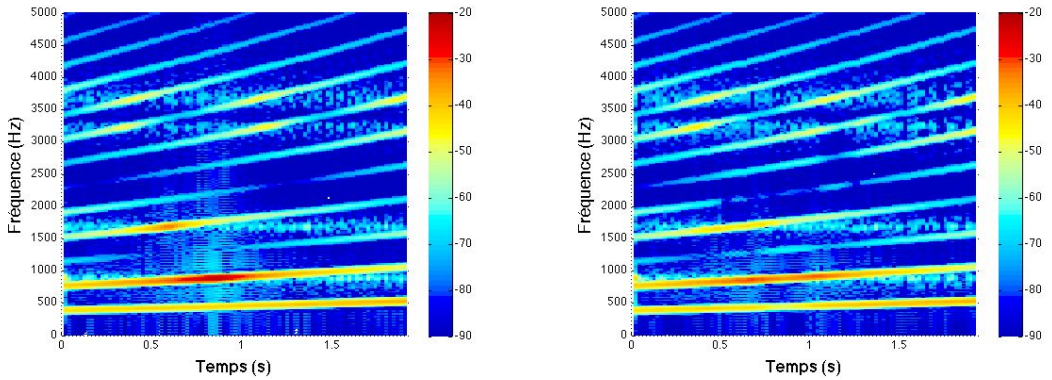
2.4.1 Atténuation des résonances en fonction de F_0

Voir fichiers audios / vidéos 2

Notre système créait des résonances non désirées à certaines hauteurs mélodiques, résonances bien trop importantes perceptivement pour être naturelles. La cause provient de l'utilisation de filtres résonnants présentant un pic fréquentiel étroit.

Pour y remédier, les amplitudes des filtres formantiques sont diminuées automatiquement et progressivement quand la fréquence fondamentale F_0 et/ou ses premières harmoniques se rapprochent de la fréquence centrale du filtre formantique concerné (parmi les cinq). On prend en compte les huit premières harmoniques de F_0 pour des raisons empiriques. Au delà, on considère les effets des résonances non gênants.

L'intervalle fréquentiel autour de la fréquence centrale du filtre formantique où le changement d'amplitude a lieu est rendu dépendant de F_0 : plus F_0 est élevée, plus l'intervalle est agrandi, afin d'avoir un rapport à peu près constant entre intervalle et F_0 quelle que soit F_0. La figure 2.9 illustre dans le domaine spectral la différence avec et sans dépendance source-filtre. Les deux sons ont été produits en faisant varier F_0 de Fa♯3 à Do3 sur 2 secondes. On observe sur la figure 2.9 (a) que lorsque l'harmonique 2 (respectivement 4) se rapproche de la fréquence centrale du premier formant (respectivement du deuxième formant), l'intensité de ces harmoniques augmente trop fortement par rapport à des productions naturelles. Ces résonances sont toujours présentes avec l'atténuation en (b), comme en voix naturelle, mais avec une amplitude moindre.



(a) *Sans atténuation automatique* (b) *Avec atténuation automatique*

FIGURE 2.9 – *Spectrogramme d'une voyelle synthétisée, où F_0 évolue avec le temps, (a) sans ou (b) avec l'atténuation automatique des résonances. Voir fichiers audios / vidéos 2*

L'amplitude des filtres formantiques des voyelles a été ajustée de manière à ne pas avoir de saturation et d'avoir une certaine cohérence de niveau sonore entre les différentes voyelles.

Joliveau *et al.* [JSW04] et Henrich *et al.* [HSW11] ont montré que les chanteurs modifient la valeur des premières fréquences de résonance de leur conduit vocal en fonction des harmoniques de la fréquence fondamentale. Ici, on adapte également les filtres formantiques à la fréquence fondamentale et ses harmoniques, mais avec un effet inverse, c'est à dire que nous diminuons l'effet de résonance avec la fréquence fondamentale et ses harmoniques. On peut voir cette atténuation comme un modèle d'amortissement non linéaire de l'appareil vocal qui atténue les résonances au-dessus d'un seuil et ce d'autant plus qu'elles sont importantes. On verra dans la partie 2.4.3 comment modéliser les effets décrits par Henrich *et al.* [HSW11].

Enfin, nous avons utilisé un limiteur pour compléter cette atténuation, en diminuant ainsi les écarts de volume sonore en fonction de F_0. Cependant, celui-ci n'est pas une dépendance source-filtre à proprement parler puisqu'il agit seulement sur le signal de sortie.

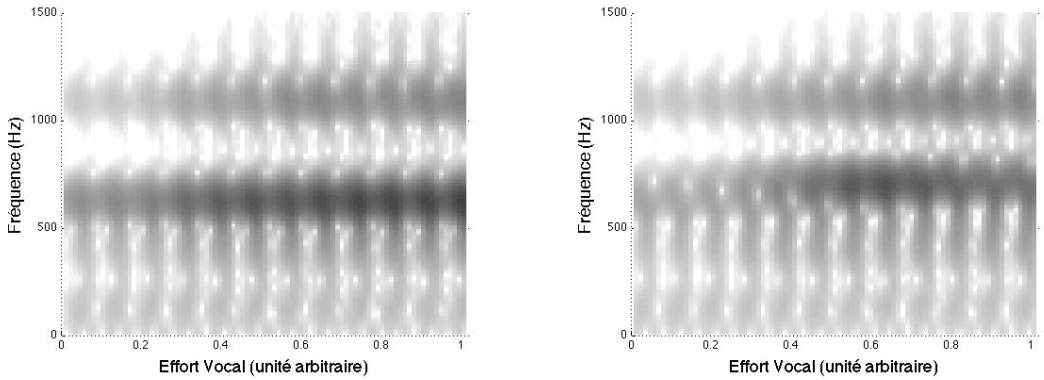
2.4.2 Fréquence du premier formant et effort vocal

Voir fichiers audios / vidéos 3

Dans cette section, on explique la modélisation de la dépendance entre l'effort vocal et la fréquence F_1 du premier formant.

D'après Lienard et Di Benedetto [LDB99], l'augmentation de l'effort vocal est liée à l'élévation de F_0 et de F_1. F_0 est un paramètre qu'on veut pouvoir contrôler finement donc on n'intervient pas dessus automatiquement. F_1 augmente en moyenne de $3,5Hz/dB$ pour les voyelles orales du français isolées dans l'expérience de [LDB99].

En supposant qu'on peut extrapoler cette information à la voix chantée et que le signal aux lèvres varie d'environ 30 dB entre effort vocal faible et important dans l'expérience de [LDB99], on arrive donc à un changement maximum de $3.5 \times 30 \approx 100Hz$ pour la fréquence du premier formant entre deux efforts vocaux faible et fort. Le premier formant se situant dans la gamme $200 - 700Hz$ (de moyenne $400Hz$) selon la voyelle, cela correspond à faire varier linéairement F_1 jusqu'à un facteur de $\pm 25\%$ entre ses valeurs extrêmes autour de sa valeur de référence. Dans la pratique, la variation à été ramenée à 15% pour rester acceptable perceptivement. La cause de cette différence peut résider dans les conditions expérimentales



(a) *Sans dépendance automatique entre F_1 et F_0* (b) *Avec atténuation automatique*

FIGURE 2.10 – *Spectrogramme de la voyelle /a/ de synthèse, où l'effort vocal augmente suivant l'axe des abscisses, (a) sans ou (b) avec la dépendance entre la fréquence centrale du premier formant et de F_0. Voir fichiers audios / vidéos 3*

de l'expérience de [LDB99], en particulier le fait que F_0 augmentait avec l'effort vocal, ce qui a pu induire une augmentation de F_1. D'autre part, les sujets de l'expérience devaient produire des voyelles isolées dans un contexte non musical.

Dans notre modèle avec la voix de ténor à $F_0 = 200Hz$, il faut faire passer le paramètre d'effort vocal de $VE = 1$ (valeur maximale) à $VE = 0.6$ pour avoir un changement d'environ 30 dB. La valeur de référence $F1$ de la base de données correspondant à celle du ténor pour un effort vocal moyen ($VE = 0.8$), on cherche donc la relation entre F_1 et VE tels que :

$$\begin{cases} F_1(VE = 0.8) = & F1 \\ F_1(VE = 0.6) = & 0.925F1 \\ F_1(VE = 1.0) = & 1.075F1 \end{cases} \tag{2.7}$$

soit la relation suivante entre F_1 et VE :

$$F_1(VE) = \begin{cases} (0.7 + 0.375 \cdot VE) \cdot F1 & \text{si } VE \geq 0.6 \\ 0.925 \cdot F1 & \text{si } VE < 0.6 \end{cases} \tag{2.8}$$

L'implémentation de cette dépendance est illustrée par la figure 2.10 qui montre l'augmentation de F_1 avec l'effort vocal pour la voyelle /a/. Il n'a pas été reporté par [LDB99] de variation de la fréquence centrale des formants supérieurs avec l'effort vocal, d'où notre restriction à F_1.

2.4.3 Adaptation des deux premières résonances à F_0

Voir fichiers audios / vidéos 4

Cette section explique l'implémentation de la dépendance entre les résonances du conduit vocal et la fréquence fondamentale F_0 de vibration de la source glottique.

Le premier type de dépendance est celui des résonances R_1 et R_2 (correspondant aux deux premiers formants) et la fréquence fondamentale F_0. Cette dépendance est la conséquence d'un apprentissage des chanteurs pour optimiser la puissance de la voix par une adaptation des résonances du conduit vocal afin que l'énergie issue de la source glottique y soit utilisée de façon plus efficace. Ainsi, Joliveau *et al.* [JSW04] et Henrich *et al.* [HSW11] ont montré que les chanteurs adaptent leur deux premières résonances en fonction de la position fréquentielle de F_0 et de ses harmoniques. Plus précisément, leur travail a porté sur des chanteuses sopranos, avec les voyelles /ɑ,ɔ,u,ɜ/ en effort vocal faible. Ils ont établi qu'elles ajustaient leur première résonance R_1 avec F_0 et que la plupart d'entre elles ajustaient R_2 avec la deuxième harmonique de F_0 ($2F_0$). Garnier *et al.* [GHWS10] montrent l'implication des lèvres dans cette adaptation.

Dans le programme CHANT [RPB84], Rodet, Potard et Barrière ont contraint le premier et le deuxième formant de leur voix de synthèse à suivre respectivement le premier et deuxième harmonique, au-dessus d'un certain seuil sur F_0.

Dans le Cantor Digitalis, nous avons implémenté la modification des fréquences centrales des filtres formantiques en fonction de F_0 de la façon suivante :

$$F_1(F_0) = \begin{cases} F1 & \text{si } F_0 \leq F1 - 50 \text{ Hz} \\ F_0 + 50Hz & \text{sinon} \end{cases} \tag{2.9}$$

$$F_2(F_0) = \begin{cases} F2 & \text{si } 2F_0 \leq F2 - 50 \text{ Hz} \\ 2F_0 + 50Hz & \text{sinon} \end{cases} \tag{2.10}$$

où F1 et F2 sont les fréquences centrales des formants de référence.

Cet effet est illustré par deux spectrogrammes à la figure 2.11 tracés en fonction de F_0, l'un sans et l'autre avec la dépendance en F_0. L'exemple a été réalisé sur la voyelle /u/ afin que la dépendance intervienne pour un F_0 peu élevé, /u/ étant caractérisée par deux premiers formants de fréquences centrales basses. On y voit bien notamment la fréquence du premier formant F_1 rester constante jusqu'à ce que F_0 atteigne une valeur proche, où F_1 suit F_0.

En réalité, le suivi des résonances à F_0 décroche quand F_0 devient trop élevé, comme on peut l'observer sur les mesures de [JSW04] reproduites à la figure 2.12.

D'après l'expérience de [JSW04], les résonances supérieures $R3$, $R4$ et $R5$ ne semblent pas autant modifiées par la variation de F_0 que $R1$ et $R2$, mais on observe quand même environ de 5 à 10% de variation pour $R3$, $R4$ et $R5$ entre 200 et $1000Hz$. Les auteurs associent cette évolution au probable changement de forme du larynx dans les F_0 élevées.

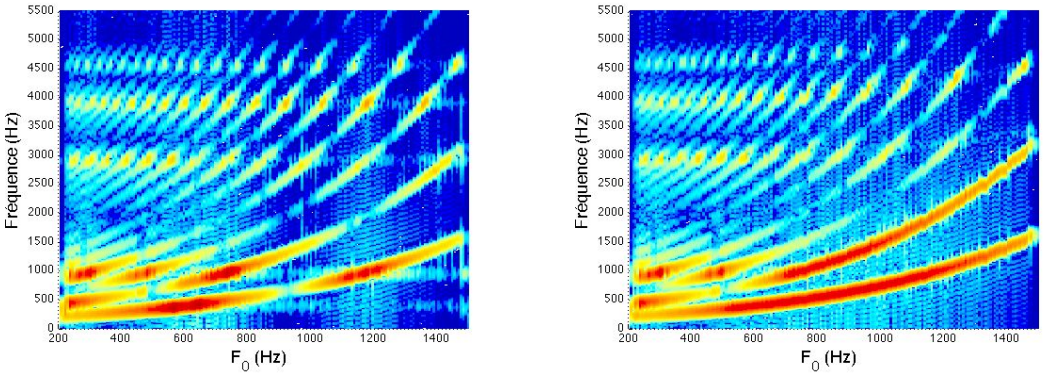
On applique ce résultat en multipliant la fréquence centrale des cinq premiers formants (formants 1 et 2 compris) à un facteur K dépendant de F_0 :

$$K(F_0) = 1.25 \cdot 10^{-4} F_0 + 0.975 \tag{2.11}$$

de façon à avoir les valeurs de référence de la base de données des formants du synthétiseur à $F_0 = 200Hz$ et les observations de [JSW04] :

$$\begin{cases} K(200Hz) & = 1 \\ K(1000Hz) & = 1.1 \end{cases} \tag{2.12}$$

On peut observer cet effet sur la figure précédente 2.11.



(a) *Sans dépendance* (b) *Avec dépendances*

FIGURE 2.11 – *Spectrogrammes de la voyelle /a/ de synthèse, où F_0 augmente avec le temps, (a) sans ou (b) avec les dépendances entre les fréquences centrales des formants et de F_0. Voir fichiers audios / vidéos 4*

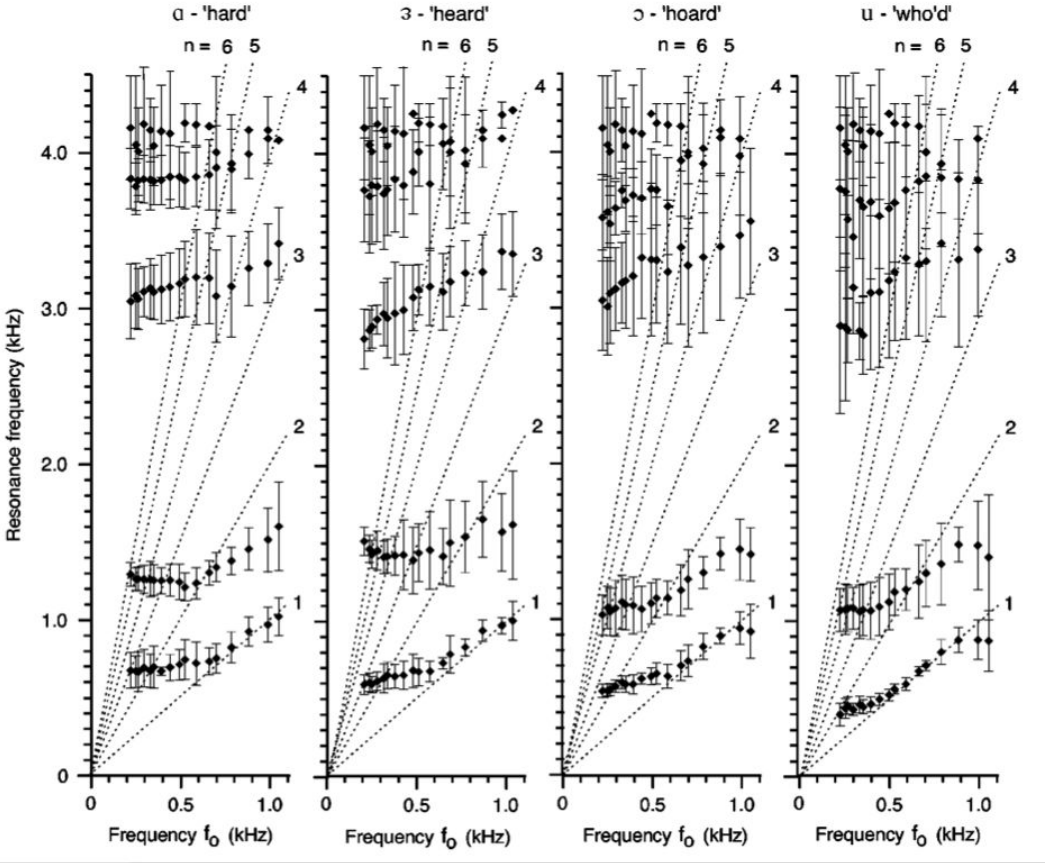


FIGURE 2.12 – *Les résonances du conduit vocal pour quatre voyelles naturelles de soprano, en fonction de F_0. Les lignes en pointillées indique la relation entre les résonances et les harmoniques nF_0. D'après Joliveau et al. [JSW04].*

2.5 Personnalisation des voix

L'utilisation conjointe de différentes longueurs de conduit vocal et de différentes tessitures et qualités vocales permet de développer une gamme de différents chanteurs. Ainsi, on met à disposition une présélection de voix de plusieurs chanteurs de type classique (basse, ténor, alto et soprano), d'enfant, de bébé, et de voix de « monstres » en poussant la valeur de certains paramètres à leur limite naturelle. Chacun de ces paramètres est configurable indépendamment pour construire de nouvelles voix.

2.5.1 Taille du conduit vocal et tessiture

Voir fichiers audios / vidéos 5

Nous faisons l'hypothèse qu'à partir de l'estimation des formants d'une seule voix réelle, nous pouvons obtenir autant d'ensembles de formants cohérents entre eux que l'on veut de voix différentes. En effet, les résonances du conduit vocal sont liées à sa taille. Plus le conduit vocal est grand, plus les fréquences de résonance seront petites et inversement.

A partir de nos valeurs formantiques de référence données dans le tableau 2.1 de la partie 2.3.1 et correspondant à une voix de ténor, on applique un facteur aux fréquences centrales des filtres formantiques pour changer la taille du conduit vocal. Ce facteur sera donc plus grand que 1 pour des voix type basse, et plus petit que 1 pour des voix de type alto ou soprano. Par ce raisonnement, on peut passer d'une voix de bébé avec un conduit vocal très petit à une voix de « géant » avec un conduit vocal très grand. Les valeurs utilisées de ce facteur sont données dans le tableau 2.3.

En changeant les valeurs des fréquences des formants, les voyelles ne sont pas modifiées. En effet, la perception d'une voyelle est plus liée au rapport des fréquences des formants qu'à leurs valeurs absolues. C'est ce qui permet à un enfant et à une grande personne de pouvoir prononcer la même voyelle tout en ayant un conduit vocal de taille très différente. Ainsi, dans une certaine mesure, utiliser un facteur identique pour tous les formants est une bonne approximation de la taille du conduit vocal.

Cependant, la différence de taille entre un enfant et un adulte n'est pas seulement un changement global de taille. La transformation n'est pas isomorphique, en particulier après la puberté chez les hommes qui fait descendre le larynx, augmentant ainsi le conduit vocal plus dans sa longueur que dans sa largeur. On ne prend pas en compte ici ces changements de forme, mais nous simplifions cette transformation par un agrandissement global en agissant sur les fréquences centrales des filtres formantiques (bande-passante et amplitude non modifiées).

Le conduit vocal n'est pas le seul à changer de taille. La source glottique aussi est modifiée. Il faut alors pour chacune des voix y associer la bonne tessiture, c'est à dire la gamme de fréquence que la source peut parcourir. Bien que la tessiture des chanteurs soit environ de 2 octaves, nous avons choisi de l'élargir à 3 octaves pour profiter de la liberté que nous donne la modélisation numérique (tableau 2.2).

D'après Kitamura, Honda et Takemoto [KHT05], la forme des sinus piriformes varie avec le locuteur, sans en préciser le lien avec la taille du conduit vocal. Il est probable que la longueur du conduit vocal soit liée à la taille des sinus piriformes. Nous avons décidé de lier la fréquence du filtre coupe-bande modélisant l'anti-résonance du sinus piriforme à la taille

	Tessiture naturelle moyenne	Tessiture dans le synthétiseur
Basse	Mi2-Mi4	Sol♯1-Sol4
Ténor	Do3-Si4	Sol♯1-Sol4
Alto	Fa3-Mi5	Sol♯2-Sol5
Soprano	Si3-Do6	Sol♯3-Sol6

Table 2.2 – *Tessiture des chanteurs naturels et synthétiques (La3=440 Hz). Voir fichiers audios / vidéos 5*

du conduit vocal. De la même manière que les formants, nous exposons la fréquence centrale du filtre au facteur de taille du conduit vocal.

2.5.2 Qualité vocale : de la voix soufflée aux voix monstrueuses

Voir fichiers audios / vidéos 6

On dispose de trois paramètres de haut-niveau pour le contrôle de la qualité vocale (explicités dans la partie 1.7) :

- la quantité de souffle moyenne
- la tension vocale moyenne (reliée aux paramètres temporels de l'ODG : valeur du coefficient d'asymétrie α_m et plage de valeurs de O_q)
- les apériodicités période à période de la vibration de la source (jitter/shimmer)

L'avantage d'un modèle de type signal est qu'on peut extrapoler facilement les valeurs des paramètres. On parvient à obtenir des « rugissements de fauve » crédibles avec une voix bruitée, apériodique et tendue, ajoutée à une longueur de conduit vocal très grande et une tessiture très basse.

La cohérence des paramètres de source n'est plus forcément présente quand on crée des voix de type monstre. Par exemple, on peut avoir à la fois une tension vocale non nulle et un souffle non nul alors que chacun de ces deux paramètres sont, dans une voix naturelle, la continuité de l'un et de l'autre. En effet, le paramètre « souffle » correspond au relâchement des plis vocaux de façon qu'un bruit turbulent apparaisse à leur niveau, alors que le paramètre « tension » correspond à la tension des plis vocaux. On ne peut donc alors pas concevoir en voix naturelle d'avoir à la fois une voix soufflée et une voix tendue, ou du moins on ne qualifiera pas la voix de soufflée si un bruit apparaît. Ce bruit sera a priori d'une autre nature.

2.5.3 Résumé des paramètres des différentes voix

Le tableau 2.3 récapitule toutes les caractéristiques des voix préconfigurées dans notre synthétiseur. Les formants des voyelles de référence sont calculés à partir de ceux de la voix de ténor donnée plus haut au tableau 2.1 et modifiés en fonction du facteur sur la longueur du conduit vocal, ainsi que des différentes dépendances sources-filtres décrites dans la partie 2.4. On dispose ainsi d'un ensemble de chanteurs (basse, ténor, alto, alto soufflé, soprano, soprano soufflé), de voix d'enfants, et quelques monstres vocaux. L'utilité d'une différenciation des voix par leur utilisation en chorale est présenté en annexe D.

	Tessiture	**Facteur de taille du conduit vocal**	**Apério-dicités [0 1]**	**Souffle [0 1]**	**Tension**	
					O_q	α_m
Basse	Sol♯1-Sol4	0.9	0.	0.	[0.97-0.68]	0.75
Ténor	Sol♯1-Sol4	1.05	0.	0.	[0.97-0.68]	0.75
Alto	Sol♯2-Sol5	1.1	0.	0.	[0.84-0.62]	0.66
Alto 2	Sol♯2-Sol5	1.075	0.	0.024	[0.84-0.62]	0.66
Soprano	Sol♯3-Sol6	1.25	0.	0.	[0.84-0.62]	0.66
Soprano 2	Sol♯3-Sol6	1.3	0.	0.016	[0.84-0.62]	0.66
Enfant 1	Sol♯4-Sol7	1.5	0.	0.	[0.84-0.62]	0.66
Enfant 2	Sol♯3-Sol6	1.25	0.	0.	[0.84-0.62]	0.66
Bébé	Sol♯4-Sol7	2.	0.	0.	[0.84-0.62]	0.66
Monstre 1	Sol♯4-Sol7	1.	0.12	0.08	[0.70-0.40]	0.73
Monstre 2	Sol♯1-Sol4	0.5	1.	1.	[0.96-0.68]	0.75
Monstre 3	Sol♯1-Sol2	0.7	0.80	1.	[0.55-0.22]	0.92

TABLE 2.3 – *Personnalisation des voix de chanteurs et monstres vocaux*

2.6 Perturbations multi-échelles de la source glottique

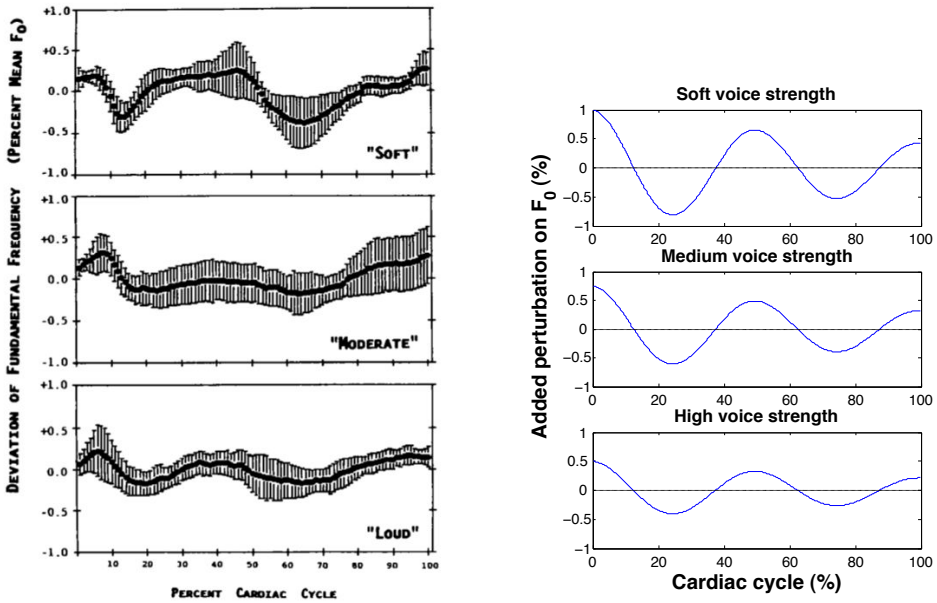
Des petites perturbations de la vibration des plis vocaux sont toujours présentes dans la voix humaines. Elles sont trop rapides pour être considérés comme des modulations d'amplitude ou d'intonation contrôlées, mais suffisamment lentes pour être perçues. Les causes peuvent être musculaires ou provenant de l'activité du corps à proximité du larynx.

Parmi les premiers synthétiseurs vocaux numériques à introduire des perturbations automatiques, on peut citer le MUSSE de Larsson [Lar77] et le programme CHANT de Rodet, Potard et Barrière [RPB84] qui agissent sur la fréquence fondamentale. Le premier ajoute sur les vibratos un bruit de basse fréquence appliqué à F_0. Le second fait intervenir un terme additif à F_0, composé de la somme de 3 bruits de périodes 0.05, 0.111 et 1.219 seconde, et dont l'amplitude, identique pour les 3 composantes, est fixée par l'utilisateur. La perturbation totale engendrée est de l'ordre de 1.1-3.7% pour le F_0 des voix de femme et de $2.0-5.7\%$ pour les voix d'hommes. Pour une étude plus poussée sur la synthèse de voix avec perturbation de fréquence fondamentale, voir les travaux de Fraj, Schoentgen et Grenez qui se placent dans une perspective médicale [FSG12].

Le jitter et le shimmer sont des perturbations respectivement de la période fondamentale et de l'amplitude de vibration des plis vocaux. Le modèle RT-CALM dispose en entrée d'un paramètre agissant sur l'amplitude du Jitter et du Shimmer, définit comme des bruits blancs calculés sur chaque période de l'ODGD. Il existe cependant des perturbations de la période fondamentale et de l'amplitude de la source glottique à des échelles temporelles plus grandes.

2.6.1 Perturbations cardiaques

Orlikoff a montré que la pulsation cardiaque interférait avec la source glottique, d'une manière d'autant plus perceptible que l'amplitude du signal acoustique est faible : à chaque pulsation, F_0 est déviée de l'ordre de 1% et l'amplitude de 3 à 14% à chaque cycle cardiaque [Orl90]. En première approximation, l'amortissement se fait sinusoïdalement à la fré-



(a) *Variabilité moyenne de F_0 pour trois niveaux sonores, d'après Orlikoff [Orl90]*

(b) *Variabilité de la perturbation selon différent SPL dans le synthétiseur*

Figure 2.13 – *Perturbations cardiaques de F_0 (a) issues de l'analyse et (b) modélisées dans le synthétiseur*

quence double de la pulsation cardiaque. Le terme perturbatif a_{pertub} (sans dimension) de l'amplitude A_0 et le terme perturbatif $f_{perturb}$ (sans dimension) de la fréquence fondamentale F_0, dépendants de l'amplitude du signal acoustique, sont modélisés de la façon suivante dans le *Cantor Digitalis* :

$$\begin{cases} a_0(t) = A_0[1 + a_{pertub}(t)] \\ f_0(t) = F_0[1 + f_{pertub}(t)] \end{cases} \tag{2.13}$$

$$\begin{cases} a_{pertub}(t) = b_{VE} \cdot \xi \cdot \exp(-\beta \cdot t) \cdot cos(4\pi \cdot f_{cardiaque} \cdot t) \\ f_{pertub}(t) = b'_{VE} \cdot \xi \cdot \exp(-\beta \cdot t) \cdot cos(4\pi \cdot f_{cardiaque} \cdot t) \end{cases} \tag{2.14}$$

où :	$t \in [0\ f^{-1}_{cardiaque}]$	Variable du temps pour chaque cycle cardiaque
	$\xi \in [0.5\ 1]$	Amplitude aléatoire uniforme
	$b_{VE}, b'_{VE} \in [0\ 1]$	Amplitudes proportionnelles, à une constante additive près, à l'opposé du niveau de pression acoustique (de façon à diminuer l'effet de la perturbation avec l'élévation de l'effort vocal), de telle sorte que la perturbation maximale de F_0 soit de 1% et de 3 à 4% pour l'amplitude A_0
	$f_{cardiaque}$	Pulsation cardiaque centrée autour de 70 pulsations par minute
	$\beta = 0.001ms^{-1}$	Facteur d'amortissement

On représente sur la figure 2.13 l'évolution du pourcentage de la perturbation additive de F_0 pour différents niveaux de pression acoustique durant un cycle cardiaque en (b), à comparer à la mesure sur voix réelle en (a).

2.6.2 Volume pulmonaire

Un autre caractère important pour rendre une voix naturelle sur une échelle de temps de plusieurs secondes est la respiration. Elle se traduit par l'extinction de la voix après l'utilisation d'un certain volume d'air $V_{inspiration}$, qu'on peut exprimer sur la durée de l'expiration par :

$$V_{expiration} = \int_{expiration} debit(t) \cdot dt \tag{2.15}$$

où $debit(t)$ est le débit d'air expiré à l'instant t de l'expiration.

Une fois que le volume d'air a été libéré, le niveau sonore décroit rapidement. Ainsi, plus l'effort vocal est important, plus le volume pulmonaire s'épuise rapidement. La durée d'extinction à la fin de l'expiration est mise en dépendance de l'effort vocal.

Dans notre modèle, aucun paramètre n'est relié directement au débit d'air absolu (l'ODG correspond au débit sans composante continue). On décide de l'associer au paramètre d'effort vocal (sans dimension) qui peut représenter approximativement une quantité d'air par unité de temps. Ainsi, en l'intégrant sur la durée de l'expiration, on obtient une grandeur qui croît avec le volume pulmonaire. Plus l'effort vocal est important et plus la durée nécessaire à l'expiration du volume pulmonaire sera petite. Au contraire un faible effort vocal augmentera la durée d'expiration du volume d'air. En pratique, on calcule l'effort vocal moyen toutes les 50 ms, et on additionne ces valeurs jusqu'à atteindre une valeur qu'on associe au volume d'air emmagasiné $V_{inspiration}$ lors de chaque inspiration. Quand cette valeur limite est atteinte, le paramètre d'effort vocal est amené rapidement à zéro afin d'arrêter la production vocale de notre modèle pour simuler l'expiration de tout le volume pulmonaire. Cette valeur limite est fixée empiriquement de manière à avoir une durée d'expiration de l'ordre de quelques secondes avec un effort vocal maximal.

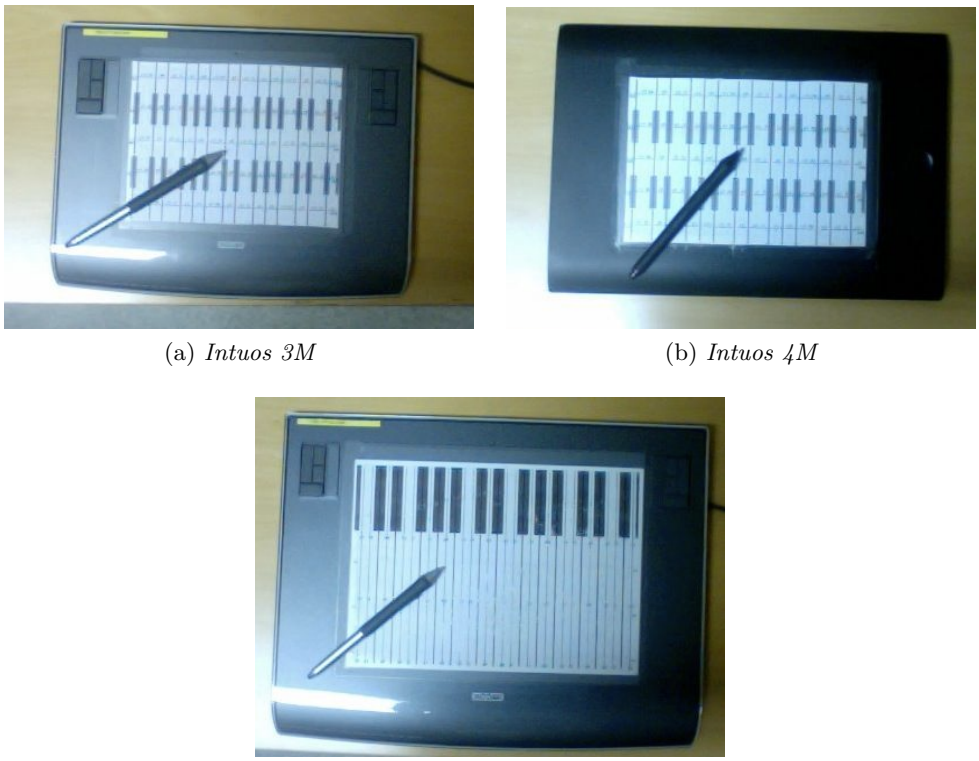
Toutefois, cette limitation de la voix dans le temps n'a pas été utilisée. On préfère garder la liberté de ne pas disposer d'une limitation dans la durée de phonation, difficile à anticiper. En voix naturelle, on ressent le volume d'air demeurant dans les poumons. La limitation du volume pulmonaire peut aussi être simulée par le geste de contrôle de l'utilisateur, simplement en évitant de jouer de trop longues séquences de notes continues.

2.7 Contrôle gestuel des voyelles chantées synthétiques

Notre but est de concevoir des instruments de synthèse vocale. Ainsi, au modèle de production présenté dans les parties ci-dessus s'ajoute un modèle de contrôle. Cette distinction est propre aux modèles numériques, vu que cette séparation est la plupart du temps presque impossible dans le cas des instruments acoustiques où l'interface gestuelle prend part directement à la génération du son (par sa vibration, son couplage ...) [WD04].

Il s'agit de réfléchir :

- au choix de l'interface pour capter les gestes ;
- à la transformation des données de l'interface en paramètres de contrôle ;
- au choix des paramètres du modèle de production qui doivent être contrôlés en temps-réel et ceux dont le contrôle gestuel n'est pas nécessaire (sélection avant de commencer à jouer) ;
- la manière de connecter les paramètres de contrôle à ceux de la production.



(a) *Intuos 3M* (b) *Intuos 4M*

(c) *Intuos 3L*

Figure 2.14 – *Les différents modèles de tablette graphique Wacom utilisées*

2.7.1 Une tablette graphique augmentée d'un clavier continu

Nous utilisons une tablette graphique Wacom Intuos (version 3, 4 ou 5, en format M ou L) comme interface de contrôle. Ces tablettes détectent notamment la position d'un stylet sur un plan et sa pression exercée sur la tablette (voir figures 2.14 pour l'apparence de ces différents modèles), et la position des doigts en contact avec la tablette pour les versions Intuos 5.

Plusieurs critères sont entrés en jeu dans la sélection de cette interface de contrôle. Tout d'abord, la latence propre à l'interface doit être minimum afin que la latence totale (du geste de contrôle au son émis en passant par le modèle de production) soit inférieure à 10-20 ms. En effet, nous envisageons un contrôle continu et instantané (perceptivement) des paramètres du synthétiseur, comme c'est le cas avec les articulateurs de l'appareil vocal. Autrement dit, nous envisageons des gestes de contrôle et non des gestes de sélection. La résolution temporelle des tablettes graphiques Intuos 3, 4 et 5 est de 5 ms (ordre de grandeur de la période d'échantillonnage généralement utilisée dans les instruments numériques gestuels) et la latence est celle de la transmission USB, négligeable dans notre cas. Cela permet d'avoir l'impression que le son et le geste sont unis par une relation de causalité directe comme avec les instruments mécaniques [Gen99].

Ensuite, l'interface doit posséder une résolution spatiale haute pour permettre la continuité du geste de contrôle en évitant d'entendre les pas de quantification. Les tablettes utilisées ont une résolution spatiale d'environ $0,25$ mm et de 2048 niveaux de pression.

Enfin, l'interface doit permettre un geste précis, reproductible et intuitif. La tablette graphique avec son stylet a été initialement conçue pour dessiner sur ordinateur. Le geste

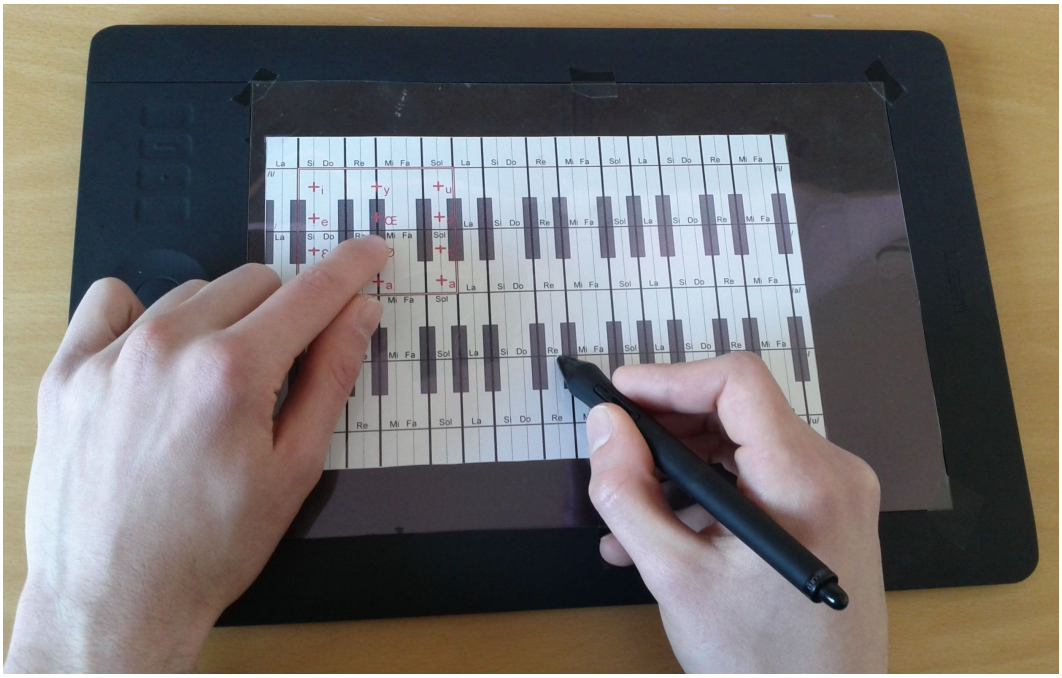


FIGURE 2.15 – *La tablette graphique munie de son calque comme interface de contrôle*

d'écriture est très précis, reproductible, et intuitif car maîtrisé depuis notre enfance.

Pour augmenter la facilité de jeu de la tablette graphique, un calque est superposé à la zone active de la tablette avec des repères de hauteur de note le long de l'axe X (figure 2.15). Des repères de voyelles le long de l'axe Y permettent un contrôle monomanuel (le même stylet utilisé que pour le contrôle de F_0), ou un contrôle tactile avec la main secondaire sur une portion de la tablette.

Les repères de notes le long de l'axe X sont inspirés d'un clavier de piano avec alternance de touches noires et blanches (voir figure 2.16). Le clavier standard qu'on connait consiste en la superposition de 2 niveaux de touches à mouvement vertical, diatonique majeur à l'avant et chromatique à douze sons à l'arrière. Il a été conservé depuis 600 ans, montrant ainsi son efficacité [Hau99]. La particularité de notre clavier réside dans la continuité et la linéarité de la hauteur mélodique, et dans la disposition des « touches » du clavier dessinées. Dans notre clavier, les intervalles mélodiques sont proportionnels aux intervalles spatiaux de chacune des touches. Ces dernières ne sont pas discrètes et c'est leur centre qu'on vise pour émettre une hauteur de note. Chaque touche a un point commun avec la précédente et la suivante. Les lignes verticales épaisses correspondent aux séparations des touches blanches du piano standard, et pour respecter l'égalité entre intervalle musical et spatial, la séparation entre les touches blanches Mi-Fa et Si-Do du clavier de piano n'apparaît pas. Par continuité, les lignes épaisses correspondent aux cibles des touches noires, à savoir Do#, Re#, Fa#, Sol# et La#. Les lignes fines correspondent aux cibles des notes des touches blanches, à savoir Do, Re, Mi, Fa, Sol, La , Si et Do. Le tout est répété deux fois suivant l'axe Y pour pouvoir jouer à différentes positions sur la tablette sans cacher les repères de sa main.

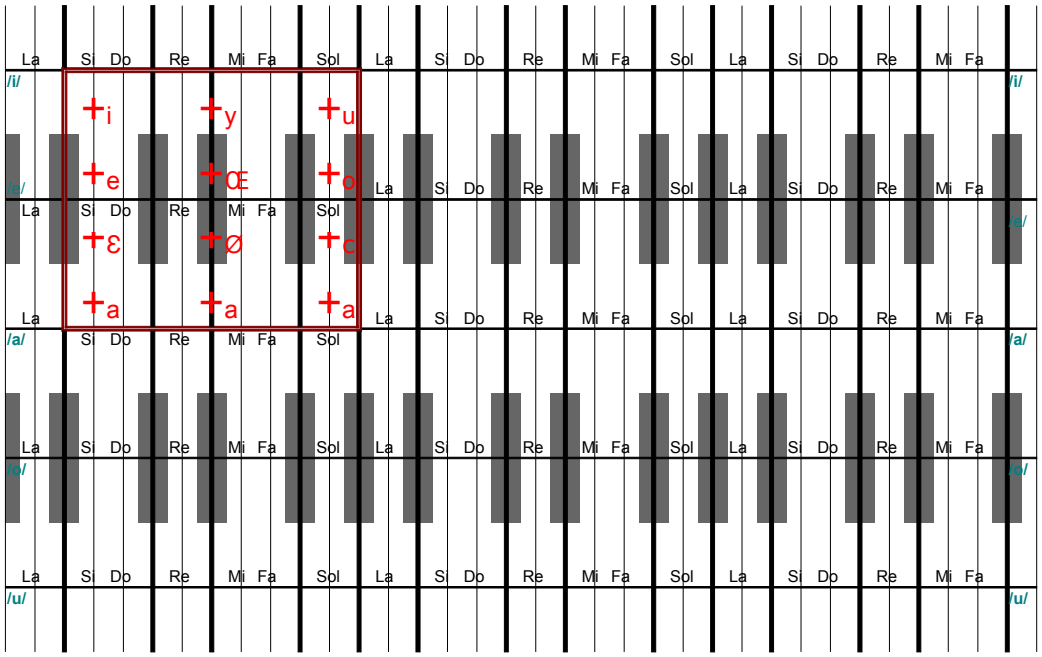


Figure 2.16 – *Le calque de la tablette graphique*

2.7.2 Contrôle du modèle de source glottique

Les paramètres contrôlables du modèle de source glottique sont les suivants :

- la fréquence des impulsions glottiques (F_0) ;
- l'effort vocal, paramètre de haut-niveau relié à la pente spectrale T_l et au quotient ouvert O_q de l'ODGD ;
- la proportion de souffle dans la voix ;
- la tension de la voix, paramètre de haut-niveau agissant sur le coefficient d'assymétrie α_m de l'ODG et au quotient ouvert O_q de l'ODGD ;
- la quantité de Jitter et Shimmer.

Dans une perspective musicale, nous avons choisi de limiter le contrôle gestuel à F_0 et l'effort vocal. Le contrôle de F_0 permet de modifier la mélodie de la voix et l'effort vocal selon les nuances et le début et la fin des sons. On considère que les autres paramètres de la source glottique peuvent être réglés avant le jeu, en les considérant propre au chanteur et constants pendant le jeu. On peut cependant rajouter pour la main secondaire une deuxième interface pour contrôler la qualité vocale (tension/souffle et jitter/shimmer), par exemple à l'aide des deux axes d'un joystick comme proposé par Le Beux [LB09].

L'effort vocal ne nécessite pas une grande précision, on a donc choisit de le relier à la pression du stylet sur la tablette. De plus, on retrouve dans le geste naturel et le geste de pression du stylet la notion d'effort : plus la pression du stylet sera importante, plus l'effort vocal sera grand. La mélodie, soit F_0, nécessite quant à elle une grande précision. D'après Flanagan [FS58], avec des voyelles synthétiques de voix d'homme de $80Hz$ à $120Hz$, le plus petit changement de F_0 perçue est compris entre 0.3% et 0.5% de F_0, soit entre 5 et 9 centièmes de demi-ton.

Si l'on fait correspondre F_0 à la position du stylet suivant un des deux axes de la tablette et qu'on dispose de 0.65 cm par demi-ton, il faut donc être précis au 1/2 millimètre près, ce qui

FIGURE 2.17 – *Contrôle mono-manuel*

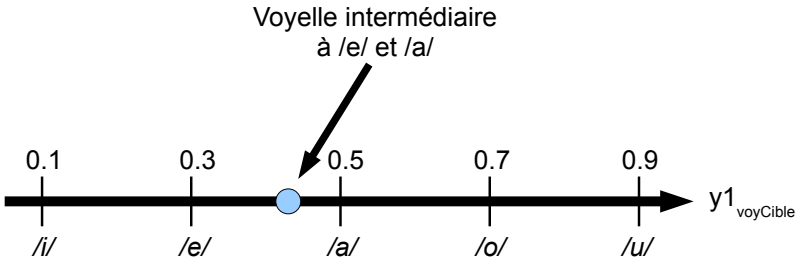


FIGURE 2.18 – *Voyelle cible intermédiaire aux deux voyelles de référence les plus proches*

est possible à la fois par rapport aux caractéristiques techniques de la tablette et par rapport au geste. En effet, toute personne sachant écrire a acquis la faculté de viser très précisément à l'aide d'un stylo, et tout cela avec une dynamique complexe qu'on peut retrouver d'une certaine manière en musique.

2.7.3 Contrôle de l'espace vocalique

On propose deux types de contrôle de l'espace vocalique, l'un avec la même main que celui de la source glottique, l'autre avec la main secondaire séparant ainsi le contrôle de la source et du conduit vocal par l'usage de mains différentes.

a) Contrôle mono-manuel de la source glottique et du conduit vocal

Ce premier mode de contrôle utilise la même tablette graphique Wacom pour contrôler l'ensemble du synthétiseur. Il permet de se focaliser sur sa main principale qui tient le stylet, mais limite l'espace vocalique à un espace 1-D (figure 2.17).

On dispose de 5 voyelles de référence /i,e,a,o,u/ définies par la fréquence centrale, largeur de bande-passante et amplitude de chacun des cinq filtres formantiques, mais une infinité de voyelles sont possibles. Le paramètre $y1_{voyCible}$ contrôle l'évolution continue des voyelles sur un axe où sont présentes les 5 voyelles de référence. Il correspond à la voyelle cible, calculée par interpolation linéaire des valeurs formantiques (fréquence centrale $F_{voyCible}$, amplitude $A_{voyCible}$ et bande passante $B_{voyCible}$) des deux voyelles de référence les plus proches sur

l'axe des voyelles (Exemple à la figure 2.18). Plus formellement, les valeurs formantiques sont données par les relations suivantes :

$$\begin{aligned} F_{voyCible}(y1_{voyCible}) &= F_{voyRefInf}(y1_{voyCible}) \\ &\quad + y1_{voyCible} \cdot [F_{voyRefSup}(y1_{voyCible}) - F_{voyRefInf}(y1_{voyCible})] \\ A_{voyCible}(y1_{voyCible}) &= A_{voyRefInf}(y1_{voyCible}) \\ &\quad + y1_{voyCible} \cdot [A_{voyRefSup}(y1_{voyCible}) - A_{voyRefInf}(y1_{voyCible})] \\ B_{voyCible}(y1_{voyCible}) &= B_{voyRefInf}(y1_{voyCible}) \\ &\quad + y1_{voyCible} \cdot [B_{voyRefSup}(y1_{voyCible}) - B_{voyRefInf}(y1_{voyCible})] \end{aligned} \tag{2.16}$$

où :
$F/A/B_{voyRefInf}$ sont des fonctions qui renvoient la valeur des fréquences centrales / amplitudes / bandes-passantes des formants de la voyelle de référence de valeur inférieure sur l'axe 1-D de $y1_{voyCible}$.
$F/A/B_{voyRefSup}$ sont des fonctions qui renvoient la valeur des fréquences centrales / amplitudes / bandes-passantes des formants de la voyelle de référence de valeur supérieure sur l'axe 1-D de $y1_{voyCible}$.

On a alors le contrôle de l'articulation entre deux voyelles de référence, les formants évoluant continûment entre les deux ensembles de formants cibles correspondant à la voyelle de départ et celle d'arrivée. L'ordre des cinq voyelles a été établi en maximisant la similarité des voyelles adjacentes sur l'axe $y1_{voyCible}$ pour avoir les transitions les plus naturelles, et l'étendue de l'espace vocalique des voyelles orales du français (celui des fréquences centrales des formants). Ainsi, l'ordre des voyelles sur l'axe 1-D correspond à la projection des 2 arêtes du triangle vocalique opposées à l'axe du deuxième formant (voir figure 2.19). Les 2 arêtes projetées sont alors une combinaison linéaire des deux premiers formants. Il n'y aura donc pas d'interpolation possible entre /i/ et /u/ et la levée du stylet sera nécessaire pour passer d'un /i/ à un /u/ de façon naturelle.

On fait correspondre $y1_{voyCible}$ à la dimension spatiale non utilisée de la tablette, c'est à dire l'axe Y. On contrôle donc à la main principale :
- F_0 selon la position du stylet le long de l'axe X ;
- la couleur vocalique $y1_{voyCible}$ selon la position du stylet le long de l'axe Y ;
- l'effort vocal selon la pression du stylet sur la tablette.

b) Contrôle bi-manuel de la source et du conduit vocal

Dans ce deuxième mode de contrôle, on dispose d'un stylet pour le contrôle fin de la source glottique, et d'un doigt de la main secondaire pour le contrôle du conduit vocal. La main principale est alors réservée à la source glottique qui nécessite plus de précision à cause de F_0. La position des deux mains est illustrée sur la figure 2.15.

La main principale munie du stylet contrôle les mêmes paramètres que dans le mode monomanuel à la seule différence que la dimension spatiale selon Y n'est plus reliée à la couleur vocalique $y1_{voyCible}$. La main secondaire permet de se mouvoir dans l'espace vocalique en déplaçant son doigt dans un carré (indiqué en rouge sur le calque, voir figure 2.16)

Peterson et Barney [PB52] sont les premiers à utiliser l'espace des 2 premières fréquences centrales des formants pour contrôler l'espace vocalique. On s'inspire de la représentation de ce triangle vocalique pour représenter la position des différentes voyelles. La position du

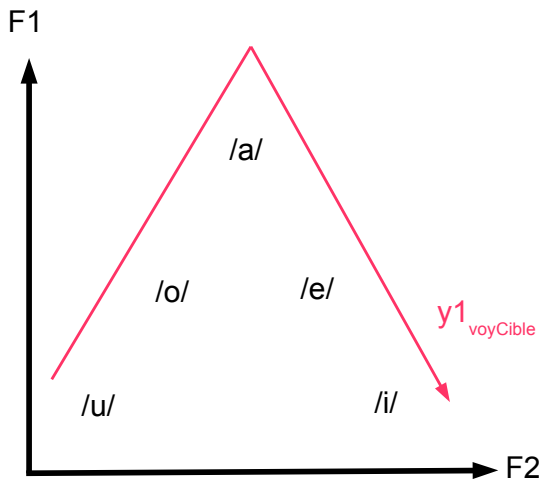


FIGURE 2.19 – *Trajectoire de l'axe 1D choisi dans l'espace vocalique des formants F1-F2*

doigt dans ce carré va déterminer la valeur des formants, en interpolant les valeurs des voyelles cibles du carré. La différence avec un triangle vocalique est double :

- d'une part, la représentation est carrée. Le triangle est projeté sur un carré en transformant un de ses sommets (/a/) sur une arête du carré, afin d'avoir un axe correspondant à l'ouverture de la bouche et un autre orthogonal correspondant à la position antéro-postérieure de la langue.
- d'autre part, tous les formants sont interpolés linéairement, et pas seulement les deux premiers comme la représentation en triangle vocalique pourrait le suggérer. Les deux dimensions spatiales du carré correspondent approximativement à l'évolution linéaire des fréquences de chacun des deux premiers formants, et les autres formants sont interpolés suivants les voyelles ciblées.

Pour arriver à ces fins, nous utilisons 4 voyelles cibles /a,i,y,u/ que nous séparons en deux dimensions : la dimension *antérieure-postérieure* /i,y,u/ et la dimension *ouverte-fermée* /a,i/. Puis on fait correspondre chacune de ces dimensions à la position du doigt de la main secondaire. La surface permettant de contrôler l'espace vocalique sur la tablette est ramenée à un carré de 5x6 cm de façon à pouvoir le parcourir avec l'index sans lever le poignet. On peut alors se concentrer plus facilement sur la main principale qui nécessite une précision plus grande (contrôle de F_0).

Pour résumer, on a pour la main principale (stylet) les correspondances suivantes :

- F_0 selon la position du stylet le long de l'axe X ;
- l'effort vocal selon la pression du stylet sur la tablette.

Et pour la main secondaire (index) :

- l'axe X est relié à la dimension « voyelle postérieure – voyelle antérieure » (interpolation des formants entre ceux de /i,y,u/) ;
- l'axe Y est relié à la dimension « voyelle ouverte – voyelle fermée » (interpolation des formants entre ceux de /a,i/).

c) Contrôle gestuel du chant diphonique

Voir fichiers audios / vidéos 7

Le chant diphonique permet de faire émerger plusieurs hauteurs mélodiques audibles à partir d'une seule voix. On trouve cette pratique à travers le monde, surtout en Asie centrale, de l'Oural et l'Altaï à la Sibérie orientale et au Tibet [Léo04].

Il consiste à accentuer une résonance du conduit vocal suffisamment fortement pour qu'elle soit perçue comme une hauteur mélodique distincte de la fréquence fondamentale de la source glottique, et non comme une simple résonance excitée par la source glottique qu'on percevrait comme un changement de timbre.

Dans ce mode d'utilisation, notre instrument permet de contrôler trois fréquences distinctes : la fréquence F_0 de vibration des plis vocaux, et celles des deux premiers formants F1 et F2.

Pour un contrôle précis de la fréquence et des amplitudes des deux premiers formants, on utilise une deuxième tablette graphique. L'une contrôle l'effort vocal, la durée et la hauteur mélodique comme dans le contrôle bi-manuel précédent. L'autre tablette contrôle dans le plan X-Y la fréquence du premier formant suivant l'axe X, celle du deuxième formant suivant l'axe Y, et l'amplitude du premier ou du deuxième formant avec la pression du stylet suivant le choix du bouton du stylet.

2.8 Résumé et conclusions

Nous avons présenté le *Cantor Digitalis*, un instrument de synthèse de voyelles chantées contrôlables en temps réel. Bien qu'utilisant une technique ancienne qu'est la synthèse par formants, nous parvenons à des résultats permettant une synthèse convaincante et expressive grâce à :

1. une amélioration du modèle de base par son réglage fin et l'ajout de plusieurs dépendances entre ses paramètres : seuil de phonation, anti-résonances des sinus piriformes, dépendances sources-filtres ($F_1(VE)$, $F_i(F_0)$, $A_i(F_0)$), perturbations automatiques de la source ;
2. une prise en compte de l'individualité des voix de chanteurs et de leur réglage fin ;
3. une interface et un choix de paramètres de contrôle adaptés au jeu musical où la hauteur mélodique doit être contrôlée avec justesse ;
4. l'aboutissement de l'instrument sous forme d'application logicielle (détaillée dans l'annexe C).

Un des inconvénients du modèle de contrôle du *Cantor Digitalis* réside dans la difficulté à jouer juste dès que le jeu est rapide et présente de grands intervalles, car les grands intervalles musicaux se traduisent par des distances grandes à parcourir avec le stylet le long de l'axe X de la tablette. Ce serait comparable à une guitare ou un violon qui n'aurait qu'une seule corde. Un saut d'une octave correspond avec notre instrument à un déplacement de 8 cm. Réduire l'échelle de l'intonation rendrait plus rapide le passage d'une note à une autre mais diminuerait la justesse, une petite erreur de position du stylet résultant en une erreur plus grande sur la fréquence fondamentale.

D'autres interfaces de contrôle pourraient être utilisées, en fonction de l'application envisagée. Par exemple, une idée intéressante seraient de contrôler la voix de synthèse avec sa

propre voix dont on analyserait la hauteur, le volume, la qualité vocale, etc. On pourrait alors chanter aisément en duo, comprenant sa propre voix et une voix synthétique qui suivrait cette dernière. Quelques règles simples permettraient ainsi d'harmoniser sa propre voix suivant une loi donné.

Il s'en suit logiquement l'extension des voyelles aux consonnes, c'est ce que nous verrons au chapitre 3 avec ses problématiques de contrôle qui sont d'un autre ordre de difficulté. L'interface proposée (tablette graphique et calque de repères) doit être évaluée, en particulier sa capacité à « chanter » juste et précisément. Les études seront décrites dans la partie II.

Chapitre 3

Digitartic, un instrument de synthèse de syllabes chantées

Sommaire

3.1 Introduction

3.1.1 Contexte

La production de syllabes est complexe et met en jeu une synchronisation des différents éléments de l'appareil vocal. Sur des durées brèves (jusqu'à un ordre d'une dizaine de millisecondes), les différents articulateurs que sont lèvres, langue, mâchoires ou luette, doivent se synchroniser pour changer dynamiquement la forme du conduit vocal, modifiant l'onde acoustique issue de la source glottique. La synchronisation se fait également entre les articulateurs et la vibration des plis vocaux, notamment pour gérer le voisement de certaines consonnes.

Cette synchronisation est issue de l'apprentissage de la parole depuis l'enfance et est spécifique à la langue et de l'accent du locuteur. Par analogie, on peut avancer que l'aspect de l'apprentissage sera prépondérant pour parvenir à un système de contrôle de production de voix artificielle.

On est amené à modéliser d'une part le « moteur » de synthèse de la voix, et d'autre part la manière de le contrôler. Cependant, ces deux composantes ne sont pas indépendantes. Suivant le modèle de synthèse utilisé, il sera ou ne sera pas possible de contrôler certains paramètres de l'articulation. Les raisons peuvent être liées au modèle de synthèse qui ne fournit pas certains paramètres, ou liées au modèle de contrôle qui ne permet pas d'avoir accès à autant de paramètres que nécessaire ou dans des durées suffisamment petites.

Plus notre système artificiel nous permettra de produire des situations diverses, plus notre modèle pourra être considéré comme proche de l'appareil vocal réel. Nous avons choisi de nous limiter à des buts musicaux. Ainsi, nous ne considérons pas la production de toutes les unités phonologiques d'un langage pour ainsi restreindre le nombre élevé de combinaisons et donc de coarticulations de phonèmes possibles. Contrairement à l'instrument Cantor Digitalis du chapitre 2, on se concentre sur l'articulation et l'intonation qui peuvent devenir un paramètre moins important. Cela nous mène plutôt vers des applications de voix de type percussion, ou des voix de type *scat* si l'intonation nécessite comme avec le Cantor Digitalis une justesse importante. L'utilisateur doit donc pouvoir contrôler de manière précise la temporalité de l'articulation, de l'attaque jusqu'à l'arrêt d'un segment de voix, ce qui pose plusieurs problèmes comme cela sera explicité dans ce chapitre.

3.1.2 Contraintes musicales et originalité de notre recherche

Pour une approche musicale, on peut ne pas chercher à reproduire tous les phonèmes d'une langue donnée. En effet, les expériences passées ont montré que les instruments de synthèse de voix parlée sont très difficiles à contrôler. Les opérateurs du VODER [DRW39] devaient être entrainés pendant au moins un an avant de pouvoir synthétiser des phrases en public. Plus récemment, le Glove-talkII [FH98], qui relie les mouvements des mains à un synthétiseur par formant, semble nécessiter beaucoup d'heures de pratique avant de pouvoir parler de façon à peu près intelligible, même si les derniers environnements utilisant le Glove-TalkII permettent un apprentissage plus rapide [PF06] [FPL09]. Notre objectif est de contrôler un maximum de paramètres de production avec le minimum de paramètres de contrôle.

En musique, une contrainte importante pour l'appareil vocal est la synchronisation rythmique. Les articulateurs doivent constamment anticiper leur position cible pour être calés temporellement avec la métrique. Comme tout instrument de musique, les instruments numériques de voix de synthèse ne doivent pas posséder une latence supérieure à 10-20 ms de façon à pouvoir être joués avec un tempo fixe extérieur.

Une deuxième contrainte est l'expressivité de l'articulation. De la même manière qu'un contrôle continu et fin de F_0 est nécessaire pour obtenir une intonation expressive et adaptative, nous pensons que l'expressivité générale de la voix de synthèse sera améliorée si on peut contrôler l'articulation continûment en temps réel. Ainsi, un large éventail de types d'articulations sera disponible, tels le degré d'articulation (hypo- à hyper-articulation), ou la modification de la durée et de l'intensité des stades de l'articulation.

Ces contraintes requièrent un modèle de production efficace et une interface de haute résolution. Les systèmes existant basés sur des modèles physiques évolués demandent trop de puissance de calcul pour être utilisés en temps-réel. La synthèse utilisant des bases de données de voix réelles, par concaténation [SBVB06] ou basée sur l'apprentissage et la reconstruction par HMMs [AdP+12], introduisent au moins un phonème de retard quand les segments de parole sont choisis en temps réel, la plus petite séquence synthétisée se faisant sur un diphone. Parmi les systèmes destinés à un contrôle de l'articulation temps réel, quelques-uns sont destinés à l'articulation de voyelles comme le VOICER de Kessous [Kes02], le Handsketch de d'Alessandro et Dutoit [dD07] ou notre instrument Cantor Digitalis présenté au chapitre 2 [FLBd11], mais ils ne sont pas capables de contrôler l'articulation. Le Glove-TalkII de Fels et Hinton [FH98] permet la production de syllabes, mais celle-ci est faite par gestes de sélection, c'est-à-dire sans contrôle fin pendant le processus de l'articulation. De plus, l'utilisation de gants haptiques présente l'inconvénient d'une latence importante, de l'ordre de 10 à 20 ms comme le mentionne Kunikoshi et al. dans leur système de synthèse temps réel destiné à la communication pour personnes muettes [KQS+11]. Le seul instrument trouvé dans la littérature qui propose un contrôle continu de l'articulation est le SqueezeVox LISA de Cook et Leider [CL00] qui utilise un modèle de tube acoustique, mais cet aspect de l'instrument est très peu documenté[1]. Contrairement aux systèmes cités ci-dessus, notre travail permet un contrôle fin de l'articulation des phonèmes grâce à des gestes de modification (comme définis par [Cad88]), tout en permettant une haute précision temporelle et un contrôle précis sur l'intonation.

D'autre part, un instrument de synthèse vocale qui ne contrôlerait que le déclenchement des différents stades de l'articulation et non le contrôle de la phase d'articulation est voué à ne pas pouvoir être joué dans des musiques nécessitant une haute précision temporelle, comparable au seuil de perception de la non simultanéité de deux événements, de l'ordre de la dizaine de millisecondes. En effet, hormis les occlusives pour lesquelles l'attaque musicale (au sens de son « centre perceptif », ou P-center [Gor87] [MMF76]) se situe au début de la transition CV en même temps que l'explosion, l'attaque musicale d'une syllabe se situe en fin de transition CV qui a une longueur de l'ordre de plusieurs dizaines de millisecondes (30 à 50 ms chez les semi-voyelles ou liquides). Ainsi, le déclenchement ne sera pas simultané à l'attaque, causant alors des difficultés à se synchroniser avec des éléments musicaux extérieurs et à jouer rapidement.

Si l'on s'intéresse seulement au contrôle des syllabes et leurs articulations, comme par exemple les récitations onomatopéiques des percussions indiennes, alors le contrôle de F_0 ne nécessite pas une grande précision. Mais s'il on veut « scatter », c'est à dire chanter en utilisant des onomatopées, alors le contrôle de F_0 devra être précis. L'interface devra alors au moins inclure les contrôles suivants :

- F_0 et lieu d'articulation (continu) ;
- phase d'articulation (position articulatoire entre deux phonèmes, continu) ;
- effort vocal (continu) ;

1. Voir la vidéo de démonstration http://www.cs.princeton.edu/sound/listen/LisaBig.wmv, consulté le 20 juin 2013

– mode d'articulation (discret) ;
– voyelles (continu pour une plus grande richesse).

Le modèle de production du *Digitartic* est détaillé dans la partie 3.2. Dans la partie 3.3, le principal modèle de production basé sur le contrôle continu de l'articulation est présenté. Enfin, un modèle alternatif avec des gestes de sélection sur des interfaces tactiles multipoints est développé.

3.2 Modèle de production VCV

Dans cette partie est présenté le Digitartic, un synthétiseur articulatoire de type VCV (Voyelle-Consonne-Voyelle) utilisant la synthèse par formant, et le modèle de source RT-CALM [DdH03] [ddLBD06]. La langue de référence est le français, bien qu'il soit possible de créer des consonnes avec des lieux d'articulation intermédiaires. Il est l'extension de l'instrument Cantor Digitalis, étant donné qu'il utilise la même technologie de base, mais avec l'articulation des consonnes en plus. On ne reprendra donc pas la présentation entière du modèle de production de l'instrument donnée dans le chapitre 2, mais seulement ce qui a été ajouté. La figure 3.1 donne une représentation simplifiée du fonctionnement du Digitartic. La partie nouvelle par rapport au *Cantor Digitalis* est la partie gauche du schéma, composée des règles d'articulation, de la source de bruit consonantique et des liens entre les différents modules dont ceux du Cantor Digitalis.

3.2.1 Structure temporelle / phonémique

En plus des contrôles du *Cantor Digitalis*, notre instrument permet d'articuler des syllabes de type $V_1 - C - V_2$, où

– V_i désigne la voyelle i contenue dans l'ensemble des voyelles du français /i,e,a,o,u/ et de leurs interpolations suivant l'axe /i,e,a,o,u/ (voir section 2.7.3)
– C désigne une consonne contenue parmi les consonnes du français /p,b,t,d,k,g,f,v,s,z,ʃ,ʒ,w,ɥ,j,m,n,ŋ/ et de leurs interpolations suivant le lieu d'articulation pour un même mode d'articulation et même voisement.

Toute combinaison de $V_1 - C - V_2$ comme $V_1 - C_1 - V_2 - C_2 - V_3$, ou toute sous-partie de $V_1 - C - V_2$ comme $C - V$ peut être produite, ainsi qu'en y combinant des silences. Les voyelles utilisées sont les mêmes qu'avec le *Cantor Digitalis* en mode « mono-tablette » (voir section 2.7.3). Quant aux consonnes, elles sont issues de 4 modes d'articulation, à savoir les occlusives, les fricatives, les nasales et les semi-voyelles. Pour les occlusives et les fricatives, on peut choisir entre voisé et non voisé. Pour un mode d'articulation, on groupe les lieux d'articulation de référence comme il suit : bilabiale ou labiodentale ; alvéolaire ; post-alvéolaire ou palatale ou vélaire. Ainsi, pour des raisons de simplicité de l'interface, on aura 3 lieux d'articulation pour chaque mode d'articulation, comme indiqué au tableau 3.1.

Mode d'articulation	Groupe 1	Groupe 2	Groupe 3
Occlusives	bilabiale	alvéolaire	palatal ou vélaire
Fricatives	labiodentale	alvéolaire	post-alvéolaire
Nasales	bilabiale	alvéolaire	alvéolaire et palatale
Semi-voyelles	bilabiale	alvéolaire	palatale

Table 3.1 – *Groupement des lieux d'articulation dans le synthétiseur*

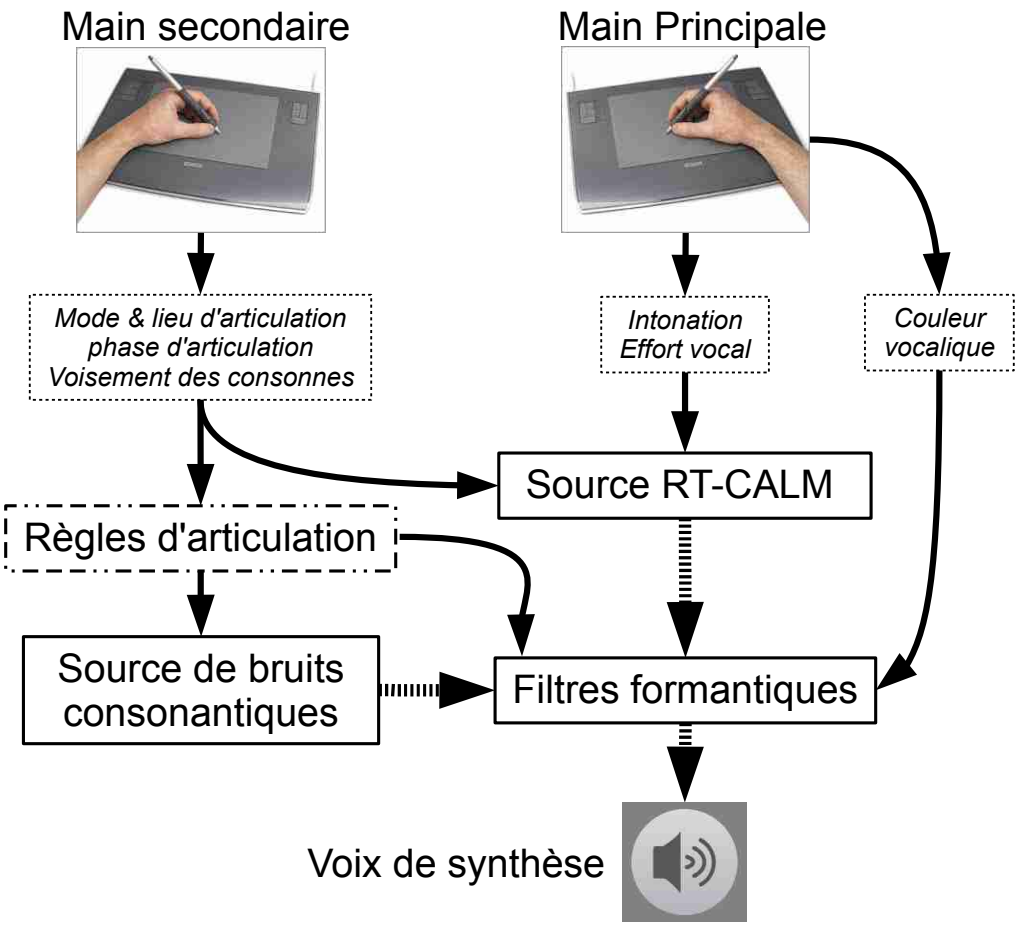


FIGURE 3.1 – *Représentation schématique du fonctionnement du synthétiseur Digitartic*

On peut classifier la structure temporelle des dissyllabes VCV en deux catégories suivant leur mode d'articulation :

- les fricatives, les semi-voyelles et les nasales qui présentent une structure temporelle plutôt symétrique autour de la phase médiane de l'articulation d'une séquence VCV, qu'on nommera *consonne tenue*. Les fricatives sont les seules parmi ces 3 modes d'articulation à disposer d'un bruit consonantique. Celui-ci apparaît principalement sur la consonne tenue au moment où les articulateurs obstruent suffisamment le conduit vocal pour qu'un bruit se crée au niveau de l'obstruction.
- les occlusives qui présentent une dissymétrie par rapport à la consonne tenue (ici un silence, ou une barre de voisement si voisées et articulées rapidement). En effet, l'explosion dûe au relâchement brusque de l'air par la disparition de l'obstruction complète des articulateurs ne se manifeste que lors de la phase CV et non lors de la phase VC.

Ainsi on traitera de manière symétrique la synthèse des syllabes CV et VC dont la consonne est une fricative, une semi-voyelle ou une nasale. Concernant les occlusives, la partie voisée de la synthèse sera traitée de manière symétrique tandis que la partie de bruit d'explosion sera traitée de manière asymétrique entre les syllabes VC et CV.

3.2.2 Cibles formantiques

A partir de valeurs de formants données dans la thèse de Garnier-Rizet [GR94], de Klatt [Kla80], par l'université de Laval sur son site internet [Lav], d'indications données par Stevens [Ste98], d'analyses personnelles, et de nombreux réglages « à l'oreille », nous avons établi le Tableau 3.2. Il comprend les valeurs des filtres formantiques (fréquence centrale, bande-passante et amplitude) correspondant aux formants cibles atteints en hyper-articulation. Le passage d'une cible à une autre sera traité dans la section 3.2.4.

Toutes ces valeurs ont été ajustées avec la voyelle /a/, en situation /a/-C-/a/, et on garde ces valeurs cibles pour des contextes vocaliques autres. Ceci est une approximation assez grossière étant donné que la forme du conduit vocal sur une consonne tenue dépend de la voyelle précédent ou suivant cette consonne. Phonologiquement, cette consonne est perçue comme identique, mais phonétiquement, la consonne peut être différente, présentant entre autres des lieux d'articulation distincts. Par exemple, /ki/ et /ku/ présentent la même consonne /k/, mais phonétiquement le lieu d'articulation de la première est bien plus antérieure que la seconde. Les formants et le spectre des bruits consonantiques résultant de la position des articulateurs, ils sont donc distincts pour un /k/ suivi ou précédé de voyelles différentes.

Les valeurs des filtres formantiques des voyelles associées sont les mêmes qu'avec le *Cantor Digitalis* et données dans le Tableau 2.1 de la section 2.3.1.

A ces résonances sont ajoutées des anti-résonances pour les consonnes nasales (on ne met pas de voyelle nasale à disposition). En utilisant des valeurs de [GR94] et de [Ste98], et en ajustant à l'oreille, nous avons choisi les valeurs données dans le tableau 3.3 pour un filtre coupe-bande de type *biquad* (2 zéros et 2 pôles).

3.2.3 Bruits consonantiques

Parmi les modes d'articulation des consonnes du français, on observe deux types de bruit consonantique suivant les modes d'articulation occlusif et fricatif :

- Le bruit de friction provient du passage de l'air entre deux articulateurs suffisamment proches l'un de l'autre pour créer des turbulences dans l'écoulement et occasionner alors un bruit sonore. Il a comme source la constriction située entre les incisives supérieures et la lèvre inférieure pour les consonnes /f,v/, ou entre l'alvéole et la pointe de la langue

	Fréquences F_i (Hz)	**Bande-passante B_i (Hz)**	**Amplitude A_i (dB)**
/p/	310. 900. 2000. 3150. 3600.	10. 18. 20. 30. 20.	-13. -23. -2. 1. -40.
/b/	460. 1550. 2570. 2980. 3600.	10. 15. 20. 30. 20.	-1. -3. -2. -2. -5.
/t/	700. 1200. 2500. 2800. 3600.	13. 13. 40. 60. 40.	0. 0. -5. -7. -30.
/d/	440. 880. 2160. 2860. 3600.	10. 12. 20. 30. 20.	-6. -1. -18. -10. -37.
/k/	290. 750. 2300. 3080. 3600.	10. 10. 20. 30. 20.	-12. -9 -14. -11. -11.
/g/	460. 1550. 2570. 2980. 3600.	10. 15. 20. 30. 20.	-1. -3. -2. -2. -5.
/f/	700. 1200. 2500. 2800. 3600.	13. 13. 40. 60. 40.	0. 0. -5. -7. -30.
/v/	440. 880. 2160. 2860. 3600.	10. 12. 20. 30. 20.	-6. -1. -18. -10. -37.
/s/	290. 750. 2300. 3080. 3600.	10. 10. 20. 30. 20.	-12. -9 -14. -11. -11.
/z/	460. 1550. 2570. 2980. 3600.	10. 15. 20. 30. 20.	-1. -3. -2. -2. -5.
/ʃ/	700. 1200. 2500. 2800. 3600.	13. 13. 40. 60. 40.	0. 0. -5. -7. -30.
/ʒ/	440. 880. 2160. 2860. 3600.	10. 12. 20. 30. 20.	-6. -1. -18. -10. -37.
/w/	290. 750. 2300. 3080. 3600.	10. 10. 20. 30. 20.	-12. -9 -14. -11. -11.
/ɥ/	460. 1550. 2570. 2980. 3600.	10. 15. 20. 30. 20.	-1. -3. -2. -2. -5.
/y/	700. 1200. 2500. 2800. 3600.	13. 13. 40. 60. 40.	0. 0. -5. -7. -30.
/m/	440. 880. 2160. 2860. 3600.	10. 12. 20. 30. 20.	-6. -1. -18. -10. -37.
/n/	290. 750. 2300. 3080. 3600.	10. 10. 20. 30. 20.	-12. -9 -14. -11. -11.
/ŋ/	290. 750. 2300. 3080. 3600.	10. 10. 20. 30. 20.	-12. -9 -14. -11. -11.

TABLE 3.2 – *Valeurs des formants cibles des consonnes de la voix de Tenor synthétisée*

	Fréquences (Hz)	**Bande-passante (Hz)**
/m/	1300.	325.
/n/	2000.	500.
/ŋ/	2800.	700.

TABLE 3.3 – *Valeurs des anti-résonances des consonnes nasales du synthétiseur*

pour les consonnes /s,z/, ou bien entre le palais et le dos de la langue pour les consonnes /ʃ,ʒ/. La friction dure autant de temps que le passage de l'air dans le conduit vocal perdure.

– Le bruit d'explosion intervient lors du relâchement de la constriction totale des articulateurs au tout début de la phase CV. Le bruit est créé au niveau des lèvres pour les consonnes /p,b/, ou entre l'alvéole et la pointe de la langue pour les consonnes /t,d/, ou bien entre le palais et le dos de la langue pour les consonnes /k,g/. Contrairement à celui de la friction, il est bref et de durée peu contrôlable de l'ordre de 10-30 ms. Son spectre dépend du lieu d'articulation et des phonèmes adjacents. L'écartement des articulateurs nécessite une durée suffisamment grande pour que le bruit dû au relachement de l'occlusion soit toujours composé d'un bruit de friction bref après l'explosion.

On fait l'hypothèse, en première approximation, que le spectre du bruit consonantique, indépendamment de son évolution dans le temps, va dépendre de seulement trois paramètres :

1. le ou les lieu(x) d'articulation qui vont agir sur le spectre de la source du bruit en divisant le conduit vocal en plusieurs cavités autour de la ou des constrictions. Le lieu d'articulation dépend de la consonne qu'on veut produire ainsi que du contexte phonétique.
2. la configuration vocalique pour un même lieu d'articulation, c'est à dire les phonèmes précédant ou suivant une consonne au sens phonétique. Elle va modifier la position des articulateurs tout en gardant le même lieu d'articulation, et ainsi modifier la forme globale du conduit vocal et filtrer la source de bruit consonantique.
3. le fonctionnement des plis vocaux. S'ils sont en vibration pendant la production du bruit occlusif ou fricatif, le bruit est modulé en amplitude par le son voisé issu de la source glottique.

On fait alors l'approximation suivante : pour un même lieu d'articulation, le bruit de friction et d'explosion auront à peu près le même spectre. La différence vient essentiellement de sa durée et de l'évolution de son intensité. Pour la synthèse du bruit consonantique, on procède de la façon suivante (schématisée sur la figure 3.2) :

1. Pour chacun des trois groupes de lieux d'articulation (bilabiale ou labiodentale ; alvéolaire ; post-alvéolaire ou palatale ou vélaire), on associe une source de bruit de même contenu spectral. Il est réalisé à partir d'un bruit blanc uniforme filtré par un passe-bande de type *butterworth* d'ordre 2. La réponse en fréquence des filtres pour les trois groupes de lieux d'articulation est donnée à la figure 3.3. La forme du conduit vocal s'adaptant avec F_0 en élevant la fréquence centrale des formants [JSW04], on fait dépendre les fréquences de coupure inférieure et supérieure du filtre passe-bande à F_0. On décide d'appliquer un facteur linéaire valant 1 à 200 Hz et 1.2 à 1000 Hz comme indiqué sur la figure 3.3.
2. L'amplitude du signal temporel du bruit est ensuite modulée selon la phase d'articulation (i.e. la position de l'articulation entre les deux phonèmes visés), le mode d'articulation, et le voisement. La phase d'articulation est reliée au degré de constriction chez les occlusives et les fricatives, on n'aura donc pas de bruit consonantique pour une phase d'articulation proche de la voyelle. Pour un même lieu d'articulation, le bruit sera plus court et plus intense pour un mode d'articulation occlusif que pour un mode d'articulation fricatif.

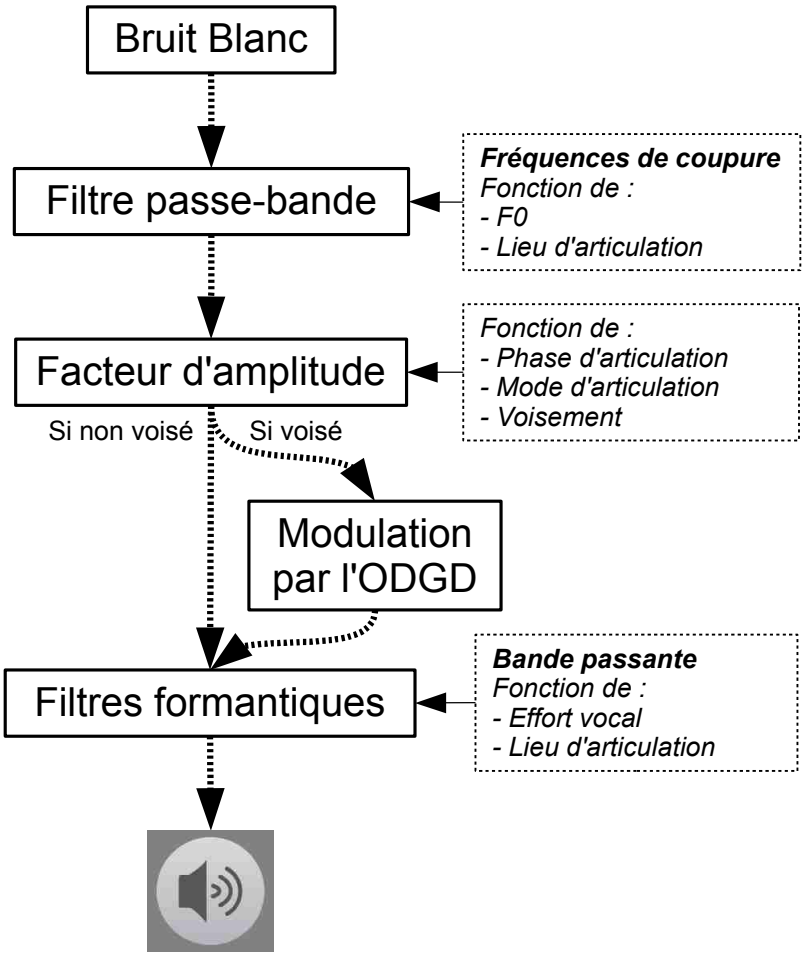
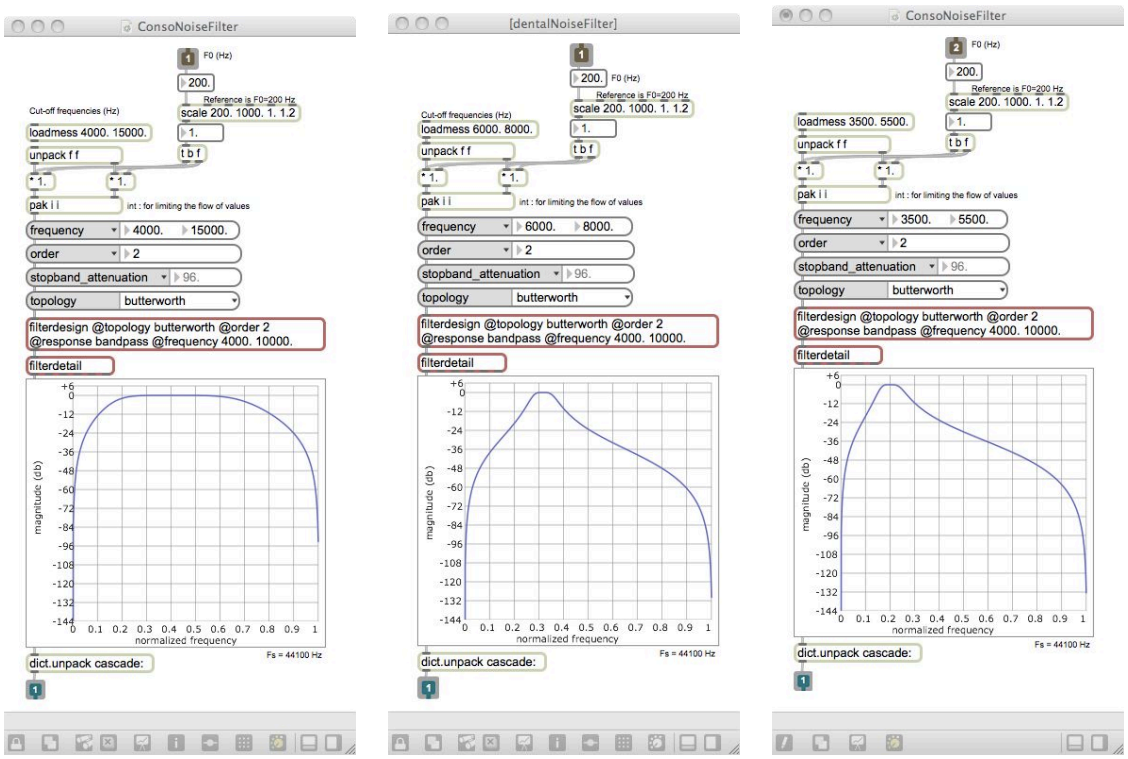


FIGURE 3.2 – *Production d'un bruit consonnatique (friction ou explosion) dans le digitartic*



(a) *bruit bilabial et labiodental* (b) *bruit alveolaire* (c) *bruit post-alveolaire, palatal et vélaire*

Figure 3.3 – *Réponse en fréquence (0-22kHz) et implémentation dans le synthétiseur des sources de bruit consonantique*

3. Dans le cas des consonnes voisées, on module de nouveau le bruit, mais cette fois-là par l'onde de débit glottique. En effet, le débit d'air nécessaire au bruit de friction et d'explosion provient de la source glottique qui vibre pour les consonnes sonores.
4. Afin d'apporter une signature spectrale des voyelles adjacentes à la consonne, le bruit est filtré différemment selon la voyelle précédente en fin de phase VC et selon la voyelle suivante en début de phase CV où le bruit est encore présent (voir section 3.2.4). Le filtrage utilisé repose sur les formants mesurés sur ces consonnes ; les valeurs de référence (voix de ténor, F_0=200 Hz) des formants des consonnes sont données au tableau 3.2. Cependant, la zone de filtrage formantique se situe entre 100 et 5000 Hz, et la zone 3000-5000 Hz varie peu avec les voyelles. Or le contenu spectral du bruit consonantique utilisé débute vers 3000 Hz. Donc l'effet se fera sentir principalement pour les consonnes voisées dont le bruit est modulé par l'ODG et qui sont donc caractérisés par un spectre sonore non nul en dessous de 3000 Hz. Pour reproduire l'effet de l'effort vocal sur les bruits consonantiques, il nous a semblé judicieux de modifier la bande-passante des filtres formantiques avec l'effort vocal : plus l'effort vocal est faible et plus la bande-passante des filtres formantiques est augmentée (*passe-tout* pour un effort vocal faible, jusqu'aux valeurs du tableau 3.2 pour effort vocal maximal). L'effet est une réduction des hautes fréquences pour un effort vocal faible et une signature des formants plus importante. L'implémentation des filtres formantiques pour les bruits fricatifs et occlusifs dans le *Digitartic*, ainsi que la dépendance de l'effort vocal avec leur bande-passante, sont données à la figure 3.4.

3.2.4 Règles sur les transitions entre consonne et voyelle

Les règles de base du *Digitartic* ne portent que sur des positions ou durées relatives à la longueur de la transition. Elles s'y rapportent quelle que soit sa longueur.

a) Visualisation des règles

Pour la production des syllabes, les paramètres évoluent continûment entre les voyelles / consonnes cibles : les valeurs des coefficients des 5 filtres formantiques (avec pour chacun sa fréquence centrale, sa bande-passante et son amplitude) ; l'amplitude du bruit consonantique ; le taux de voisement ; l'aspiration.

Les évolutions de ces paramètres est regroupée sous un paramètre de haut-niveau, qu'on appelle *phase d'articulation*. Il correspond à l'emplacement des articulateurs entre 2 positions articulatoires de référence (i.e. voyelle ou phase médiane de la consonne).

Chacun de ces paramètres est borné entre 0 et 1 et on représente leur évolution temporelle de la voyelle vers la consonne. L'évolution de ces paramètres est symétrique par rapport à la consonne, c'est-à-dire qu'elles sont les mêmes pour la phase VC que pour la phase CV, sauf pour les bruits d'explosion qui interviennent seulement lors de la phase CV. Par exemple, la figure 3.5 affiche l'évolution des paramètres pour l'articulation VCV avec /p/ pour consonne (à lire de gauche à droite pour la phase VC et de droite à gauche pour la phase CV) :

- l'amplitude du bruit consonantique (courbe jaune). Elle est nulle sur la partie correspondant à la voyelle tenue (gauche) et sur la consonne tenue (droite) et on observe un pic assez bref correspondant à l'explosion qui intervient au début de la phase CV. Dans le cas de /p/, la consonne tenue correspond à un silence.
- le voisement (courbe bleue). Il vaut 1 sur la voyelle (comme sur n'importe quelle voyelle) et 0 sur la consonne tenue, la consonne étant une occlusive sourde. Le voisement se

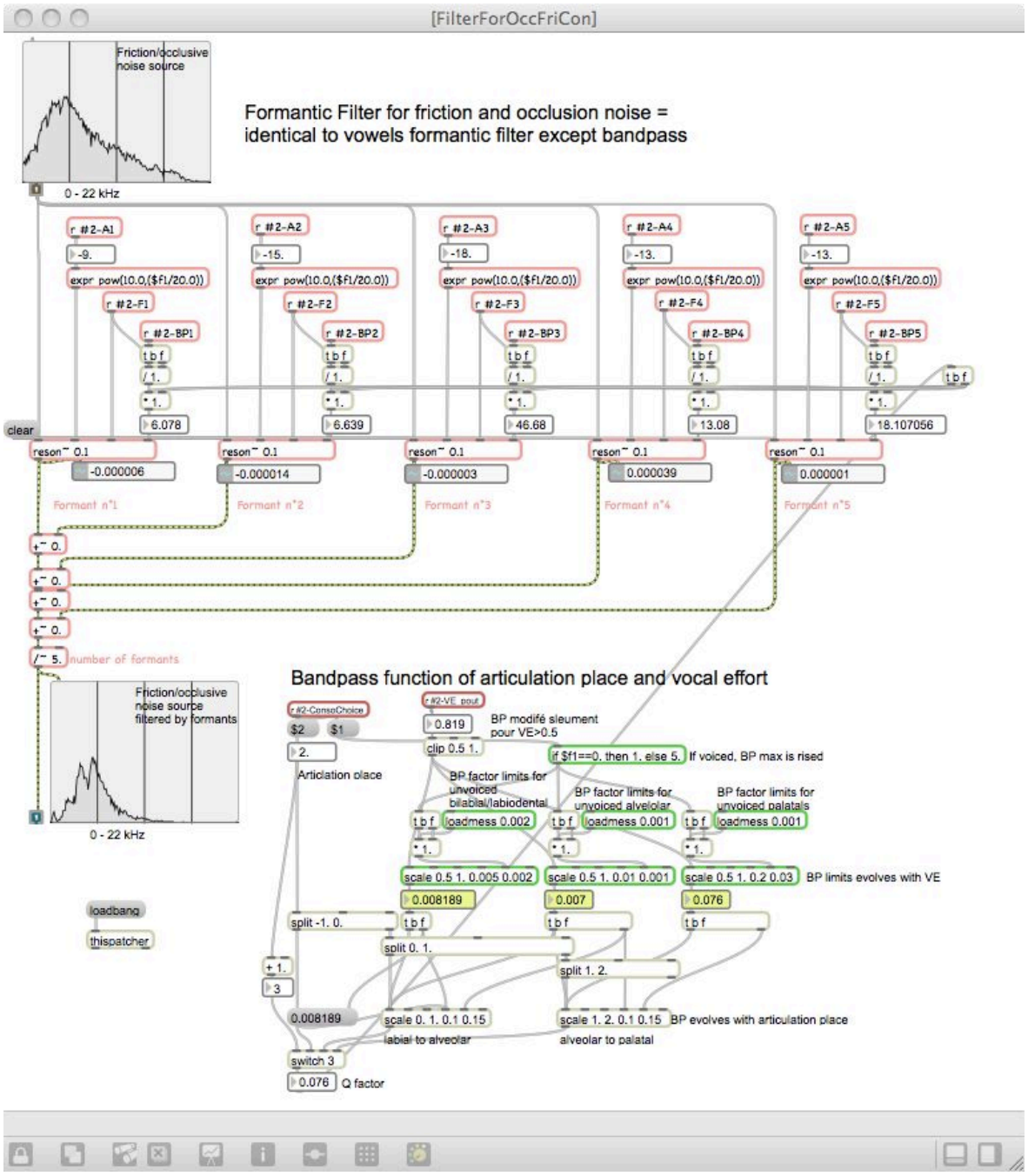


Figure 3.4 – *Implémentation des filtres formantiques pour le bruit consonantique et la dépendance de leur bande-passante avec l'effort vocal VE*

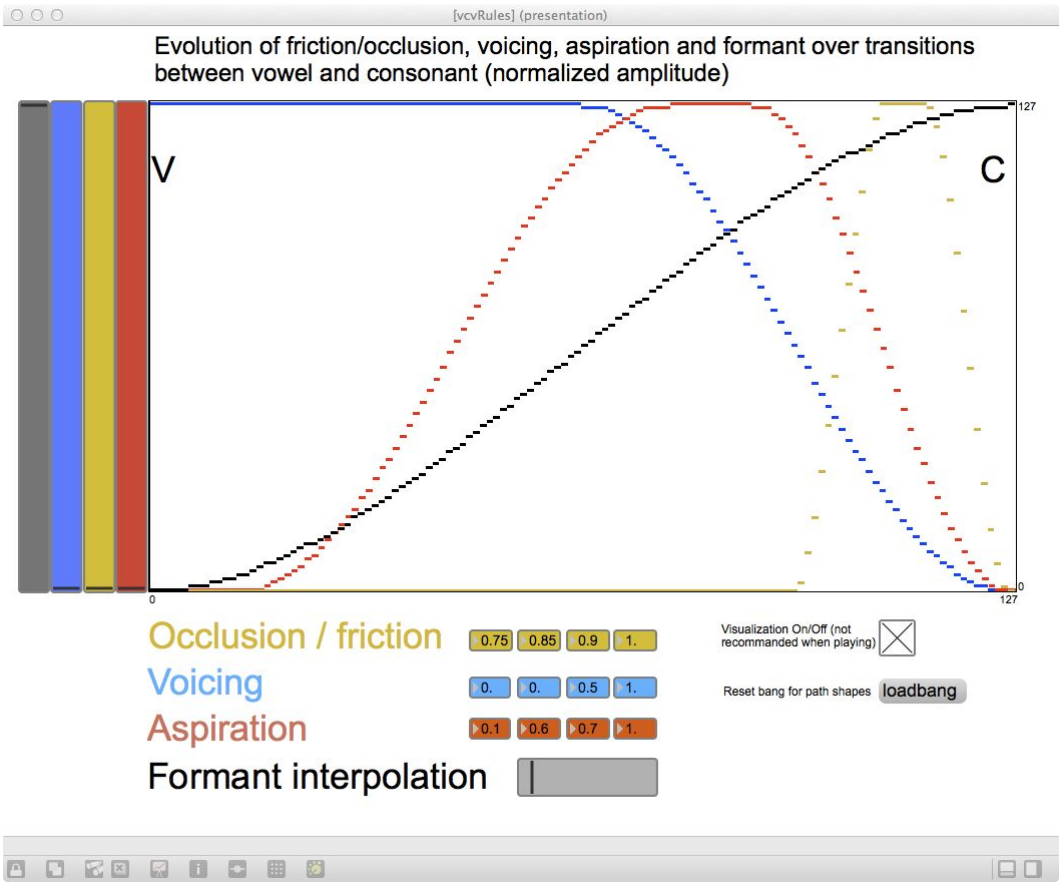


FIGURE 3.5 – *Exemple d'évolution des paramètres de transition articulatoire VC (lecture de gauche à droite) et CV (lecture de droite à gauche) avec la consonne /p/*

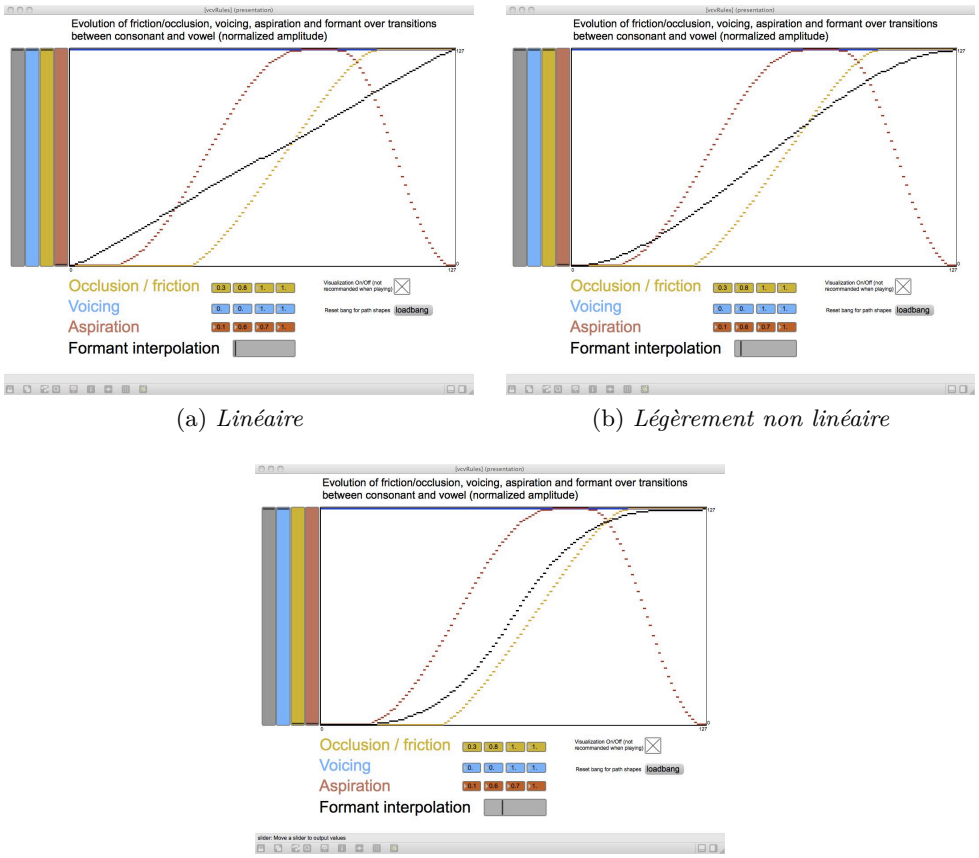
rétablit après l'explosion dans le sens CV (de droite à gauche), l'occlusive /p/ étant sourde.
- l'aspiration consonantique (courbe rouge). Elle intervient au moment où les articulateurs sont proches, mais pas suffisamment pour créer un bruit de friction ou d'explosion. On l'a placé de telle sorte qu'il soit après le bruit d'explosion dans le sens CV.
- le passage d'une configuration de formants à une autre (courbe noire), 0 correspondant aux formants des voyelles, 1 à ceux de la consonne tenue.

L'affichage des valeurs temps-réel se fait par des curseurs à gauche du graphique. Les valeurs en dessous du graphique permettent l'édition de chacune des courbes et correspondent à la position de ses points critiques.

b) Trajectoires des formants

Concernant les valeurs des filtres formantiques (fréquences centrales, bandes-passantes et amplitudes), nous disposons d'une base de données pour les voyelles cibles (figure 2.1) et pour les consonnes cibles (figure 3.2). Le passage d'une configuration à une autre (voyelle à consonne ou consonne à voyelle) se fait par interpolation. On peut régler la linéarité de la courbe d'interpolation de l'évolution des formants, comme par exemple sur la figure 3.6, où l'on passe (a) d'une interpolation linéaire à (c) fortement non linéaire.

La base de formants pour les consonnes cibles ayant été réalisée dans un contexte /a-C-a/,



(a) *Linéaire* (b) *Légèrement non linéaire*

(c) *Non linéaire*

Figure 3.6 – *Trois exemples de transition pour les trajectoires formantiques entre voyelle et consonne, plus ou moins linéaire (courbes noires)*

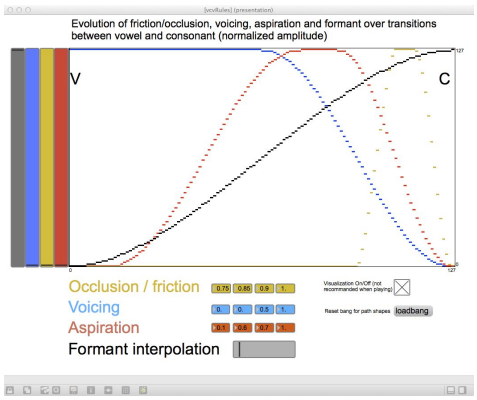
la discussion des différentes interpolations ne prend sens que dans le cas du contexte /a-C-a/. En effet, la forme de la trajectoire des formants paraît secondaire vis-à-vis des valeurs cibles.

c) Evolution du bruit et du voisement pour les consonnes occlusives et fricatives

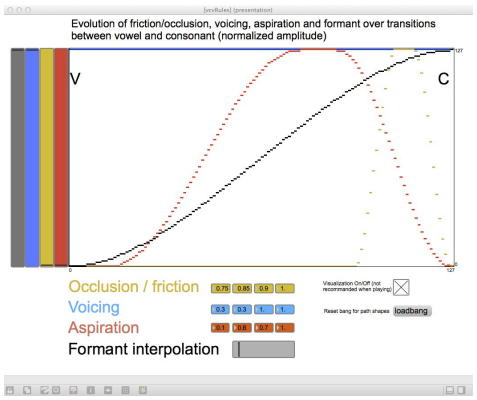
Les trajectoires de l'interpolation du bruit et du voisement entre les cibles phonétiques des voyelles et des consonnes occlusives sont données respectivement par les courbes jaunes et bleues de la figure 3.7. La différence entre occlusives sourdes et sonores (respectivement colonne de gauche et colonne de droite sur la figure 3.7) se manifeste par la position du voisement, qui se situe avant le bruit d'explosion pour les occlusives sonores dans le sens CV. Le bruit n'intervient qu'en phase CV et brièvement par rapport à la longueur de la transition (d'environ 15 à 25% selon le lieu d'articulation).

Les trajectoires de l'interpolation du bruit et du voisement entre les cibles phonétiques des voyelles et des consonnes fricatives sont données respectivement par les courbes jaunes et bleues de la figure 3.8 pour les fricatives. La différence entre fricatives sourdes et sonores (respectivement colonne de gauche et colonne de droite sur la figure 3.8) ne se manifeste ici que par le voisement, en ne s'interrompant pas pour les fricatives sonores.

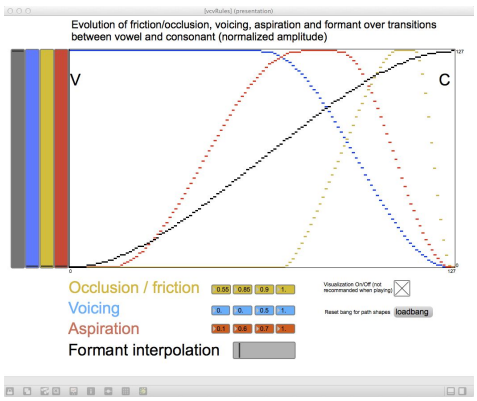
Pour les occlusives et les fricatives, le bruit débute avant la consonne cible afin que le bruit soit modifié par son contexte vocalique. Les figures 3.7 et 3.8 mettent en évidence que le bruit



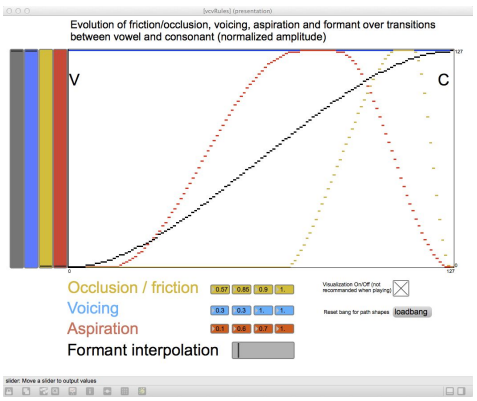
(a) *V-/p/ (ou /p/-V suivant les paramètres et le sens de lecture)*



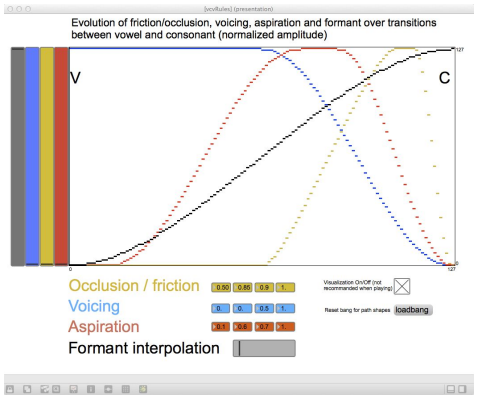
(b) *V-/b/ (ou /b/-V suivant les paramètres et le sens de lecture)*



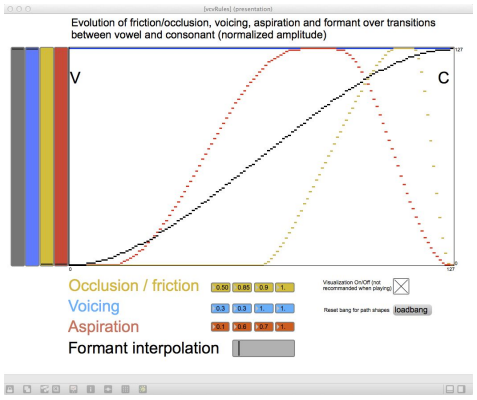
(c) *V-/t/ (ou /t/-V suivant les paramètres et le sens de lecture)*



(d) *V-/d/ (ou /d/-V suivant les paramètres et le sens de lecture)*

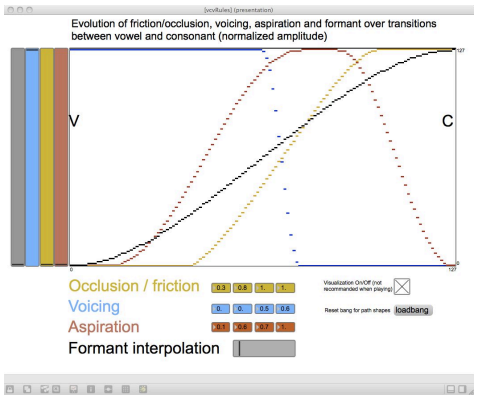


(e) *V-/k/ (ou /k/-V suivant les paramètres et le sens de lecture)*

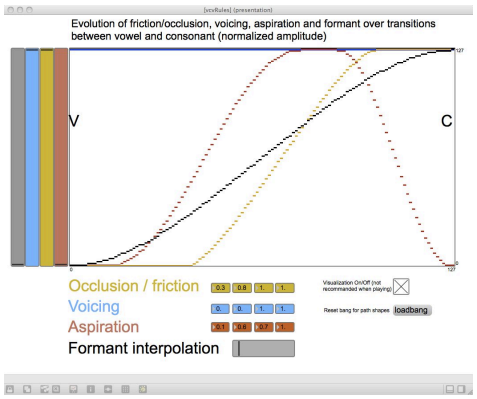


(f) *V-/g/ (ou /g/-V suivant les paramètres et le sens de lecture)*

FIGURE 3.7 – *Trajectoires des paramètres entre voyelles et consonnes occlusives cibles*



(a) *V-/f,s,ʃ/ (ou /f,s,ʃ/-V suivant les paramètres et le sens de lecture)*



(b) *V-/v,z,j/ (ou /v,z,j/-V suivant les paramètres et le sens de lecture)*

FIGURE 3.8 – *Trajectoires des paramètres entre voyelles et consonnes fricatives cibles*

intervient à un moment où la courbe d'interpolation des formants est en cours de transition. Ainsi, les filtres formantiques modifiant les bruits consonantiques (voir section 3.2.3) seront propres à la voyelle adjacente. Cependant, l'influence des filtres formantiques sur les bruits consonantiques est très faible étant donné qu'on se situe dans la partie haute de l'interpolation, correspondant aux valeurs des consonnes cibles, et donc loin des valeurs des voyelles cibles.

d) Amplitude des bruits consonantiques

A chacune des consonnes est associée une amplitude de bruit consonantique de référence $AmpBruitConso_{Ref}$ regroupées dans le tableau 3.4. Elles ont été réglées à l'oreille de façon à ce qu'elles correspondent à l'amplitude du bruit à effort vocal maximal. On remarque les faibles amplitudes pour les occlusives bilabiales et les fricatives labiodentales.

L'amplitude des bruits consonantiques évolue avec l'effort vocal. Empiriquement, nous l'avons réglée de telle sorte que l'amplitude effective $AmpBruitConso(VE)$ soit modulée par l'effort vocal VE de la façon suivante :

$$AmpBruitConso(VE) = \begin{cases} (2 \cdot VE - 1) \cdot AmpBruitConso_{Ref} & \text{si } VE \geq 0.5 \\ 0 & \text{si } VE < 0.5 \end{cases} \quad (3.1)$$

Occlusives	/p/	/b/	/t/	/d/	/k/	/g/
Amplitude de bruit	0.6	0.3	3.0	1.5	4.2	2.1
Fricatives	/f/	/v/	/s/	/z/	/ʃ/	/ʒ/
Amplitude de bruit	0.5	0.25	5.0	2.5	9.0	4.5

TABLE 3.4 – *Amplitude des bruits consonantiques pour un effort vocal maximal (unité arbitraire)*

ampConsoAspi ampConsoNoise

e) Évolution de l'aspiration

L'aspiration des consonnes, comme Fant [Fan60] la définit, est décrite comme le bruit créé par la ou les constrictions partielles des articulateurs (dont la source glottique), qui intervient lorsque les articulateurs ne sont pas suffisamment proches pour produire un bruit de friction. Elle intervient après ou en chevauchement sur la période de friction pour les fricatives et les plosives dans le sens CV.

On modélise l'aspiration des consonnes en jouant sur l'amplitude du souffle glottique, modélisé par un bruit blanc filtré entre 1000 et 6000 Hz et modulé par l'ODGD (voir section 1.7). Une même évolution temporelle de l'aspiration est donnée dans les syllabes VC et CV, mais l'amplitude maximale diffère suivant les consonnes. L'amplitude $AmpAspiConso(VE)$ est donnée au tableau 3.5 et est reliée à l'effort vocal VE par la formule réglée empiriquement :

$$AmpAspiConso(VE) = 0.75 + 0.25 \cdot VE^{15} \cdot AmpAspiConso_{Ref} \tag{3.2}$$

Occlusives	/p/	/b/	/t/	/d/	/k/	/g/
Amplitude d'aspiration	0.24	0.12	0.42	0.21	0.72	0.36
Fricatives	/f/	/v/	/s/	/z/	/ʃ/	/j/
Amplitude d'aspiration	0.03	0.015	0.06	0.03	0.12	0.06
Semi-voyelles		/w/		/ɥ/		/j//
Amplitude d'aspiration		0.012		0.012		0.06
Nasales		/m/		/n/		/ŋ/
Amplitude d'aspiration		0.012		0.012		0.06

TABLE 3.5 – *Amplitude maximale d'aspiration pour un effort vocal maximal (unité arbitraire)*

A part pour les semi-voyelles pour lesquelles l'aspiration est maximale sur la consonne tenue (là où la constriction est maximale sans être obstructive ou fricative), l'évolution temporelle de cette amplitude est réglée de manière identique pour toutes les consonnes comme indiquées par les courbes rouges sur la figure 3.8.

3.3 Modèle de contrôle : contrôle continu de la position articulatoire

Dans la partie ci-dessus a été décrit le modèle de production de notre synthétiseur de syllabes. Maintenant, nous traitons du modèle de contrôle de l'instrument, agissant sur des paramètres de la voix permettant d'en faire un instrument de musique. On vise donc à rendre possible une certaine expressivité, ce qui ne sera pleinement possible qu'en supprimant tout retard perceptible entre le geste et la production afin de se rapprocher du cas de la voix naturelle.

3.3.1 Une deuxième tablette graphique comme interface de contrôle articulatoire

Une première tablette graphique nous permet de contrôler les voyelles et la source glottique de la même manière que indiqué dans le Chapitre 2 dans le cas du contrôle mono-manuel.

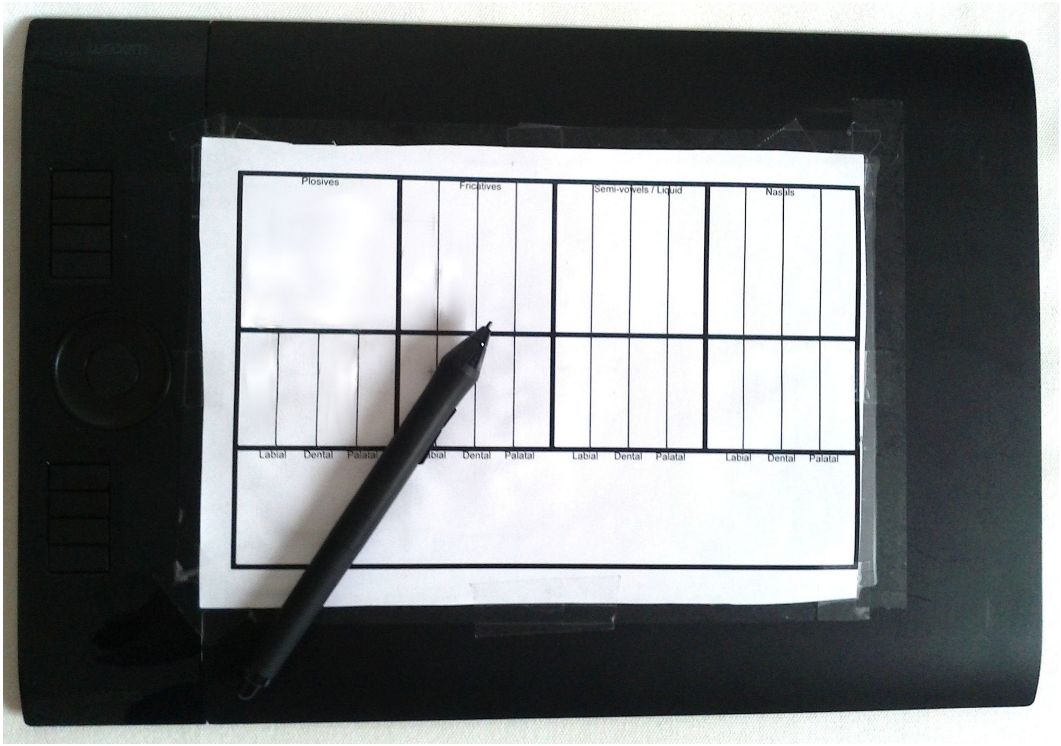


Figure 3.9 – *Tablette graphique avec son calque pour contrôler l'articulation VCV*

Une deuxième tablette graphique Wacom Intuos de type 5M est utilisée comme interface de contrôle des consonnes, en complément de la première tablette.

La tablette nous permet de capturer des gestes suffisamment rapides pour reproduire les gestes articulatoires. Le déplacement rapide du stylet sur la tablette sur de courtes distances peut se faire aisément en quelques dizaines de millisecondes et donc reproduire correctement la dynamique de l'articulation. D'autres contrôleurs ont été testés (trackpad, wiimote, accéléromètre) ou envisagés (capteurs optiques), mais n'ont jamais égalé la résolution, la latence ou la facilité d'utilisation de la tablette graphique.

Un modèle de contrôle constitué d'un pad de percussion électronique comme interface et de gestes de sélection pour le déclenchement de syllabes a été réalisé, mais il présente moins de qualité expressive étant donné ses gestes de sélection et non de contrôle.

Un calque est ajouté à la tablette (figure 3.9) afin d'avoir des repères de contrôle qui sont détaillés dans la partie suivante.

3.3.2 Paramètres de contrôle de haut-niveau

Dans le modèle de contrôle, pour un mode d'articulation donné (occlusives orales, fricatives, semi-voyelles, nasales, voisées ou non), on dispose de deux paramètres articulatoires principaux de haut-niveau (bornés entre 0 et 1) :

– $x2_{conCible}$ correspondant à la position relative du lieu d'articualtion de la consonne tenue cible, calculée par interpolation des valeurs de paramètres des deux consonnes tenues de référence les plus proches sur un axe correspondant au lieu d'articulation, et pour un même mode d'articulation. Ces paramètres correspondent à la valeur des formants consonantiques cibles (fréquence centrale $F_{conCible}$, amplitude $A_{conCible}$ et bande-

passante $B_{conCible}$), des coefficients de filtres modélisant les bruits consonantiques, ainsi que la valeur de paramètres caractérisant l'évolution temporelle du voisement, de l'amplitude des bruits et de l'aspiration. Par exemple, les filtres formantiques sont caractérisés ainsi :

$$\begin{aligned}
F_{conCible}(x2_{conCible}) &= F_{conRefAnt}(x2_{conCible}) \\
&\quad + x2_{conCible} \cdot [F_{conRefPost}(x2_{conCible}) - F_{conRefAnt}(x2_{conCible})] \\
A_{conCible}(x2_{conCible}) &= A_{conRefAnt}(x2_{conCible}) \\
&\quad + x2_{conCible} \cdot [A_{conRefAnt}(x2_{conCible}) - A_{conRefPost}(x2_{conCible})] \\
B_{conCible}(x2_{conCible}) &= B_{conRefAnt}(x2_{conCible}) \\
&\quad + x2_{conCible} \cdot [B_{conRefAnt}(x2_{conCible}) - B_{conRefPost}(x2_{conCible})]
\end{aligned} \tag{3.3}$$

où :

$F/A/B_{conRefAnt}(x)$ est une fonction qui renvoie la valeur des fréquences centrales / amplitudes / bandes-passantes de la consonne de référence dont le lieu d'articulation est antérieure à $x2_{conCible} = x$

$F/A/B_{conRefPost}(x)$ est une fonction qui renvoie la valeur des fréquences centrales / amplitudes / bandes-passantes de la consonne de référence dont le lieu d'articulation est postérieure à $x2_{conCible} = x$

Le mode d'articulation et le voisement étant connus, $x2_{conCible}$ permet de déterminer la consonne en lui donnant le lieu d'articulation.

– $y2_{articu}$, correspondant à la phase d'articulation relative, c'est-à-dire à la position articulatoire entre les deux états « voyelle cible » et « consonne cible tenue », agissant notamment sur les valeurs formantiques du son produit (fréquence centrale $F(x2_{conCible}, y2_{articu})$, amplitude $A(x2_{conCible}, y2_{articu})$ et bande-passante $B(x2_{conCible}, y2_{articu})$) :

$$\begin{aligned}
F(x2_{conCible}, y2_{articu}) &= y2_{articu} \cdot [F_{conCible}(x2_{conCible}) - F_{voyCible}] + F_{voyCible} \\
A(x2_{conCible}, y2_{articu}) &= y2_{articu} \cdot [A_{conCible}(x2_{conCible}) - A_{voyCible}] + A_{voyCible} \\
B(x2_{conCible}, y2_{articu}) &= y2_{articu} \cdot [B_{conCible}(x2_{conCible}) - B_{voyCible}] + B_{voyCible}
\end{aligned} \tag{3.4}$$

Sur les courbes 3.5, 3.6, 3.7, 3.8, $y2_{articu}$ correspond à la projection de la courbe noire « Formant interpolation » sur l'axe des abscisses.

D'autres paramètres sont contrôlés via $y2_{articu}$, comme l'évolution temporelle du bruit fricatif et occlusif, du taux d'aspiration, du voisement, et la nasalité par l'addition d'une anti-résonance en série.

Les coefficients des filtres reproduisant le bruit consonantique sont interpolés de la même manière que les formants afin d'obtenir une continuité entre chaque lieu d'articulation.

La figure 3.10 reprend schématiquement l'explication précédente pour le passage d'une voyelle cible intermédiaire à une consonne cible intermédiaire.

3.3.3 Correspondances entre paramètres de haut-niveau et la tablette graphique

Pour l'articulation entre consonne et voyelle ou entre deux consonnes de lieux d'articulation adjacents, on utilise la tablette graphique schématisée sur la figure 3.11 et contrôlée par la main secondaire, qui s'occupe de l'aspect consonantique avec les correspondances suivantes :

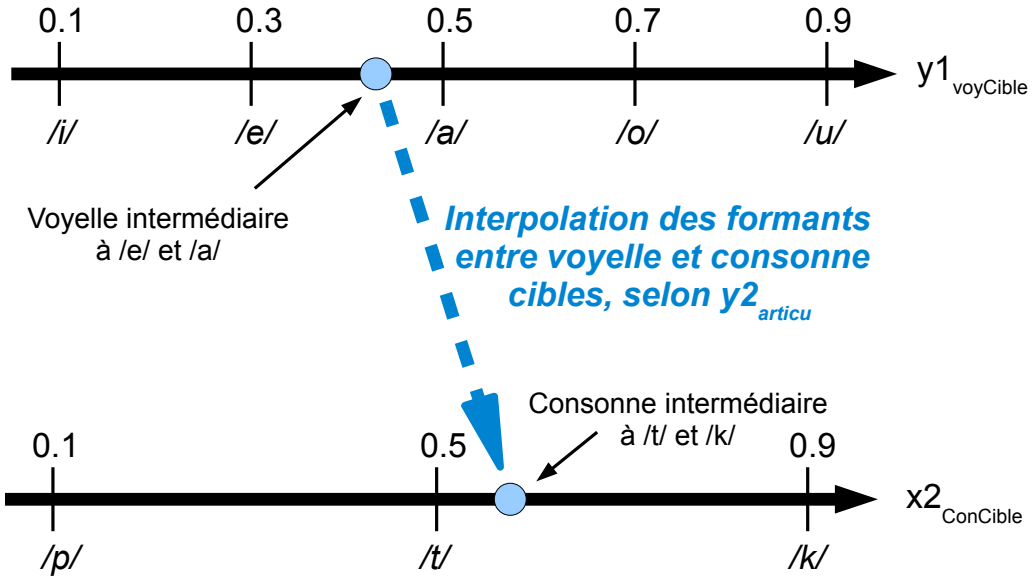


Figure 3.10 – *Exemple d'interpolation des formants entre voyelle et consonne cibles.*

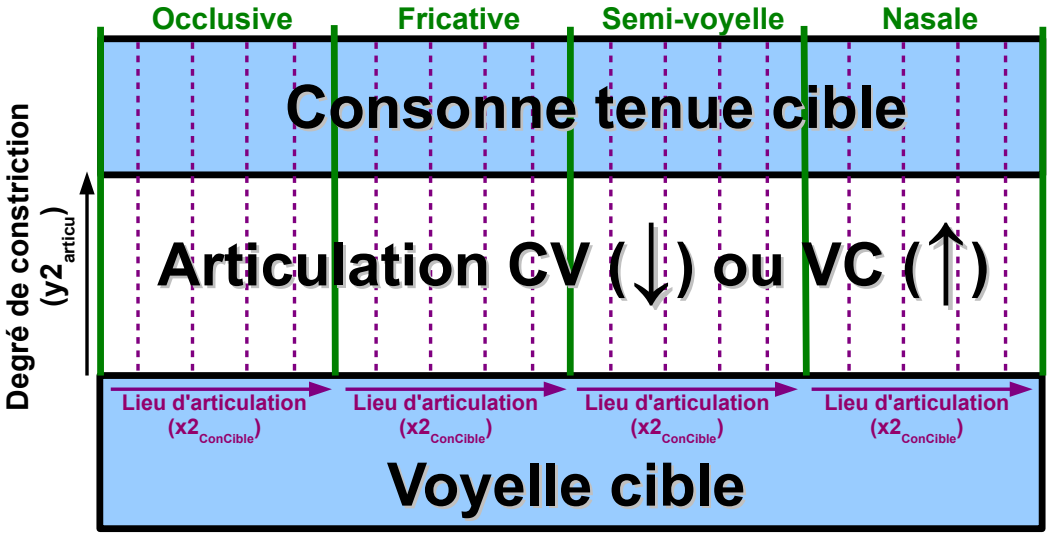


Figure 3.11 – *Représentation schématique des différentes zones de contrôle de la tablette vue de dessus*

- Le lieu d'articulation de la consonne avec la position du stylet suivant l'axe X de la tablette, avec des positions de référence correspondant aux consonnes du français, (lignes violettes verticales et en pointillés de la figure 3.11). Il correspond au paramètre $x2_{conCible}$ du modèle de contrôle.
- La phase d'articulation avec la position du stylet suivant l'axe Y (mouvement vertical dans la zone « Articulation CV ou VC » de la figure 3.11). Il correspond au paramètre $y2_{articu}$ du modèle de contrôle.
- Un facteur sur l'effort vocal avec la pression du stylet sur la tablette de la main secondaire (qui est désactivé en pratique pour limiter la latence)
- Le voisement de la consonne avec le bouton du stylet
- On a un mode d'articulation pour chaque région de la tablette, séparée suivant l'axe X (lignes vertes verticales de la 3.11)

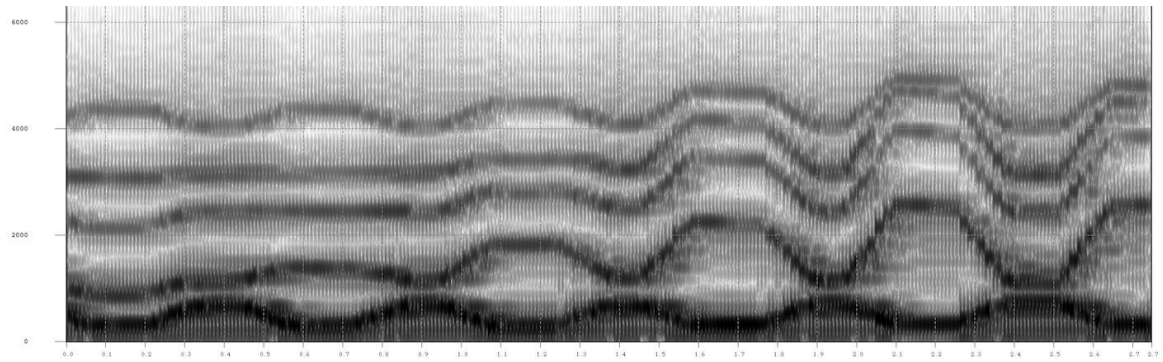
3.3.4 Visée et continuité des lieux d'articulation canonique

Pour l'instant, on a décrit notre système comme pouvant articuler exclusivement des consonnes de la langue française. Ici, on discute de l'interpolation réalisée pour avoir une continuité dans les lieux d'articulation.

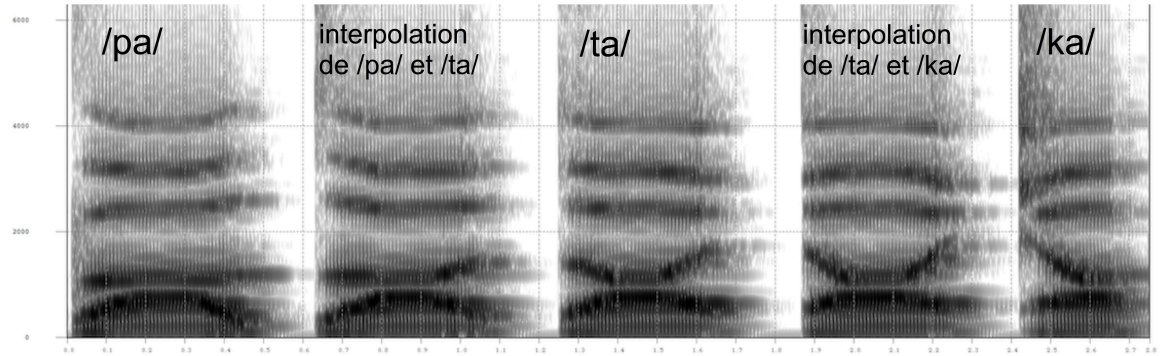
L'hypothèse proposée est qu'en première approximation, on peut faire une interpolation linéaire entre les lieux d'articulation canoniques du français, pour chaque mode d'articulation : bilabial, alvéolaire et palatal pour les occlusives ; labiodental, alvéolaire et post-alvéolaire pour les fricatives ; labio-vélaire, labio-palatal, palatal pour les semi-voyelles ; bilabial, alvéolaire, vélaire . A partir des valeurs des fréquence/amplitude/bande-passante de leurs formants, des valeurs des coefficients des filtres des bruits consonantiques et de l'amplitude d'aspiration, on peut produire des consonnes virtuelles intermédiaires sur l'axe labial-palatal. Du point de vue de la synthèse de voix humaine, l'hypothèse utilisée revient à interpoler ces valeurs pour obtenir des niveaux d'articulation intermédiaires. C'est peut être le cas en première approximation entre les occlusives alvéolaires et palatales par exemple, mais plus difficilement concevable entre les occlusives bilabiales et alvéolaires par exemple, vu la discontinuité entre ces deux lieux d'articulation. Enfin, du point de vue de la synthèse de voix pour la musique, cela permet en plus d'obtenir des sons à l'allure humaine mais non prononçable réellement. La figure 3.12 illustre cette continuité par des séquences VCV continues où la consonne passe de /w/ à /j/ en passant par /ɥ/ pour les semi-voyelles (a) et de /p/ à /k/ en passant par /t/ pour les occlusives (b).

Le voisement, ainsi que le mode d'articulation ne sont pas quant à eux contrôlables continûment. Cependant, en choisissant le mode fricatif/occlusif et en faisant varier la phase d'articulation autour de l'apparition du bruit fricatif/occlusif, on a en quelque sorte une continuité entre mode fricatif/occlusif et mode des semi-voyelles.

Le lieu d'articulation est un paramètre continu via $x2_{conCible}$. Avec une interpolation linéaire des consonnes, la visée des consonnes canoniques est assez difficile, puisque un léger décalage du stylet à droite ou à gauche tend à changer de consonne (1.4 cm sépare le centre de chacun des lieux d'articulation sur la tablette). Afin de pouvoir viser plus aisément les lieux d'articulation canoniques, on applique à $x2_{conCible}$ une non-linéarité par rapport au lieu d'articulation comme indiquée à la figure 3.13.



(a) *Semi-voyelles*



(b) *Occlusives orales*

Figure 3.12 – *Spectrogramme (*$0 - 6000$ *Hz) de successifs C-/a/ (C1-/a/-C2-/a/-C3-...) produits par le Digitartic, avec le lieu d'articulation évoluant sur l'axe bilabial - alveolaire - palatal*

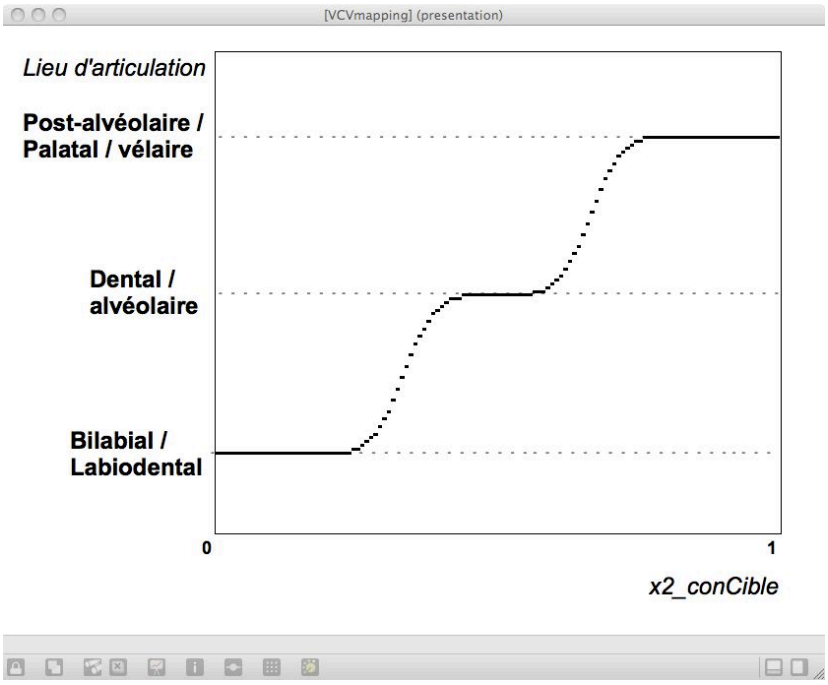
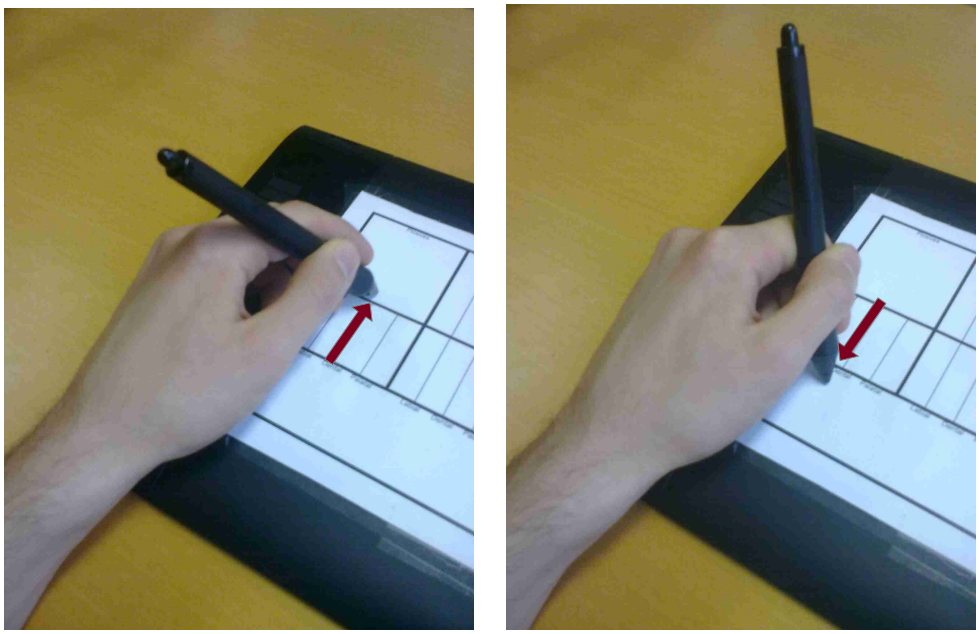


Figure 3.13 – *Non linéarité du paramètre d'interpolation des consonnes*



(a) *Position pour une consonne tenue* (b) *Position pour une voyelle*

FIGURE 3.14 – *Mouvements pour la production de syllabes VC et CV*

3.3.5 Dynamique du geste de contrôle de la phase d'articulation

Voir fichiers audios / vidéos 8

Une syllabe est le passage d'une configuration du conduit vocal à une autre. Ce passage est contrôlable via le paramètre $y2_{articu}$ comme détaillé dans la section 3.3.2.

Dans ce modèle de contrôle, nous choisissons de contrôler continûment, de manière réversible et sans latence perceptible le paramètre $y2_{articu}$ appelé *phase d'articulation*. Ce contrôle est réalisé à l'aide d'un des axes de la deuxième tablette destinée au contrôle consonantique, comme indiqué sur la figure 3.11. On peut produire n'importe quelle séquence du type $C_1V_1V_2...V_NC_2V_{N+1}V_{N+2}...V_PC_KV_{P+1}...V_QC_L$ ou toute sous-partie de cette séquence. D'autres séquences plus anecdotiques sont possibles et sont décrites dans la section 3.3.7.

La distance séparant la zone de consonne tenue de celle de la voyelle est identique quels que soient le mode et le lieu d'articulation. Or la durée de transition entre une consonne tenue et une voyelle (et vice versa) dépend de la consonne et avant tout du mode d'articulation. Ainsi, cette durée va être contrôlée par le geste de l'utilisateur. La distance étant de 4 cm, le geste à effectuer ne nécessite pas de lever le poignet et peut être réalisé très rapidement. Dans le cas des occlusives naturelles, la transition est d'environ 40 ms. Le geste utilisé, par sa rapidité et le peu de distance à parcourir, ajouté à la résolution de la tablette (5 ms), permet de contrôler continûment la transition en temps réel. La figure 3.14 représente en (a) la position de la main pour une consonne tenue et en (b) celle pour une voyelle. Le geste (indiqué par une flèche rouge) se fait rapidement par un mouvement de l'index. En (a), le geste indiqué permet la production d'une syllabe VC tandis qu'en (b) il permet une syllabe CV.

La non-linéarité des trajectoires formantiques présentée en section 3.2.4 est modifiée par le geste qui contrôle la phase d'articulation, directement relié à la trajectoire formantique.

Le geste va modifier la dynamique de cette transition non-linéaire, donc ce geste doit être adapté à la dynamique des modes d'articulation, qu'on regroupe sous deux catégories :

- Les occlusives orales qui présentent un bruit d'explosion. Le geste devra être nécessairement rapide. En effet, la durée relative du bruit d'explosion a été réglée de façon à ce qu'elle ne dure que le temps d'une explosion pour une durée de transition de l'ordre de 40 ms. Si le geste est plus lent, le bruit ne paraîtra plus explosif mais plutôt fricatif du fait de sa longueur accrue. Autrement dit, ralentir le geste de production des occlusives de synthèse n'est pas bien corrélé au ralentissement du geste articulatoire naturel des occlusives qui voudrait que l'explosion ait lieu quel que soit sa vitesse d'articulation.
- Pour les fricatives, les semi-voyelles et les nasales, la vitesse du geste de production des consonnes de synthèse est bien corrélée avec le cas de la production naturelle. Ainsi, suivant le type d'articulation qu'on veut, on modifie la dynamique en conséquence. Mais quelle qu'elle soit, elle reste naturelle. Cependant, pour produire des syllabes se rapprochant de la voix parlée, on doit effectuer le geste de transition sur environ 60 à 80 ms.

Nous montrons qu'il est possible de reproduire la dynamique des trajectoires des formants de consonnes naturelles en comparant le spectrogramme de séquences VCV pour des séquences de synthèse contrôlées gestuellement et des séquences de voix naturelles. Sur la figure 3.15, nous comparons respectivement 3 occlusives de synthèse et naturelles, et 3 semi-voyelles de synthèse et naturelles. La voix naturelle a été enregistrée dans une chambre anéchoique, et la voix de synthèse a été produite en essayant d'imiter cette voix naturelle. Tous les spectrogrammes sont de longueurs 400 ms et sont échelonnés de 0 à 6000 Hz.

Ces spectrogrammes appellent les remarques suivantes :

- la dynamique des trajectoires est bien reproduite ;
- les amplitudes des formants F_4 à F_5 des consonnes sont de façon générale trop élevées par rapport à ce qu'on peut observer sur cette voix naturelle (figure 3.15). Mais leurs valeurs sont plutôt reliées à la personnalisation de la voix et ont peu de sens phonologique ;
- le bruit d'aspiration est trop long sur la phase VC de cet exemple.

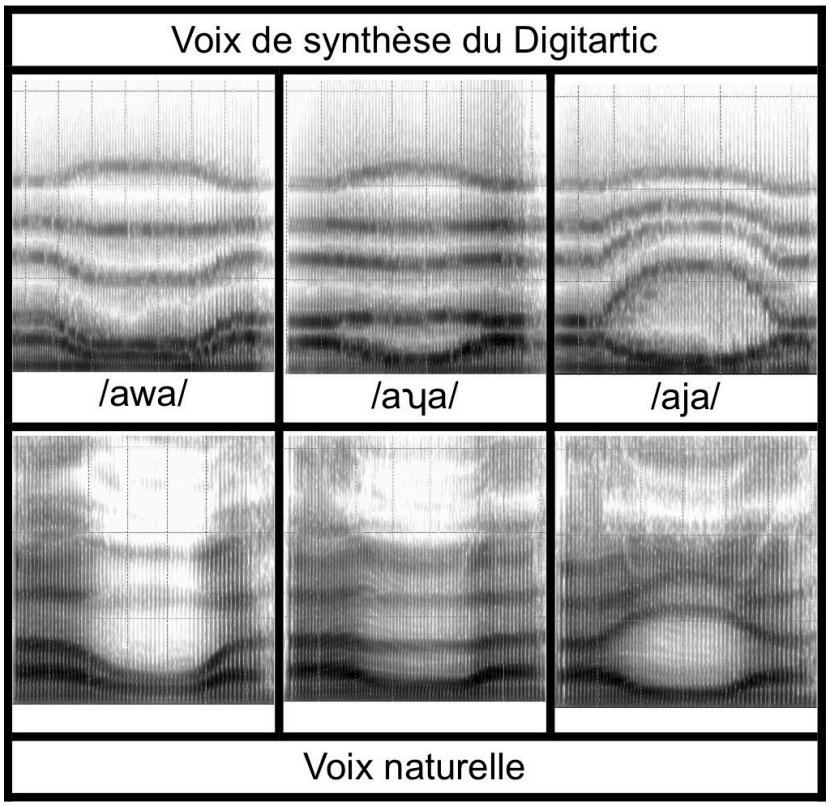
3.3.6 Contrôler l'hypo-articulation

Voir fichiers audios / vidéos 9

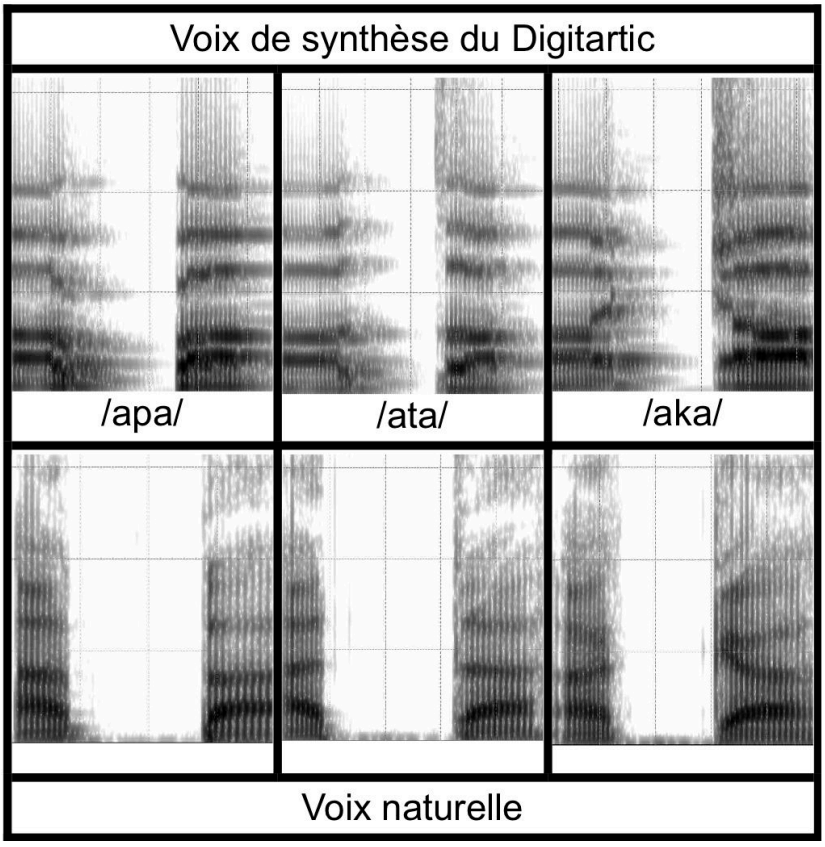
Si notre geste n'atteint pas la zone de position de la voyelle dans une syllabe CV par exemple, comme illustré par la figure 3.16, on produit alors une syllabe hypo-articulée sur la voyelle. En effet, le paramètre de phase d'articulation $y2_{consArticu}$ n'atteint pas sa valeur minimale donnant ainsi une voyelle dont les formants sont légèrement interpolés avec ceux de la consonne tenue du lieu et mode d'articulation visé (à comparer aux mouvements articulés normalement sur la Figure 3.14) .

Il en va de même pour le degré d'hypo-articulation des consonnes du *Digitartic* à l'exception des occlusives.

Les nasales, semi-voyelles et fricatives sont dans le *Digitartic* symétriques par rapport à la consonne tenue dans une séquence VCV. Toutes les caractéristiques de ces consonnes (formants, nasalité, bruit de friction) apparaissent progressivement dans la zone de transition. On peut donc hypoarticuler en visant un point dans la zone de transition, proche de la zone de la consonne tenue mais sans l'atteindre, comme indiqué à la figure 3.17

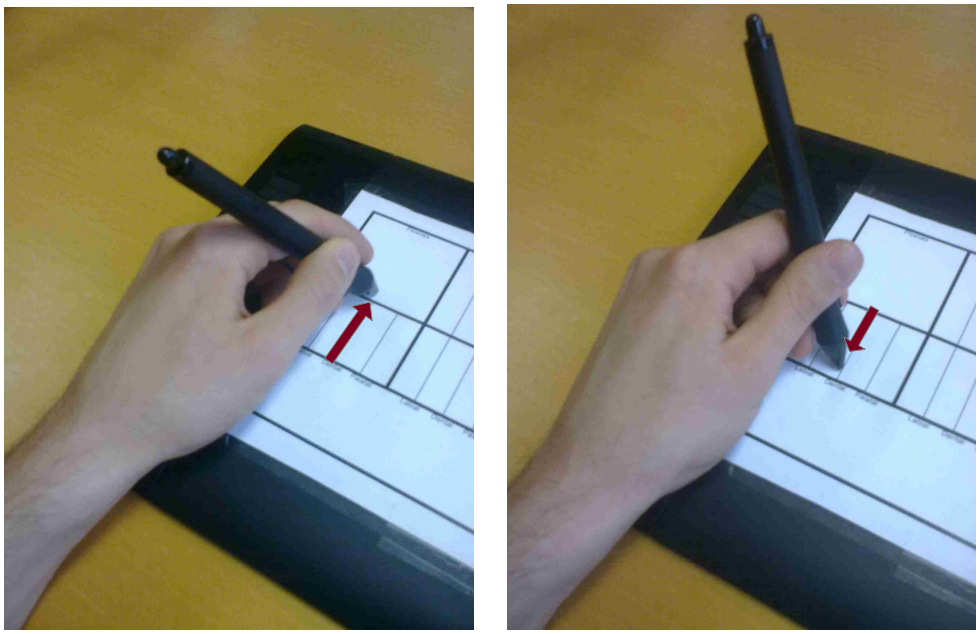


(a) *Semi-voyelles*



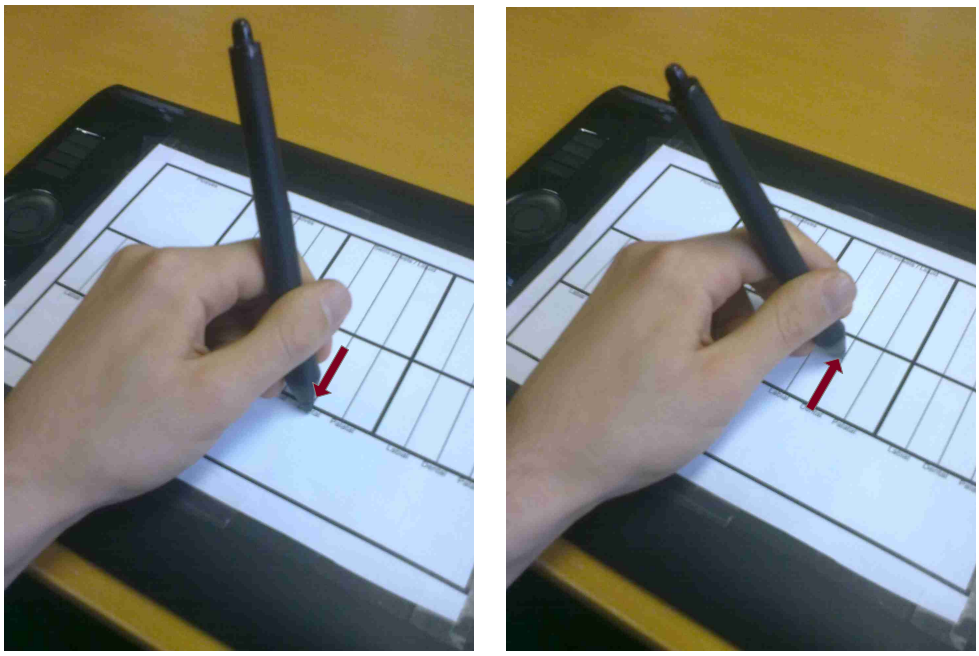
(b) *Occlusives*

FIGURE 3.15 – *Spectrogramme de séquences VCV du Digitartic et de voix naturelles (*$0-6000$ *Hz)*



(a) *Position pour une consonne (tenue)* (b) *Position pour une voyelle hypoarticulée*

Figure 3.16 – *Mouvements pour la production de syllabes VC et CV, dans le cas d'une syllabe avec voyelle hypo-articulée. La zone de la voyelle n'est pas atteinte*



(a) *Position pour une voyelle* (b) *Position pour une consonne tenue hypo-articulée*

Figure 3.17 – *Mouvements pour la production de syllabes VC et CV, dans le cas d'une syllabe avec consonne hypo-articulée. La zone de la phase médiane de la consonne n'est pas atteinte*

Pour produire l'explosion, le geste doit d'abord atteindre la zone de la tablette « consonne tenue » correspondant à l'occlusion, puisque l'explosion n'intervient pas dans le sens VC (comme indiqué par les courbes jaunes de la figure 3.7). L'hypo-articulation peut être marquée d'une autre façon, comme pour les autres consonnes : en abaissant l'effort vocal arrivé sur la consonne tenue, baissant alors le niveau sonore de l'attaque de la consonne.

L'analogie entre gestes articulatoires de l'appareil vocal et gestes articulatoires manuels est renforcé dans le sens où une hypo-articulation des gestes manuels se traduit naturellement par une hypo-articulation des syllabes synthétisées. Le trajet à effectuer manuellement ayant un sens assez physique (il correspond à la phase d'articulation), des gestes manuels rapides vont encourager à ne pas atteindre les zones de la tablette cibles et produire des syllabes hypoarticulées.

3.3.7 Séquences VCCV

A l'aide de ce modèle de contrôle, même s'il n'est pas conçu à la base pour cela, on peut produire quelques successions de consonnes de type VCCV. Les possibilités sont les suivantes :

- enchainer deux consonnes fricatives (quel que soit leur voisement), deux semi-voyelles ou deux nasales en déplaçant le stylet suivant l'axe du lieu d'articulation dans la zone d'un même mode d'articulation. L'effet sera d'autant plus réussi que les deux consonnes sont de lieux d'articulation proches afin d'éviter de devoir passer par une consonne tenue intermédiaire lors de l'interpolation entre les deux consonnes. L'interpolation se fait d'une manière continue et contrôlable en temps réel comme pour l'articulation CV ou VC
- pour les occlusives, la consonne tenue est un silence quel que soit le lieu d'articulation et le voisement ; et la consonne ne se caractérise complètement que lors de l'explosion (phase CV), contrairement aux autres modes d'articulation où la consonne se caractérise complètement avec la phase VC (car symétrique par rapport à la consonne tenue). Donc changer de lieu d'articulation sur une occlusive tenue aura pour effet d'avoir une séquence VCV où la phase de resserrement de l'obstruction correspond à la première consonne visée et où la phase de relâchement (dont bruit d'explosion) correspond à la deuxième consonne visée. L'enchaînement articulatoire est possible naturellement

Il n'est pas possible d'avoir des enchainements de consonnes qui appartiennent à des modes d'articulation différents, à moins d'avoir une voyelle entre les deux consonnes.

Un exemple d'articulation VCCV manuelle est donné à la figure 3.18 avec V-/mn/-V.

3.4 Modèle de contrôle alternatif utilisant des gestes de sélection pour le déclenchement des phases VC et CV

Ici est présenté un mode de contrôle alternatif du *Digitartic* où la transition des voyelles et des consonnes (et vice versa) n'est plus contrôlée continûment et en temps réel mais par sélection (i.e. déclenchement). La sélection des consonnes se fait à l'aide d'un système de doigtés inspirés de ceux d'une percussion indienne, les *tablas*. On utilise deux interfaces multitouch (positions de 5 doigts et surface de ces doigts) pour chacune des mains. La deuxième différence avec le premier modèle de contrôle est que le contrôle de F_0 devient secondaire et que l'attention est focalisée sur la production de consonnes, comme des récitations d'onomatopées. Ce doigté a été peu pratiqué, surtout à cause de problèmes de latence avec l'interface tactile multitouch trackpad.

Figure 3.18 – *Trajectoire du stylet de la tablette secondaire pour obtenir la séquence V-mn-V*

La figure 3.19 donne un aperçu du fonctionnement du Digitartic avec des trackpads comme interface *multitouch*.

3.4.1 Mapping entre les paramètres de contrôle du modèle et l'interface multitouch

A chacune des mains est associée une interface *multitouch*. Le premier contrôle concerne plutôt ce qui a trait au voisement (main secondaire), et le deuxième à l'articulation (main préférée). Sur la figure 3.20, une des deux interfaces est un trackpad intégré à l'ordinateur portable, l'autre est un trackpad externe envoyant les données à l'ordinateur par ondes radio (*Bluetooth*). Les données sont récupérées par l'external *fingerpinger*[2] pour Max/MSP.

L'interface associée à la main secondaire contrôle avec l'index la voyelle suivant la position sur l'axe Y, la hauteur mélodique suivant l'axe X, l'effort vocal avec la surface en contact, et avec l'annulaire (ou tout autre doigt) le voisement des consonnes (booléen), comme illustré sur la figure 3.21 qui représente le trackpad vu de dessus avec ces informations.

Les Figures 3.22 et 3.23 représentent de deux manières équivalentes le trackpad associé à la main préférée, qui contrôle l'articulation. A l'aide de l'index (ou alterné avec le majeur pour plus de rapidité), on contrôle la structure temporelle de l'articulation : le contact du doigt déclenche la transition VC (1) et le lieu de contact suivant l'axe Y est lié au lieu d'articulation (continu) ; la durée de tenue du doigt sur le trackpad correspond à la durée de la tenue de la consonne (par exemple, bruit fricatif sur /s/ ou silence sur /p/), et la surface de doigt S_c contrôle son niveau sonore dans le cas des fricatives et des semi-voyelles (2) ; le retrait

2. Fingerpinger 2009 by Michael & Max Egger, http ://www.anyma.ch/2009/research/multitouch-external-for-maxmsp/

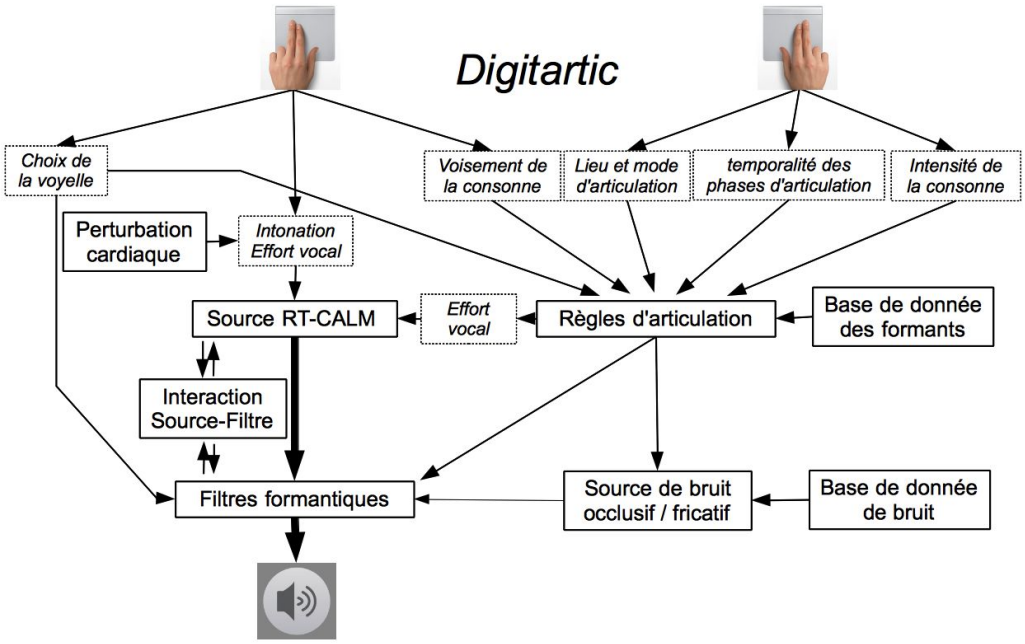


FIGURE 3.19 – *Représentation schématique du fonctionnement du synthétiseur Digitartic*

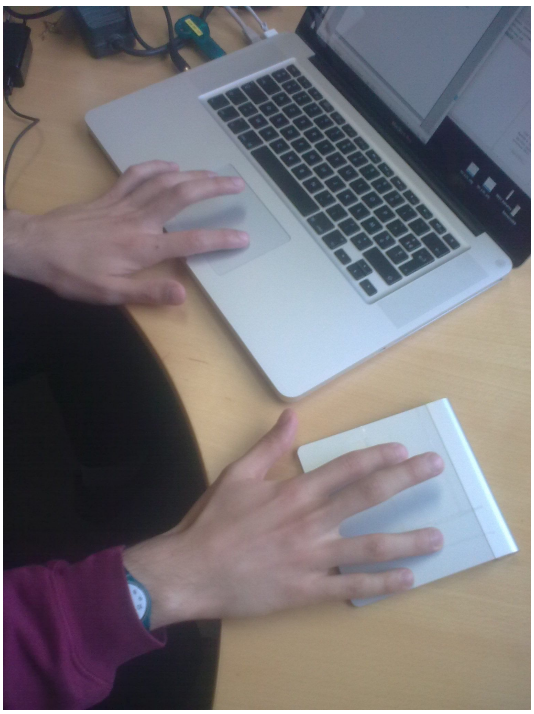


FIGURE 3.20 – *Interface de contrôle du Digitartic, constitué de deux trackpads (configuration droitier)*

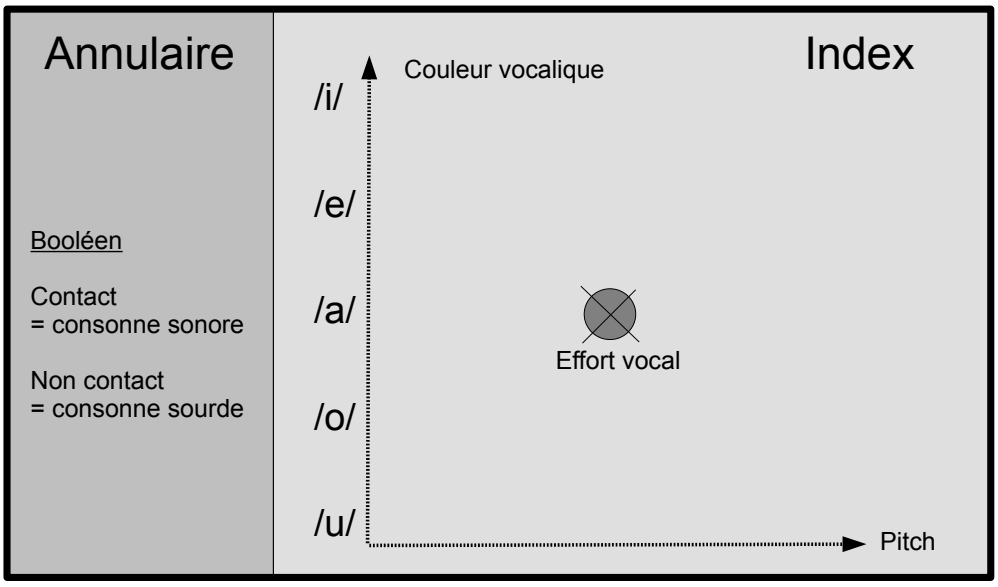


Figure 3.21 – *Agencement spatial du trackpad de la main secondaire*

du doigt déclenche la transition CV (3), dont le niveau sonore est déterminé par la dernière surface S_{cv} non nulle enregistrée par le trackpad, et par la dernière position du doigt suivant l'axe Y pour le lieu d'articulation.

La représentation b) de l'agencement spatial du trackpad de la main préférée (figure 3.23) montre les consonnes utilisables avec ce système. On remarque qu'on retrouve toutes les consonnes du français exceptées les consonnes nasales, la liquide /l/ et la fricative ʁ.

La figure 3.24 illustre concrètement par trois photos les positions des doigts pour contrôler l'articulation. On commence par la main préférée, qui contrôle lieux et modes d'articulation, le déclenchement des phases VC et CV, ainsi que l'intensité de ces transitions. Les trois photos montrent trois consonnes en phase de tenue avec des modes d'articulation distincts. L'index est maintenu en contact pour tenir la consonne, seuls l'auriculaire et l'annulaire changent de configuration : la première photo de la Figure 3.24 concerne une occlusive car aucun de ces deux doigts n'est en contact avec le trackpad ; la seconde montre l'annulaire en contact, on a donc une fricative ; enfin la troisième représente une semi-voyelle, ces deux doigts étant en contact avec le trackpad.

La figure 3.25 illustre le changement de lieu d'articulation pour un mode d'articulation donné (ici fricatif car l'annulaire est en contact), par le changement de position de l'index suivant la profondeur du trackpad. Un index situé en avant correspond à un lieu d'articulation antérieure, alors que un index situé en arrière correspond à un lieu d'articulation postérieur.

La figure 3.26 illustre le changement de lieu d'articulation comme précédemment mais avec pour mode d'articulation les semi-voyelles, c'est à dire avec à la fois l'annulaire et l'auriculaire en contact avec le trackpad. On remarque ici contrairement aux illustrations précédentes que l'annulaire et l'auriculaire se courbent pour atteindre les lieux d'articulation antérieur : c'est une position alternative qui est équivalente, car seul compte le bout du doigt qui est en contact.

La figure 3.27 représente deux mêmes configurations de consonne fricative alvéodentale mais où la tenue de la consonne est réalisée à l'aide de doigts différents (résultat équivalent).

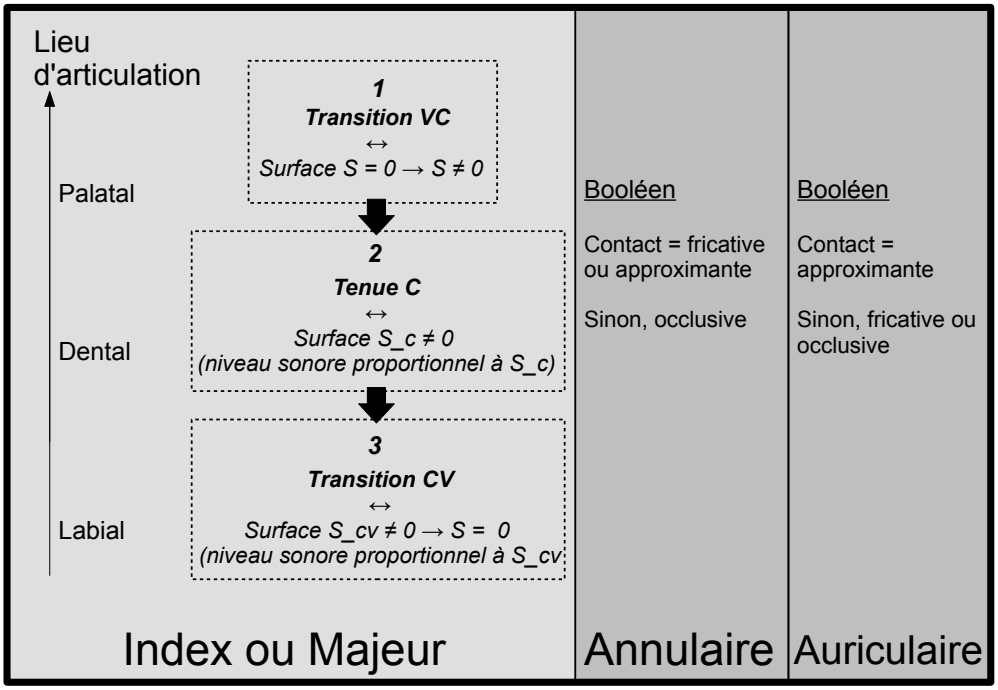


FIGURE 3.22 – *Agencement spatial du trackpad de la main préférée - représentation a)*

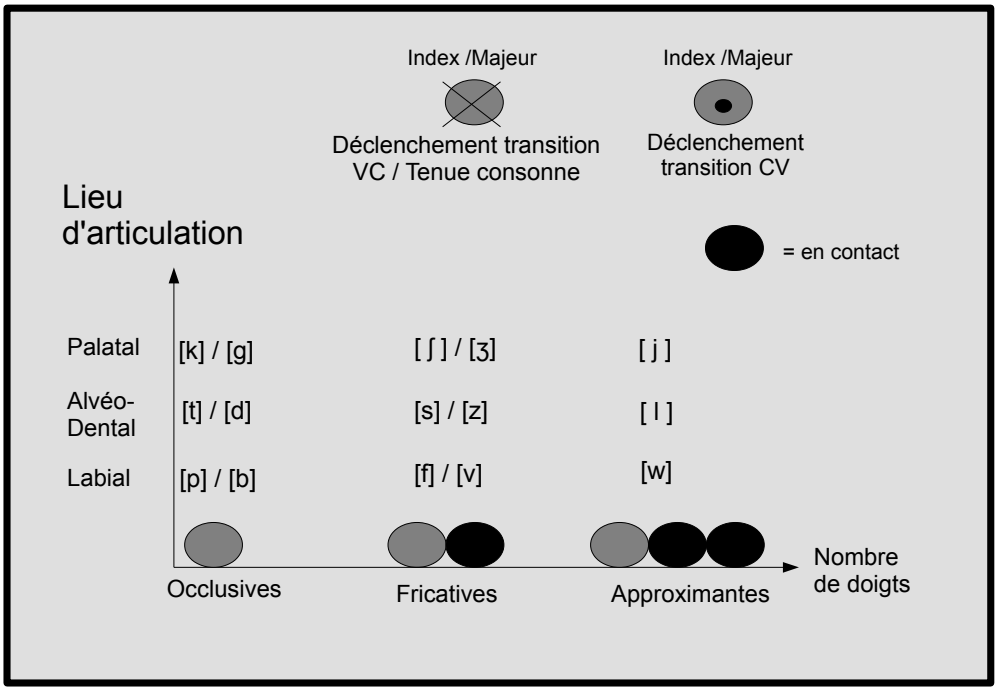
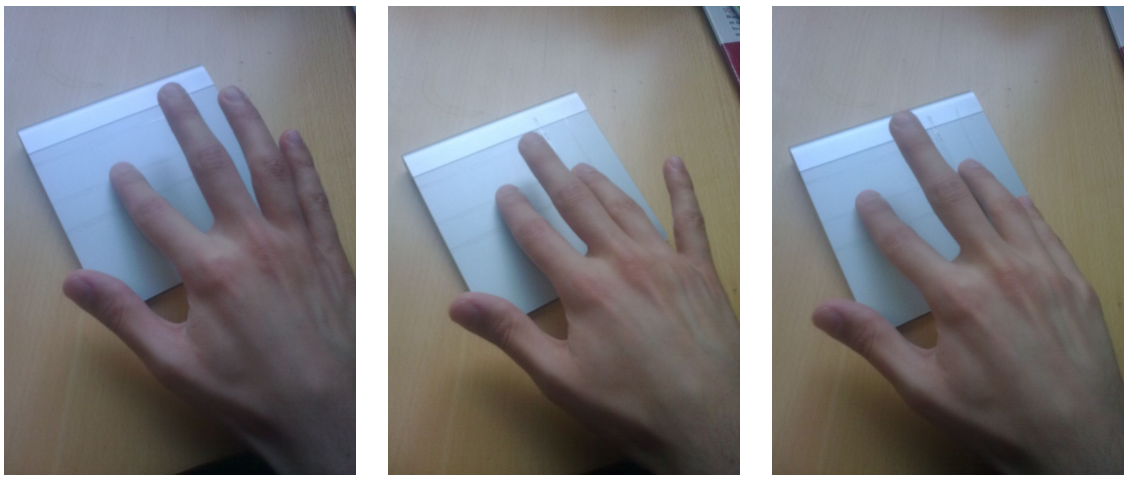
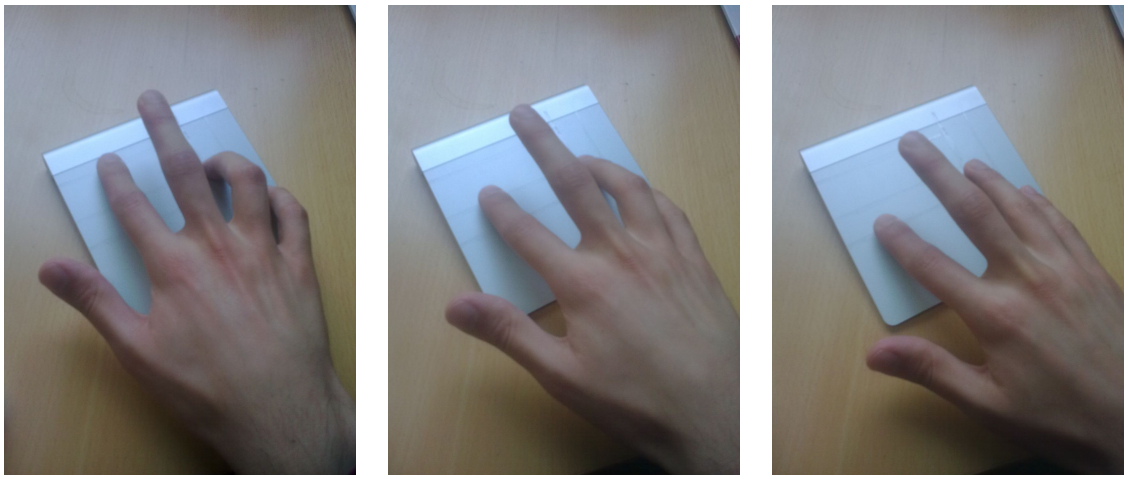


FIGURE 3.23 – *Agencement spatial du trackpad de la main préférée - représentation b)*



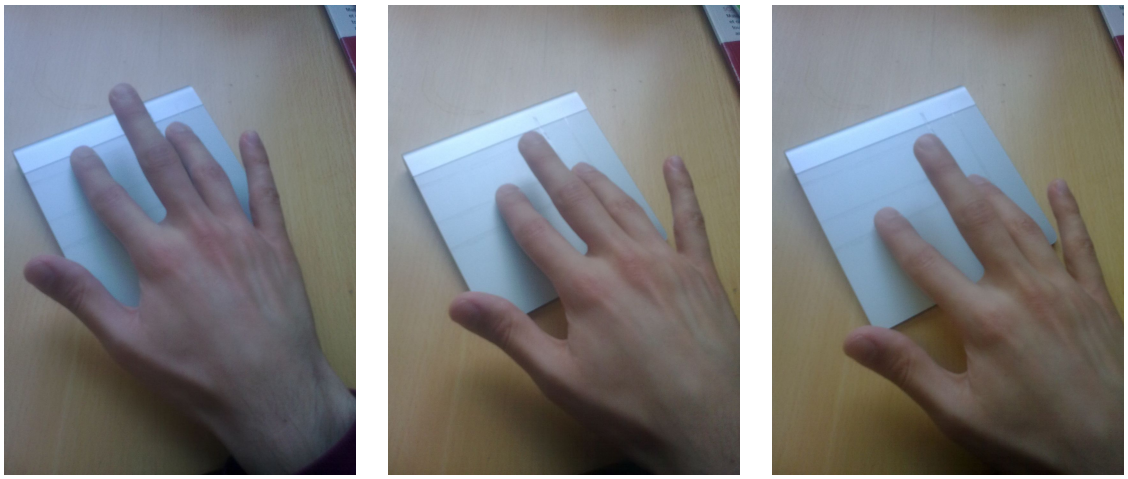
(a) *Pour une occlusive tenue* (b) *Pour une fricative tenue* (c) *Pour une semi-voyelle tenue*

Figure 3.24 – *Position de l'annulaire et l'auriculaire de la main préférée pour des consonnes tenues de différents modes d'articulation*



(a) *Pour une fricative labiale tenue* (b) *Pour une fricative alvéolaire tenue* (c) *Pour une semi-voyelle palatale tenue*

Figure 3.25 – *Position de l'annulaire et l'auriculaire de la main préférée pour des consonnes fricatives tenues de différents lieux d'articulation*



(a) *Pour une semi-voyelle labiale tenue*

(b) *Pour une semi-voyelle alvéolaire tenue*

(c) *Pour une semi-voyelle palatale tenue*

FIGURE 3.26 – *Position de l'annulaire et l'auriculaire de la main préférée pour des semi-voyelles tenues de différents lieux d'articulation*

Cela permet de déclencher deux syllabes identiques rapidement en alternant l'index et le majeur.

Enfin, la figure 3.28 montre, quant à elle, la main secondaire qui s'occupe des voyelles et du caractère voisé des consonnes. La position de l'index sur l'axe suivant la profondeur détermine la voyelle et le contact de l'annulaire avec le trackpad correspond à une prochaine consonne voisée.

Pour bien comprendre ces figures, il faut se reporter aux représentations schématiques 3.21, 3.22 et 3.23.

3.4.2 Structure temporelle du modèle : les différentes phases de VCV

Ce modèle de contrôle permet de contrôler une séquence « V1 - C - V2 » aux moments suivants : tout au long de la voyelle V1, au début de la transition V1-C, tout au long de la tenue de la consonne, et au début de la transition C-V2.

La transition Voyelle-Consonne (VC)

Avant le début de la transition, l'utilisateur choisit le mode d'articulation, la voyelle de départ, le caractère sourd ou sonore de la consonne, la hauteur mélodique et l'effort vocal par une configuration de position des doigts comme décrit ci-dessus. Ensuite, à l'aide de l'index ou du majeur de la main préférée, l'utilisateur va déclencher la transition en posant son doigt sur le lieu d'articulation désiré. A partir de là, rien n'est contrôlable jusqu'à la phase suivante de l'articulation, la tenue de la consonne. L'ensemble de ces paramètres caractérise la consonne désirée et envoie les valeurs des formants cibles et la durée de transition.

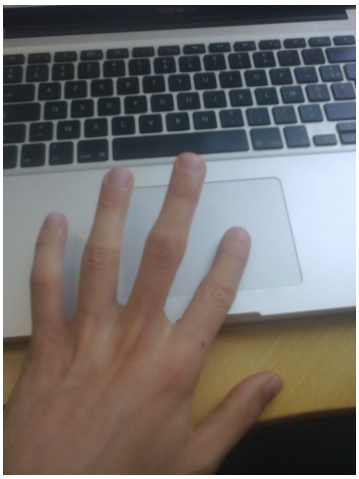
La tenue de la consonne centrale

L'utilisateur peut changer de lieu d'articulation pendant la tenue, ainsi que son niveau sonore. Il peut également changer de mode d'articulation, mais sans coarticulation entre les deux modes.

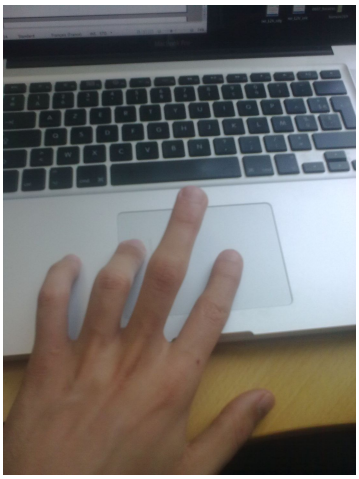
(a) *Fricative alvéodentale tenue avec l'index*

(b) *Fricative alvéodentale tenue avec le majeur*

Figure 3.27 – *Stratégie pour enchaînement de deux syllabes identiques, en alternant index et majeur*



(a) *prochaine consonne sourde*



(b) *prochaine consonne sonore*

Figure 3.28 – *Position de l'index/annuaire de la main secondaire pour la voyelle /a/ et selon le voisement de la prochaine consonne*

La transition Consonne-Voyelle (CV)

La transition CV est le retour à la voyelle d'origine ou le chemin vers une nouvelle. La dernière configuration de position des doigts détermine les traits de la consonne de la transition CV. Si l'on veut produire une syllabe CV seule, les gestes adéquates sont d'abord de se placer en configuration C puis en configuration CV.

3.4.3 Contrôle de l'articulation et temps réel : durée des transitions

Dans la partie précédente, on a vu quels stades de l'articulation était déclenchés. On contrôle le déclenchement d'une consonne connaissant la voyelle sur laquelle on part (VC) ou le déclenchement d'une voyelle connaissant la consonne sur laquelle on part (CV). On traite ci-dessous de sa relation avec le contrôle temps réel pour des applications musicales.

Notre but est d'avoir un lien temporel très proche entre le geste et le rendu audio. En musique, l'accent rythmique est porté sur le bruit d'explosion de la consonne dans le cas des syllabes CV avec des occlusives, mais sur la voyelle dans le cas des syllabes CV avec des fricatives ou des semi-voyelles. On prend en compte cette particularité dans ce modèle dans la limite suivante : le déclenchement de la phase CV (effectuée par la levée du doigt) enclenche le début de la transition et non le début de la voyelle à proprement parler. Le retard entre le geste et la voyelle est de la durée de la transition, soit entre 25 et 35 ms chez les occlusives et fricatives, et de 30 à 50 ms chez les semi-voyelles. Or, on estime qu'un retard est perceptible à partir de cet ordre de grandeur de temps. Dans le modèle, on a un paramètre pour contrôler globalement tous les temps de transitions, qui pourrait être utilisé dans le jeu. Bien qu'il permette de réduire ce retard en l'abaissant, il rendrait aussi l'articulation moins nette. La durée de transition doit être suffisamment petite pour ne pas percevoir de retard, mais suffisamment grande pour bien percevoir l'articulation. Il faut donc anticiper le geste dans le jeu musical pour être synchronisé avec un tempo extérieur.

A ce retard lié au modèle, s'ajoute le retard dû à la récupération des données de l'interface *multitouch*, d'autant plus important que le nombre de doigts en contact est grand (au moins égale à 20 ms avec plusieurs doigts, 8 ms avec un seul selon les développeurs de l'external Max/MSP [3]).

3.5 Résumé et conclusion

L'instrument *Digitartic* démontre qu'il est possible de contrôler précisément l'articulation de consonnes et de voyelles à l'aide de gestes manuels. L'analogie entre gestes articulatoires dans l'appareil vocal et gestes manuels est encourageant.

Le lieu et la phase d'articualtion peuvent être contrôlés continûment en temps réel sans latence perceptible. Cette caractéristique peut être utilisée pour jouer des syllabes articulées expressives et réactives dans un contexte musical de voix chantée, de scat, ou de récitation d'onomatopées (voir démonstration audio 10).

Le modèle de contrôle principal permet de reproduire n'importe quelle séquence VCV, parmi toutes les consonnnes et voyelles du français, à l'exception des voyelles nasales, de la liquide /l/ et de la fricative /ʁ/. Cependant, les cibles formantiques des consonnes sont réglées dans le contexte vocalique /a/-C-/a/ : certaines consonnes ne sonnent pas naturelles si elles sont suivies ou précédées d'une voyelle autre que /a/.

3. http ://www.anyma.ch/2009/research/multitouch-external-for-maxmsp/

En plus des comparaisons qualitatives des spectrogrammes de voix de synthèse et naturelles que nous avons établies, des tests plus formels doivent être entrepris pour évaluer la qualité de synthèse des syllabes. Un meilleur réglage de la coordination temporelle des paramètres de production consonantique, ainsi que l'amélioration des bruits fricatifs et occlusifs doivent précéder l'élaboration des tests.

Par contre, des tests formels peuvent être entrepris en ce qui concerne les possibilités d'articulation expressive (degré et dynamique d'articulation). Nous comptons pouvoir montrer une analogie entre la dynamique gestuelle manuelle et des articulateurs, comme cela a été réalisé entre l'intonation et les gestes manuels par d'Alessandro, Rilliard et Le Beux [dRLB11].

Digitartic est peu à peu introduit dans la chorale Chorus Digitalis que nous présentons au chapitre D, agrandissant les capacités de notre chorale.

Deuxième partie

JEUX INDIVIDUELS ET COLLECTIFS : ANALYSE ET ÉVALUATION

Avant-propos

Parmi les instruments numériques de synthèse vocale qui ont atteint le stade de la représentation publique, on peut citer les suivants. Le *SqueezeVox* [CL00], interface basée sur un accordéon auquel on a remplacé les éléments responsables du son de l'accordéon par différents capteurs de la soufflerie et des touches de l'accordéon, et augmenté de capteurs tactiles et d'un écran de contrôle. Une large variété de modèles de production vocale est contrôlable. Une vidéo de l'instrument au sein du *Virtual Augmented Chorale (VAChorale)*[4], est extraite d'un concert en 2001 au JBL Theater dans le cadre de la conférence *New Interfaces for Musical Expression*, est consultable en ligne[5]. Le *Voicer*[6] [Kes02], utilisant la synthèse par formant et une interface composée d'un joystick et d'une tablette graphique, est joué dans le morceau *Ici et Ailleurs* du groupe *Tutti Quanti Computing Orchestra*. Deux extraits de concert de 2002, qui montrent le mariage réussi entre instruments acoustiques et cet instrument numérique, sont consultables sur le site du groupe[7][8]. Le *DIVA*[9] [PFd+11], dérivé de l'instrument GloveTalkII utilise la synthèse par formant et un modèle de contrôle reconnaissant des gestes issus de gants haptiques. Cette vidéo[10], mise en ligne en 2010, nous montre une musicienne jouant à la fois du DIVA et de sa propre voix. L'instrument *HandSketch*[11] [dD07], permet de contrôler différents modèles de synthèse. Un duo avec un violoncelliste, des Journées d'Informatique Musicale en 2012 est visible en ligne[12] L'instrument *ChoirMob*[13] [dPW+12] composé d'un téléphone portable tactile contrôlant le playback d'une partition de voix produit par de la synthèse par formant. Un concert datant de novembre 2011 est écoutable sur le lien donné en bas de page[14].

Parmi les orchestres numériques, on peut citer le *Plork Laptop Orchestra*[15] [TCSW06], composé d'une dizaine de musiciens chacun muni d'un ordinateur et d'un haut-parleur dédié, mais n'abordant pas le contrôle gestuel de la voix de synthèse. La VAChorale citée plus haut rassemble quant à elle plusieurs voix de synthèse, mais celles-ci sont contrôlées par la même personne, alternativement. De notre coté, nous proposons la première chorale de voix de synthèse (à notre connaissance) où chacune des voix est contrôlée par un musicien. Une

4. http://voce.cs.princeton.edu/, consulté le 21/02/2013
5. http://voce.cs.princeton.edu/VAChorale/perryEMPShort.mov
6. http://tqco.free.fr/
7. http://tqco.free.fr/videos/wb_extrait_ici.mpg
8. http://tqco.free.fr/videos/cjulien_intro_ici.mpg
9. http://www.magic.ubc.ca/artisynth/pmwiki.php?n=VisualVoice.Design
10. http://vimeo.com/8983689
11. http://www.numediart.org/projects/06-4-handsketch/
12. http://vimeo.com/45080127
13. http://www.nicolasdalessandro.net/choirmob/
14. https://soundcloud.com/aura-pon/sets/intertwine-for-mobile-device
15. http://plork.cs.princeton.edu/

autre chorale de voix de synthèse, le VoxTactum [16], composée des instruments *ChoirMobs* a donné son premier concert 8 mois après notre premier concert.

Dans la partie I, nous avons décrit les instruments de synthèse de voix *Cantor Digitalis* et *Digitartic*. Nous présentons maintenant l'utilisation de ces instruments dans un contexte musical, en analysant d'abord les gestes instrumentaux des musiciens de l'ensemble de voix de synthèse *Chorus Digitalis*. Cet ensemble musical avec sa configuration, son répertoire et ses concerts et son organisation, est présenté dans l'annexe D. Ensuite, l'interface de l'instrument sera évaluée en mesurant la capacité des musiciens à jouer juste lors de tâches individuelles. Enfin, une étude préliminaire sera discutée, visant à mesurer leur faculté à jouer juste en groupe avec le *Chorus Digitalis*.

16. http://voxtactum.com/

Chapitre 4

Les gestes instrumentaux du *Cantor Digitalis* et du *Digitartic*

Sommaire

Dans la partie I, nous avons décrit les instruments de synthèse de voix chantée *Cantor Digitalis* et *Digitartic*. Dans ce présent chapitre, nous analysons l'utilisation de ces instruments dans un contexte musical, plus particulièrement en traitant les gestes des musiciens.

Nous commencerons par décrire l'ensemble des gestes utilisés avec nos deux instruments suivant une approche fonctionnelle des gestes musicaux. Après avoir explicité la manière de les enregistrer, nous décrirons ensuite les gestes du *Cantor Digitalis* et du *Digitartic* par une approche phénoménologique en les séparant suivant deux critères fonctionnels, à savoir les gestes nécessaires à la production du son, et ceux qui ne le sont pas, tout en étant présents chez les instrumentistes de haut niveau [WD04].

4.1 Analyse fonctionnelle des gestes associés au Cantor Digitalis et Digitartic

Il s'agît ici de traiter de la typologie des gestes instrumentaux du *Cantor Digitalis* et du *Digitartic*, en se basant sur l'analyse de Cadoz [Cad88]. Selon cet auteur, un geste instrumental est un geste appliqué à un objet matériel avec lequel on interagit physiquement. Le résultat de cette interaction est un phénomène physique dont l'évolution peut être maitrisée par l'utilisateur.

Dans le cas du *Cantor Digitalis*, la main principale agit comme un geste d'excitation par la pression du stylet sur la tablette. Plus la pression est forte, plus l'énergie acoustique sera importante. Et l'émission sonore est dépendante d'une pression non nulle. Le mouvement du stylet suivant les autres axes sont des gestes de modification paramétrique : fréquence fondamentale suivant la position sur l'axe X, et couleur vocalique suivant Y (de même si le contrôle de la couleur vocalique se fait avec l'autre main). Les changements de style de voix sont des gestes de sélection si les pré-configurations sont utilisées mais deviennent des gestes de modification paramétrique (voir structurel pour le bruit d'aspiration) s'ils sont réglés manuellement via les curseurs de l'interface logicielle.

Les gestes de la main principale du *Digitartic* sont identiques à ceux du *Cantor Digitalis*. La main secondaire utilise quant à elle :

- des gestes de sélection pour les modes d'articulation (choix via les zones de la tablette suivant l'axe X) et le voisement des consonnes (choix via bouton du stylet) ;
- des gestes de modification paramétrique pour modifier le lieu d'articulation pour un mode d'articulation donné (sur chaque zone du mode d'articulation suivant l'axe X)
- des gestes de modification paramétrique pour l'évolution des formants entre les cibles phonétiques, et un geste de modification structurelle pour les occlusives, les fricatives et les nasales. En effet, passer d'une voyelle à une occlusive fait intervenir une occlusion totale modifiant alors structurellement l'instrument vocal. De même, passer d'une voyelle à une fricative fait intervenir un différent mode de fonctionnement du conduit vocal en ajoutant une nouvelle source sonore au son total, en l'occurrence un bruit de friction. Les consonnes nasales sont le résultat de l'utilisation du conduit nasal, ajoutant alors un nouvel élément à l'instrument vocal.

4.2 Outils d'analyse phénoménologique du geste

Nous étudions les gestes associés à l'instrument *Cantor Digitalis* en mode mono-manuel 2.7.3. L'axe X de la tablette graphique est donc relié à la hauteur mélodique, l'axe Y à la couleur vocalique et la pression à l'effort vocal. Pour cela, nous utilisons les données renvoyées par la tablette graphique. Ainsi, ce qui est mesuré est la position de la pointe du stylet dans le plan de la tablette et sa pression sur la tablette, et non le geste corporel dans son ensemble.

On récupère 2 vecteurs. Le premier est échantillonné à 20 ms et constitué des valeurs des paramètres suivants :

- l'horloge de l'ordinateur (ms) ;
- la position du stylet suivant l'axe X ;
- la position du stylet suivant l'axe Y ;
- la pression du stylet sur la tablette.

Bien que la période d'échantillonnage de la tablette soit de 5 ms, nous avons réduit celle de l'enregistrement du premier vecteur à 20 ms pour des questions de temps de calcul, l'enregistrement étant fait sur le même ordinateur que celui qui contient le moteur de synthèse. Nous verrons que cette fréquence d'échantillonnage est suffisante pour notre analyse.

Le deuxième est échantillonné à 5 ms, synchronisé avec le premier, et constitué des valeurs des paramètres suivants :

- l'horloge de l'ordinateur (ms) ;
- le signal audio du micro externe qu'on sous-échantillonne à 5 ms, et qui a pour seul but de synchroniser plusieurs fichiers provenant d'ordinateurs différents (et donc de *Cantor Digitalis* différents).

Dans le cas où nous analysons les gestes de plusieurs musiciens, nous devons synchroniser temporellement les données issues de chacun des ordinateurs. Pour cela, nous enregistrons avec tous les ordinateurs un son bref (claquement de main) qui nous permet de synchroniser les horloges internes des ordinateurs en relevant la valeur de l'horloge sur chacun des ordinateurs à l'instant de l'impulsion sonore. Cet enregistrement se fait à une période d'échantillonnage de 5 ms comme indiqué ci-dessus, afin d'être clairement en dessous du seuil de perception de décalage de deux sons consécutifs, et correspondant environ au temps de parcours du son entre l'émetteur et les ordinateurs les plus distants, soit à l'incertitude sur la synchronisation. L'enregistrement de l'impulsion de synchronisation est réalisé avant de commencer à jouer et ne rentre donc pas en concurrence avec le calcul du moteur de synthèse.

L'enregistrement se fait via un patch Max/MSP (copie d'écran à la figure 4.1) intégré à l'application de réception des données.

4.3 Les gestes pour imiter certaines tâches musicales

Chacun des nouveaux arrivants de la chorale *Chorus Digitalis* s'est vu expliquer le fonctionnement de base de l'instrument (en l'occurrence le *Cantor Digitalis)*. Aucune technique de jeu n'a été imposée aux participants. Ainsi, chacun a développé sa propre technique, plus ou moins indépendamment des autres joueurs. Nous traitons dans cette partie de l'observation de certaines techniques, qui ont émergé de la pratique de chacun, pour accomplir des tâches musicales nécessaires à l'interprétation des morceaux du répertoire du *Chorus Digitalis*.

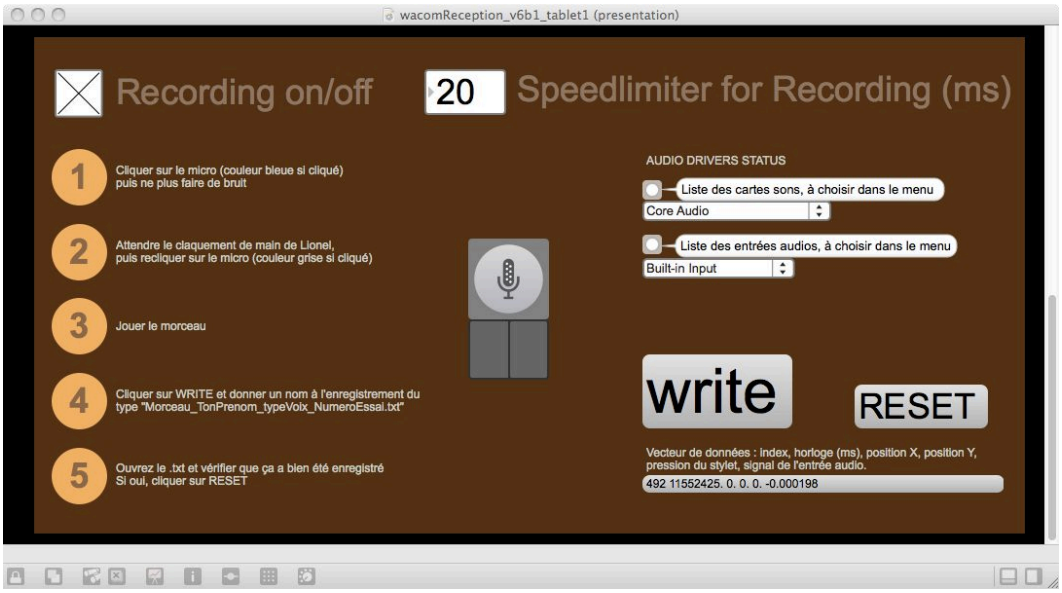


Figure 4.1 – *Copie d'écran de l'application permettant d'enregistrer les données de la tablette*

4.3.1 Portamento

Le portamento est le terme désignant le passage continu d'une note à une autre par glissando. Il s'applique donc seulement aux instruments permettant une telle continuité entre deux notes, comme dans le cas de la voix.

En ce qui concerne nos instruments de voix synthétique, que ce soit avec le *Cantor Digitalis* ou le *Digitartic*, le changement continu de F_0 est réalisé par un mouvement d'un point du plan XY à un autre, la hauteur de note étant contrôlée par l'axe X de la tablette. Différentes stratégies ont été entreprises par les membres du Chorus Digitalis. La plupart optent pour un mouvement rectiligne de la note de départ à celle d'arrivée, qui a l'avantage d'être le plus rapide. Mais d'autres préfèrent un mouvement en courbe permettant d'atteindre la note cible plus facilement. En effet, une courbe induit à la fin du mouvement un ralentissement de la composante de la vitesse suivant l'axe X (voir la trajectoire du mouvement dans le plan 2D de la tablette sur la figure 4.2).

L'enchainement de plusieurs notes peut se faire par la succession de courbes comme indiqué sur la figure 4.3 où est jouée la séquence de notes Do-Mi-Sol-Ré avec portamento entre chaque note. Afin de pointer la note, le sens de rotation du geste est toujours le même, ce qui implique une courbe "par le bas" en jouant un portamento sur une séquence de notes de hauteur croissante (Do-Mi et Mi-Sol), et une courbe "vers le haut" sur une séquence de note de hauteur décroissante (Sol-Ré). Dans cet exemple, le sens utilisé est trigonométrique. Si la vitesse d'exécution devient grande, le rayon de courbure du geste aura tendance à s'agrandir et à tendre vers un mouvement rectiligne afin de gagner en vitesse. Il en sera de même si l'intervalle à réaliser est grand, se traduisant sur la tablette par une plus grande distance à parcourir sur une même durée et donc un mouvement devant être nécessairement plus rapide.

4.3.2 Vibrato et gamak

Le vibrato est une modulation périodique de la fréquence fondamentale autour d'une valeur moyenne, qui peut s'accompagner de la modulation d'autres paramètres comme l'am-

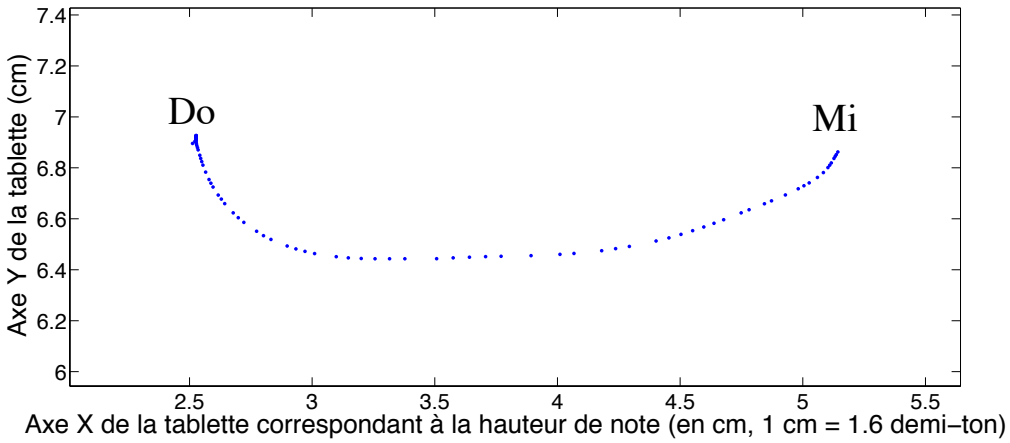


FIGURE 4.2 – *Trajectoire du stylet sur la tablette pour la succession des notes Do et Mi*

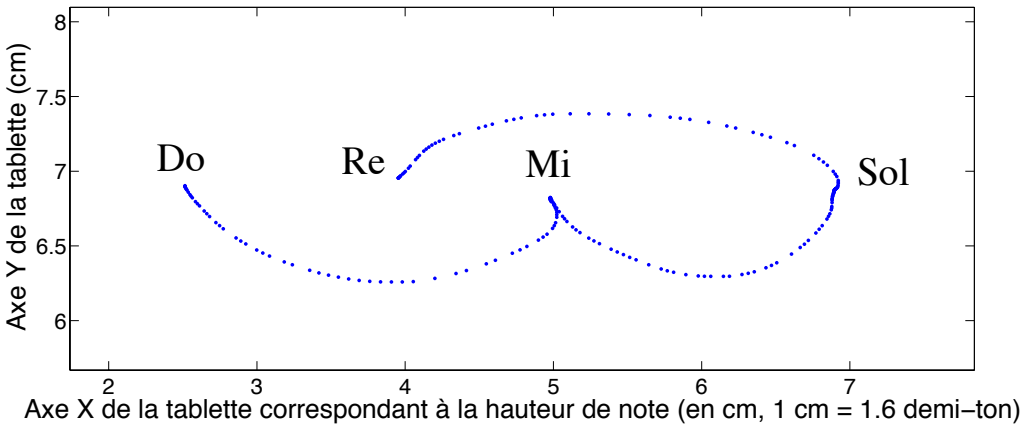


FIGURE 4.3 – *Trajectoire du stylet sur la tablette pour la succession des notes Do-Mi-Sol-Re, sans vibrato sur les notes cibles*

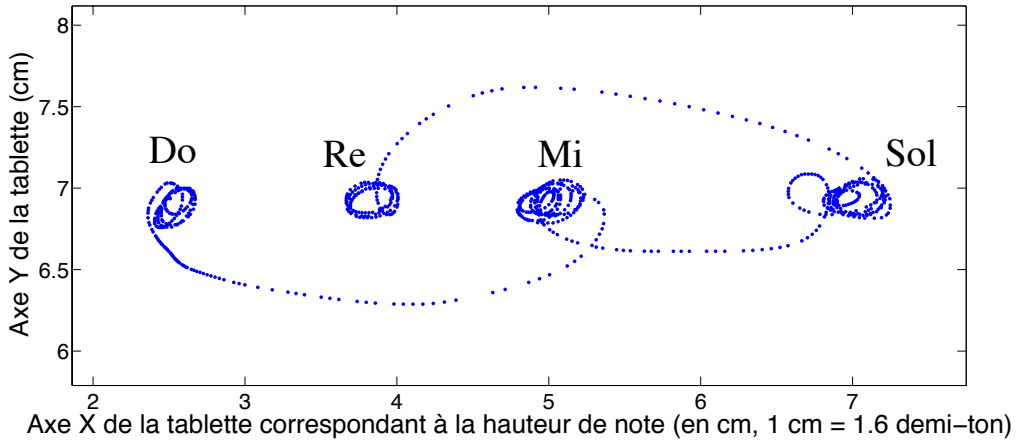


Figure 4.4 – *Trajectoire du stylet sur la tablette pour la succession des notes Do-Mi-Sol-Re, avec vibrato sur les notes cibles*

plitude du signal et le timbre. Chez des chanteurs interprétant *Ave Maria* de Schubert, Prame mesure une modulation de fréquence moyenne d'environ 6 Hz, avec une variation interindividuelle et intra-individuelle de ± 10% [Pra94]. La fin des notes présente une augmentation exponentielle de la fréquence de la modulation, d'après Bretos et Sundberg [BS03].

Dans notre modèle de contrôle, le mouvement le plus simple *a priori* pour moduler la fréquence est un mouvement périodique du stylet le long de l'axe X. Certains musiciens du *Chorus Digitalis* effectuent le geste du vibrato rectilignement et parallèlement à l'axe X, mais d'autres considèrent comme plus simple un geste formant un cercle dans le plan XY autour de la fréquence centrale. Le geste est ainsi de vitesse constante, donc a priori moins fatiguant. Il est à noter que le mouvement circulaire est réalisé par les mêmes personnes que celles utilisant un geste courbe pour le portamento. De plus, le sens de rotation reste identique à celui du portamento. Le mouvement à l'origine du déplacement suivant l'axe X se fait à l'aide du poignet (comme si on raturait un mot) tandis que celui suivant Y se fait à l'aide des doigts (comme si on écrivait la barre verticale de la lettre *I*. Le geste de vibrato en cercle est la résultante de la combinaison de ces deux mouvements, correspondant au mouvement réalisé pour écrire un *O*. Elle est d'autant plus facile à réaliser qu'on a l'habitude des gestes de boucle fréquents dans l'écriture.

La figure 4.4 présente la trajectoire du même enchaînement de notes qu'à la figure 4.3 mais avec un vibrato sur chacune des notes cibles, par un joueur utilisant la technique de rotation décrite ci-dessus. On observe bien le sens de rotation constant entre les portamentos et vibratos.

Le geste qui agit sur la fréquence fondamentale (axe X de la tablette) induit, en plus du mouvement dans le plan XY cité ci-dessus, une variation de la pression du stylet exercée sur la tablette. Il faudrait étudier si cette variation de pression est corrélée avec la modulation en fréquence de la même façon que dans la voix naturelle.

Tout comme le portamento, il y a un équilibre à trouver entre la circularité du geste et la vitesse d'exécution et l'amplitude du geste à effectuer. En chant Khayal d'Inde du nord, une ornementation appelée *gamak* est largement utilisée. Elle consiste en un vibrato d'une grande amplitude en fréquence. Pour reproduire cette tâche musicale dans un morceau de ce

style avec notre instrument, le mouvement circulaire devient trop lent et un geste rectiligne s'impose.

La fréquence du vibrato doit être correctement reproduite à l'aide de ces mouvements afin d'obtenir un résultat naturel, de part son étendue et sa fréquence moyennes, et son évolution temporelle comme l'augmentation exponentielle à la fin des notes décrite par Bretos et Sundberg [BS03].

4.3.3 Attaque de note

Une attaque de note est le début d'un son. Dans le cas de la voix et pour une attaque d'une voyelle, elle correspond à la phase de mise en vibration des plis vocaux.

Dans nos instruments, le modèle de source RT-CALM fonctionne en permanence, même pendant un silence. C'est le paramètre d'effort vocal qui en passant de 0 à une valeur non nulle (ou à partir d'un certain seuil dans le cas de l'activation du seuil de phonation) va diminuer la pente spectrale de la source et augmenter l'amplitude du signal du débit glottique. La limite de ce modèle est qu'il ne prend pas en compte la phase transitoire de mise en vibration des cordes vocales que nous n'étudions pas ici. De plus, suivant la phase du signal glottique à laquelle intervient l'attaque, on aura un effet sonore légèrement différent, qui se traduira par un petit bruit si l'attaque est enclenchée sur une partie non nulle du signal glottique. Cet effet sonore n'est pas perceptible si le seuil de phonation n'est pas activé, l'amplitude du signal commençant à 0 quelle que soit la phase du signal de source glottique.

La plus simple façon d'imiter une attaque de note avec nos instruments est de simplement poser le stylet sur la tablette, sa pression étant associée à l'effort vocal. Cependant, peut être à cause des limites de notre modèle donné ci-dessous, l'attaque manque de naturel. Pour combler cette lacune, un des joueurs a empiriquement proposé de faire évoluer rapidement la voyelle au moment de l'attaque, ce qui s'est traduit par une amélioration du caractère naturel de l'attaque. Par exemple, comme indiqué sur la figure 4.5 avec un mapping monomanuel du *Cantor Digitalis*, le joueur part de la voyelle /e/ à la fréquence fondamentale F_0 et très rapidement se déplace vers la voyelle visée /a/. On peut expliquer la réussite de cet effet par le fait qu'en voix naturelle, les articulateurs se placent en position de la voyelle visée pendant l'attaque, induisant un rapide changement des résonances du conduit vocal au début de la production d'une voyelle (effet de co-articulation). En passant de /e/ à /a/, cela correspond à une ouverture de la mâchoire pendant l'attaque. Et comme souvent en synthèse vocale, la non stabilité des paramètres de production induit une amélioration de la qualité de la synthèse, donnant de la "vie" au son en offrant à l'auditeur la variation d'une dimension supplémentaire du modèle de production.

4.4 Gestes d'accompagnement et style de jeux

Dans cette partie, nous décrivons les différents styles de jeux de quatre musiciens jouant la chorale baroque *Wie Schön leuchtet der Morgenstern* de Johan-Sebastian Bach (partition à la figure 4.6), à travers les gestes d'accompagnement, c'est-à-dire ceux non nécessaires à la production du son comme définis dans [WD04]. L'instrument utilisé est le *Cantor Digitalis* en mode mono-manuel, c'est à dire où seule la main principale est utilisée, avec les correspondances suivantes : effort vocal suivant la pression du stylet sur la tablette ; hauteur mélodique suivant l'axe X de la tablette ; couleur vocalique i.e. voyelles /i,e,a,o,u/ et leurs intermédiaires sur l'axe Y à une dimension.

Ces trois données de la tablette ont été enregistrées pour chacun des quatre musiciens et

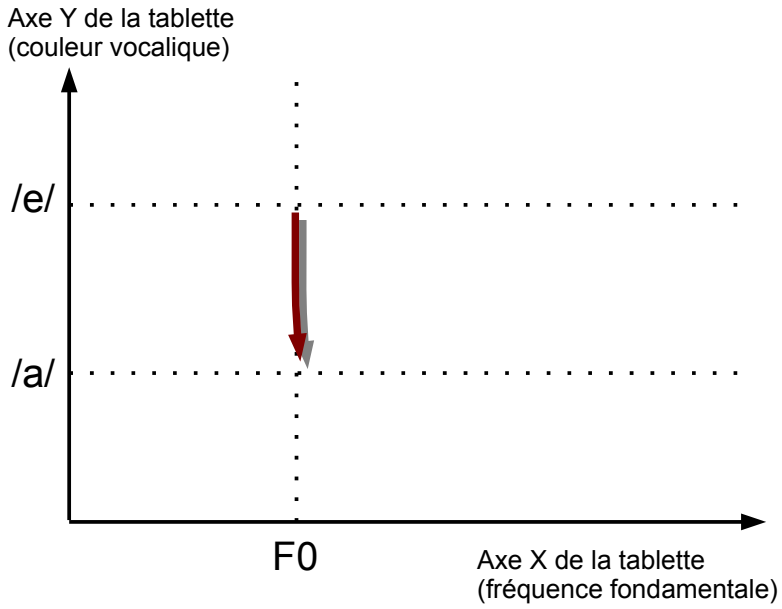


Figure 4.5 – *Trajectoire du stylet dans le plan de la tablette pendant une attaque avec pour cible la voyelle /a/* et la fréquence fondamentale F0

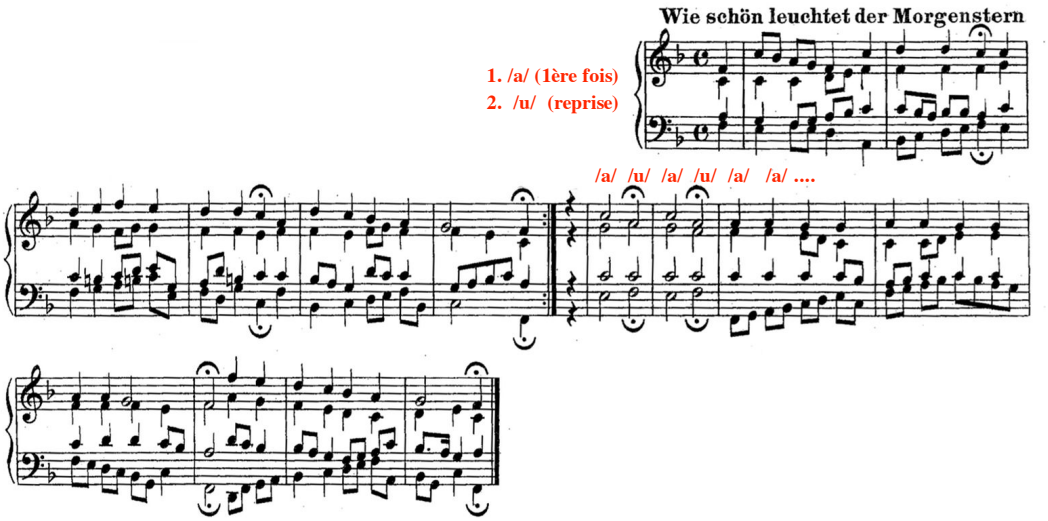


Figure 4.6 – *Partition du morceau Wie Schön leuchtet der Morgenstern (Johan-Sebastian Bach)*

synchronisées comme expliqué dans la section 4.2. On s'intéresse ici seulement à l'évolution temporelle des données de position X/Y du stylet sur la tablette.

La figure 4.7 est la réalisation gestuelle de la partition de la figure 4.6, représentant l'évolution de la hauteur mélodique (via la trajectoire du stylet suivant l'axe X) en fonction du temps. Toutes les informations de la partition y apparaissent : durée des notes, hauteur des notes et voyelle utilisée. On voit clairement la réalisation de la reprise sur chacune des voix reproduisant deux motifs à l'identique (du début à la 52$^{\text{ème}}$ seconde et de la 53$^{\text{ème}}$ à la 83$^{\text{ème}}$ seconde sur la figure 4.7). Sur cette figure comme sur la partition, on peut voir que les tâches musicales demandées à chacun des musiciens sont assez similaires : rythmes simples où les durées des notes changent peu, hauteurs de notes proches entre notes successives, évolutions temporelles qui se superposent assez bien, et même évolution de voyelle. La principale différence est l'étendue des hauteurs de notes, élevée pour la voix de basse (musicien BD), moyenne pour la voix de soprano (musicien LF) et relativement faible pour la voix d'alto et de ténor (respectivement musicien CA et HM). Sur cette figure, la position du stylet suivant l'axe Y est donnée par la couleur de la courbe, bleue pour un /a/ et verte pour un /u/. Cette représentation permet de rendre compte de la voyelle choisie, mais pas des petites variations suivant Y, suffisamment faible pour ne pas induire un changement nette de perception de voyelle. Ce sont ces petites variations qui nous intéressent, car comme introduites dans la section 4.3, elles interviennent dans la technique de certaines tâches musicales. Elles ne sont pas nécessaires à leur réalisation, mais aident à la virtuosité du geste.

Les figures 4.8, 4.9, 4.10 et 4.11 représentent la trajectoire du stylet dans le plan XY de la tablette tout au long du morceau pour chacun des quatre musiciens BD, CA, HM et LF. Pour une meilleure visualisation de chacun, les limites des axes sont fixées par l'étendue maximale des gestes, et la représentation du jeu du musicien BD voit ses axes Y et temporel inversés par rapport aux autres musiciens. Cette représentation permet de visualiser dans son ensemble les différences des gestes d'accompagnement des quatre musiciens, comme on pourrait le faire avec différents styles d'écriture, à la différence qu'un tracé 3D n'est pas nécessaire pour l'écriture pour laquelle on ne réécrit pas par dessus des lettres précédemment écrites. Nous n'avons pas relié les points de mesure afin d'avoir une information sur la vitesse du mouvement.

Pour décrire un peu plus précisément ces gestes, nous projetons les représentations 3D précédentes respectivement dans le plan XY (figure 4.12) et dans le plan Y-temps (figure 4.13). De ces deux figures, ainsi que des représentations 3D, on peut faire les remarques qualitatives suivantes :

- le degré de courbure du geste pour la réalisation des portamentos varie grandement entre chaque musicien, allant de très faible chez la voix de ténor et de soprano (musicien HM et LF) à très important chez la voix de basse (musicien BD).
- on retrouve la technique de portamento en courbe suivant Y chez la basse (musicien BD), comme décrite dans la section 4.3. Malgré la superposition des phrases, on identifie bien les notes visées et les portamentos, mais ceux-là sont presque toujours réalisés par des courbes "par le bas", c'est à dire en changeant le sens de rotation du geste pour les successions de notes ascendantes et descendantes (figure 4.12).
- L'écart à la moyenne des mouvements suivant Y joués par les musiciens HM et LF est plus grand que les deux autres. (figure 4.13).
- le musicien CA, jouant la voix d'alto, a tendance à réaliser des portamentos avec une pente négative (figure 4.9 et 4.12). On peut faire l'analogie avec les écritures de style "penché".

D'autres gestes d'accompagnement existent sûrement pour le jeu du *Cantor Digitalis*,

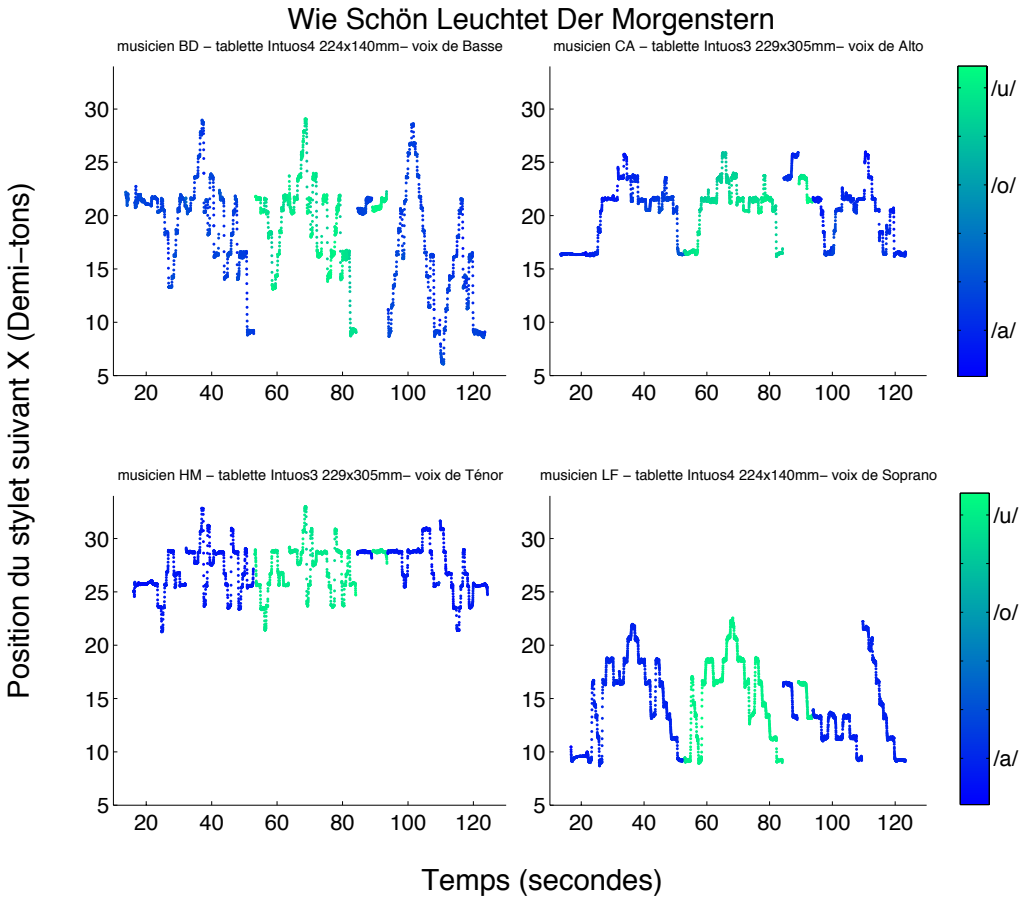


Figure 4.7 – *Enregistrements gestuels des quatre voix interprétant Wie Schön Leuchtet Der Morgenstern, dans l'espace F_0-Temps et avec la couleur vocalique en couleur*

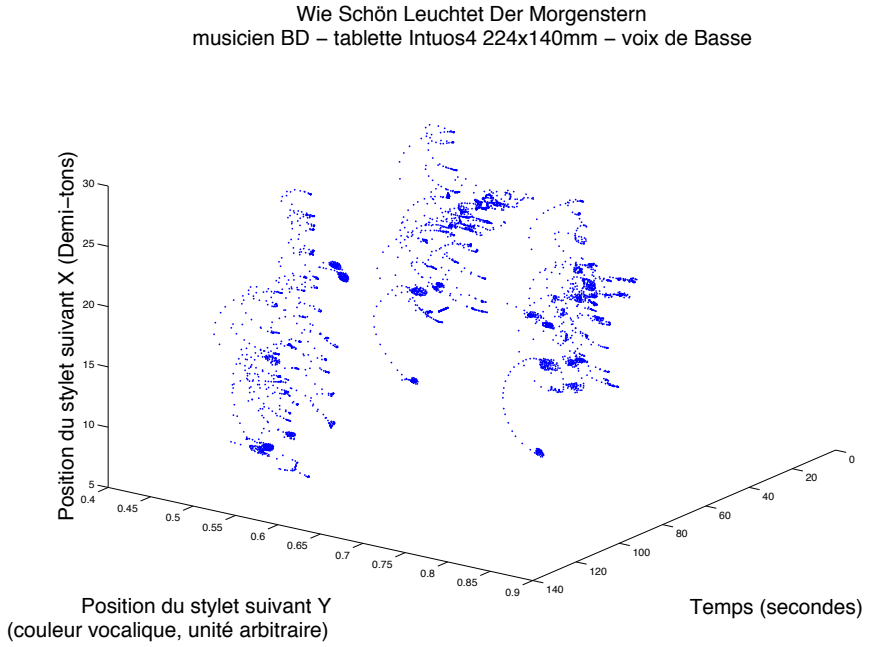


FIGURE 4.8 – *Enregistrements gestuels du musicien BD interprétant Wie Schön Leuchtet Der Morgenstern, dans l'espace F_0-Temps-Voyelle*

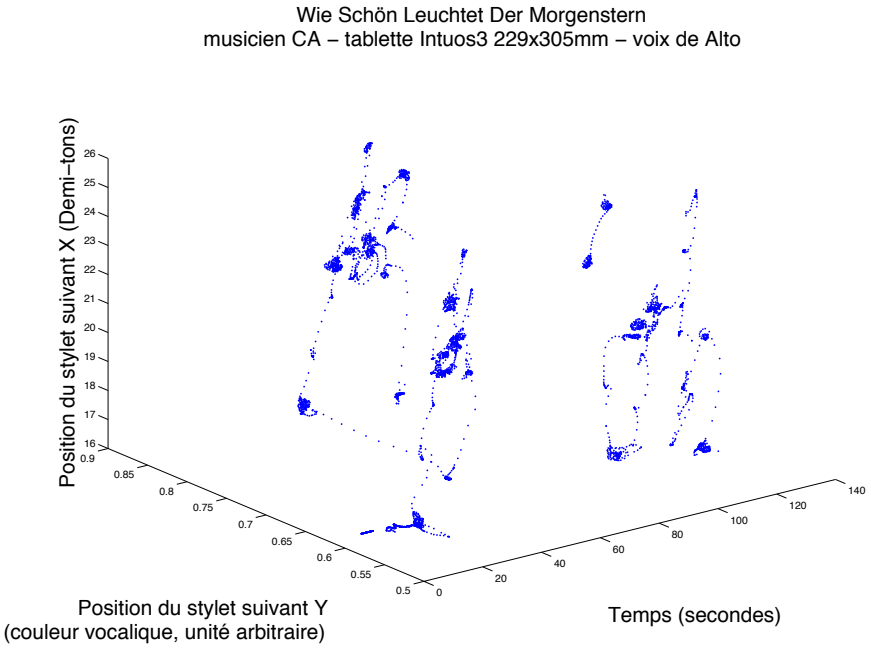


FIGURE 4.9 – *Enregistrements gestuels du musicien CA interprétant Wie Schön Leuchtet Der Morgenstern, dans l'espace F_0-Temps-Voyelle*

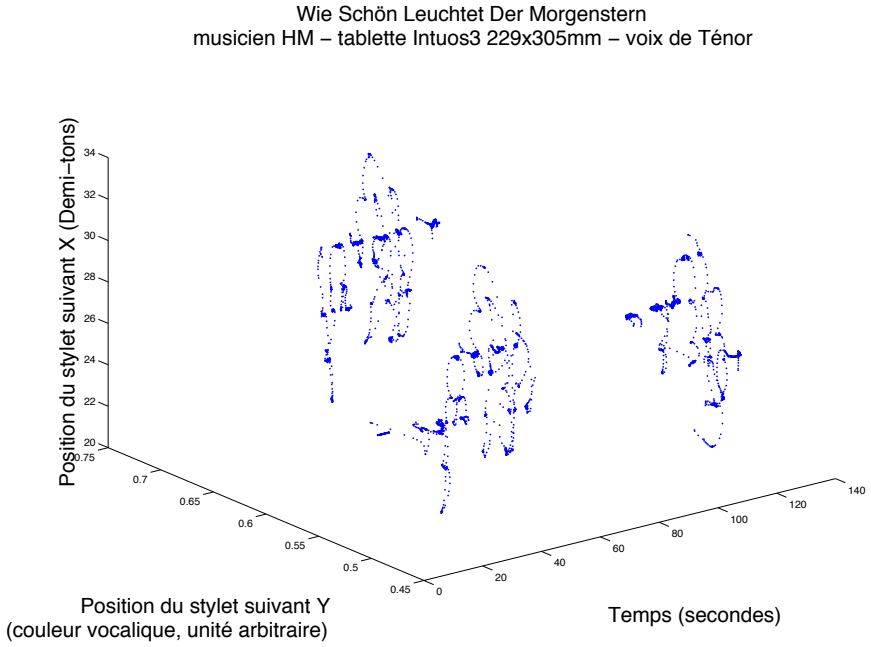


Figure 4.10 – *Enregistrements gestuels du musicien HM interprétant Wie Schön Leuchtet Der Morgenstern, dans l'espace F_0-Temps-Voyelle*

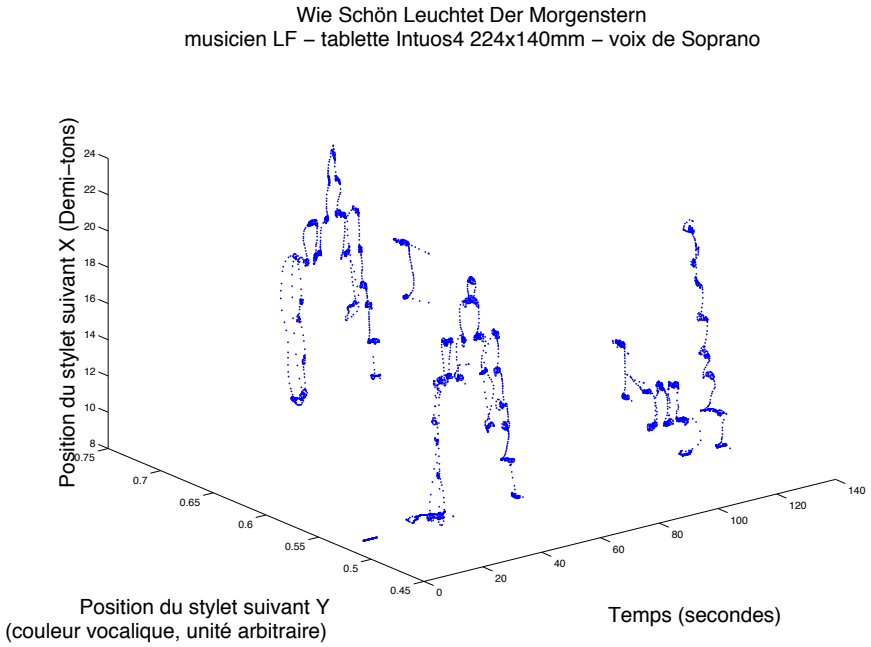


Figure 4.11 – *Enregistrements gestuels du musicien LF interprétant Wie Schön Leuchtet Der Morgenstern, dans l'espace F_0-Temps-Voyelle*

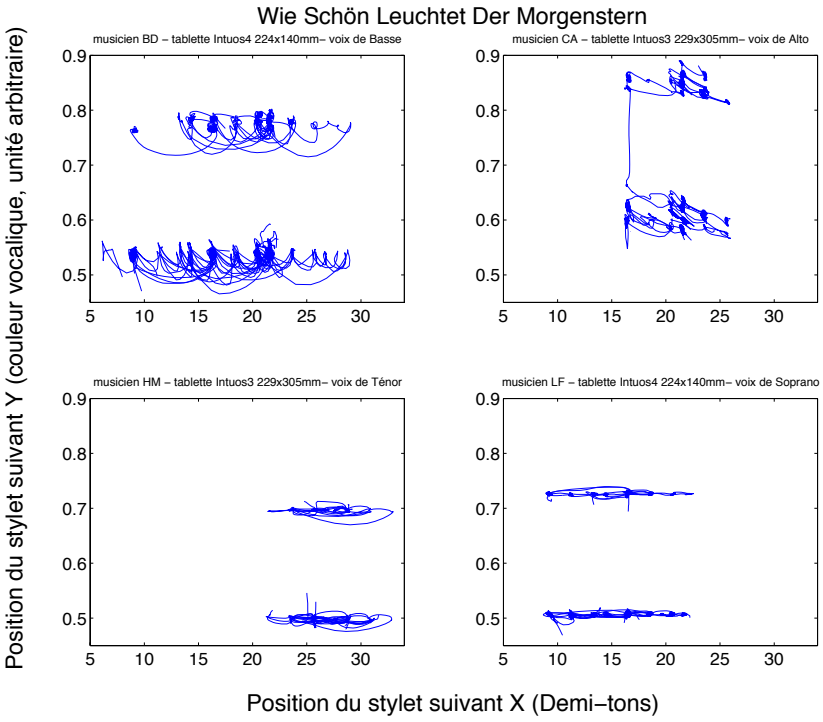


FIGURE 4.12 – *Enregistrements gestuels des quatre voix interprétant Wie Schön Leuchtet Der Morgenstern, dans l'espace 2-D de la tablette (F_0-Voyelle)*

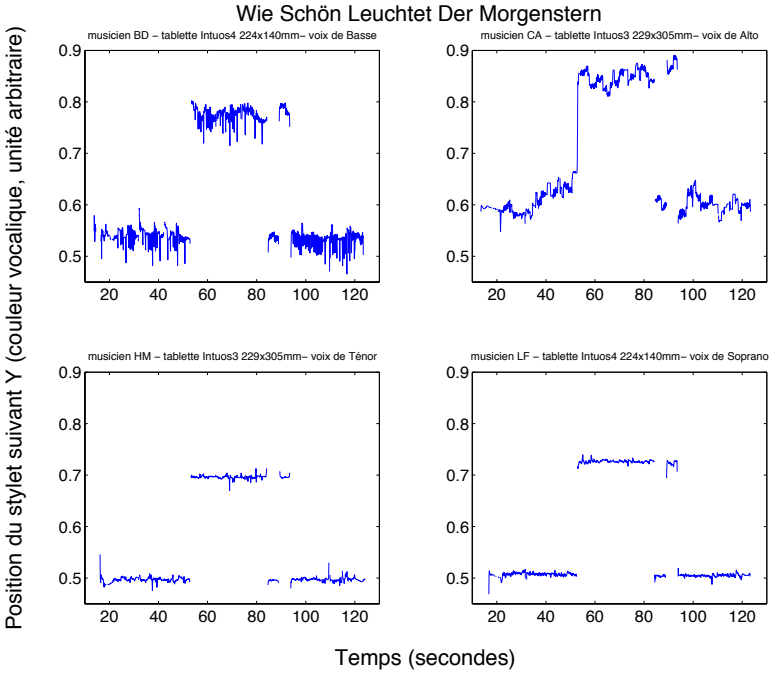


FIGURE 4.13 – *Enregistrements gestuels des quatre voix interprétant Wie Schön Leuchtet Der Morgenstern, dans l'espace Temps-Voyelle*

comme les mouvements du poignet et du bras. Nous nous sommes intéressés aux petits mouvements suivant Y car ces gestes sont facilement récupérables (puisque captés par l'interface), et car il semble évident qu'ils aident à la réalisation des tâches musicales, comme explicité dans la section 4.3.

4.5 Conclusion

Les gestes associés à nos deux instruments ont été analysés à travers deux axes, à savoir une analyse fonctionnelle plutôt théorique et une analyse phénoménologique basée sur les usages des musiciens de la chorale *Chorus Digitalis*. Les gestes pour réaliser certaines tâches musicales comme le portamento, le vibrato ou encore les attaques ont été décrits. Des tendances à développer des styles de jeu personnels ont été cernées parmi les musiciens.

D'un point de vue artistique, la richesse des trajectoires des mouvements du stylet pourrait être exploitée dans l'ensemble *Chorus Digitalis*, afin d'améliorer les aspects visuels des concerts.

Chapitre 5

Justesse et précision de l'intonation musicale chironomique et vocale

Sommaire

5.1 Introduction

Sur plus d'un millier d'étudiants d'un cours de psychologie aux états-unis, 59% d'entre eux déclarent ne pas chanter « juste » [PB07], alors que c'est en pratique peu répandu, si on se réfère au seuil de non-justesse à 1 demi-ton de la cible [PBM+10]. Deux types de mesure peuvent être faites pour mesurer la faculté à reproduire une mélodie : la mesure de la distance à la note cible, la *justesse*, et la faculté à reproduire avec constance un même motif mélodique (écart-type de cette distance), la *précision* [TS88]. La réponse des étudiants au sondage sur leur évaluation de leur capacité à chanter « juste » est en fait corrélée à des mesures sur la précision, qui donne 50-60% de chanteurs imprécis, en considérant le seuil d'imprécision à 1 demi-ton. Cependant, ces jugements sont à considérer en fonction du seuil fixé. En effet, un demi-ton correspond dans la plupart des musiques à un changement discret de hauteur de note, donc à une erreur de cible plutôt qu'à un écart trop grand à la cible.

La « chironomie », du grec *cheir* (main) et *nomos* (règle), désigne le contrôle manuel de l'intonation vocale. Les résultats de d'Alessandro *et al.* [dRLB11] montrent qu'il est possible d'imiter l'intonation d'une phrase pré-enregistrée avec une plus grande justesse à l'aide de gestes manuels sur une tablette graphique qu'avec sa propre voix.

Alors que l'intonation de la voix parlée nécessite une justesse d'environ un demi-ton (100 cents), celle de la voix chantée peut nécessiter environ 5-10 cents en l'absence de vibrato. En comparaison à la voix parlée, le chant nécessite d'inscrire la hauteur mélodique dans une série de notes à références discrètes, en plus d'une synchronisation temporelle accrue allant jusqu'à 15-20 millisecondes. Ainsi, l'objet de notre présente étude, la chironomie pour l'intonation musicale, constitue une tâche plus difficile que la voie parlée. Nous proposons de l'explorer en se focalisant sur les deux dimensions exposées ci-dessus, à savoir la justesse et précision du *chant chironomique* (chanter par le geste de la main), et sa comparaison avec la voix naturelle. Cette étude fera intervenir aussi le rôle du retour audio et de l'habileté manuelle et vocale.

La tablette graphique munie d'un stylet est une interface bien adaptées à la chironomie, car elle fait intervenir les gestes de l'écriture : ceux de la visée de cibles avec un crayon. De plus, sa haute résolution temporelle et spatiale en font un candidat adapté au contrôle temps réel et à notre capacité à percevoir de très petites variations de hauteur. Fréquemment utilisée comme interface pour les musiques électroniques actuelles [ZWMC07], nous voulons étudier sa capacité à contrôler finement la hauteur musicale, en l'augmentant d'un calque de repères mélodiques.

Le but de ces expériences est de mesurer la justesse et la précision de l'intonation musicale suivant 3 modalités : à l'aide de sa propre voix (*Voix*), à l'aide du Cantor Digitalis avec sa tablette graphique (*Tablette + Audio*), et à l'aide de la tablette graphique seule sans retour audio du synthétiseur Cantor Digitalis (*Tablette Seule*). Des sujets ont imité des intervalles musicaux dans des expériences afin d'étudier l'effet des paramètres suivants : la modalité d'imitation, le tempo, la longueur des séquences de notes à reproduire, les sujets, le stimulus à reproduire et la distance à la note cible précédente.

Nous commencerons par décrire les conditions expérimentales, dont le type de participants, les outils utilisés et le protocole expérimental. Ensuite, nous traiterons de l'analyse et de la transformation des données brutes issues de l'expérience. Enfin, seront présentés les différents résultats, avant de conclure et discuter ce travail.

5.2 Expériences en chant chironomique

5.2.1 Participants

Les expériences ont été réalisées sur 31 sujets membres du laboratoire, mais seulement 20 d'entre eux ont pu être exploités pour les expériences 1 et 2 et 28 pour l'expérience 3 (erreurs d'octave systématiques, tâches réalisées incomplètes, incapacité à chanter suffisamment juste pour l'algorithme de détection de F_0, bogue informatique menant à la perte de données).

La moyenne d'âge du groupe de 20 personnes des expériences 1 et 2 est de 31 ans, s'étalant de 22 à 49 ans et comprenant 6 femmes, 14 hommes, 4 gauchères, et 16 droitières. Quatorze d'entre elles rapportent d'une pratique musicale présente ou passée (en moyenne de 18 ans). Trois d'entre eux ont une expérience dans l'utilisation de la tablette dans le cadre d'une pratique musicale impliquant le *Cantor Digitalis*. Enfin, 12 sujets sur 20 signalent une inaptitude à chanter juste.

La moyenne d'âge du groupe de 28 personnes de l'expérience 3 (comprenant l'intégralité du groupe des expériences 1 et 2) est de 29 ans, s'étalant de 21 à 49 ans et comprenant 11 femmes, 17 hommes, 5 gauchères et 23 droitières. Les audiogrammes réalisés auprès de chacun des participants ne présentaient pas d'anomalie. Dix-huit d'entre eux rapportent d'une pratique musicale présente ou passée (en moyenne de 16 ans). Trois d'entre eux ont une expérience dans l'utilisation de la tablette dans le cadre d'une pratique musicale impliquant le *Cantor Digitalis*. Enfin, 15 sujets sur 28 signalent une inaptitude à chanter juste.

5.2.2 Matériel

a) L'instrument de synthèse utilisé dans ces expériences

Parmi les 3 modalités d'imitation, la modalité *Tablette + Audio* est mise en oeuvre à l'aide de l'instrument *Cantor Digitalis* décrit dans le chapitre 2. Parmi les paramètres contrôlables de l'instrument, seulement deux sont utilisés dans cette étude, l'effort vocal et la hauteur mélodique. Les correspondances avec la tablette sont les mêmes que dans le *Cantor Digitalis* : respectivement la pression du stylet sur la tablette et sa position suivant l'axe X (axe droite-gauche du point de vue de l'utilisateur). Notamment, la couleur vocalique était fixée sur un /a/ et la qualité de voix était relâchée sans bruit de souffle. La tessiture de la voix était d'une octave plus basse pour les hommes que pour les femmes (le Do le plus bas passe de 250 à 125 Hz, soit de 60 à 48 en notation MIDI), afin de faire correspondre au mieux les tessitures moyennes des deux genres à la voix synthétique. La tablette utilisée est l'Intuos 3L, de surface active 228.6 × 304.8 mm (9 × 12 pouces), présentant 1024 niveaux de pression au stylet, une résolution temporelle de 200 Hz, et une résolution spatiale de 0.25 mm au le stylet.

L'interface du *Cantor Digitalis* utilisée dans cette étude est légèrement différente de l'originale par le choix du calque de marqueurs superposé à la tablette, et dans l'étendue fréquentielle de la voix. Le calque est constitué ici de lignes verticales séparées les unes des autres de 1.4 cm correspondant à un demi-ton, contre 0.65 cm dans la version présentée au chapitre 2. La résolution de la tablette correspond alors à 3.8 cents, soit juste en dessous du seuil de perception. Les lignes des notes utilisées dans l'expérience sont accompagnées de leur nom (Do, Ré, Mi, Fa, Sol, La, Si, Do). Seule une octave est utilisée. Les noms des notes sont répétés suivant l'axe Y pour pouvoir jouer sur différentes parties de la tablette selon sa convenance. La tablette munie de ce calque est visible sur la figure 5.1.

Figure 5.1 – *Tablette graphique munie de son calque utilisée pour cette étude*

b) Enregistrements de la voix et des gestes de la tablette

Deux sessions expérimentales se sont déroulées en parallèle dans deux pièces très peu réverbérantes et isolées acoustiquement de l'extérieur. Pour la première modalité d'imitation *Voix*, un micro de type DPA 4006-TL capte la voix chantée et le son est enregistré en format WAVE avec un MacBook Pro muni d'une carte son Fireface 400 ou 800 suivant la salle. Pour les deux autres modalités d'imitation (*Tablette Seule* et *Tablette + Audio*), la pression et la position du stylet suivant l'axe X de la tablette sont récupérées dans un fichier texte à l'aide de l'objet Max *wacom.mxo* [1]. Toutes les données sont collectées dans le logiciel Max 5 [Max].

c) Interface logiciel

Le scénario de l'expérience a été implémenté par Sylvain Le Beux sous la forme de 4 applications, correspondant aux 3 expériences décrites ci-dessous et à la phase d'entraînement. Pour chacune des applications, le sujet doit d'abord donner son nom et son genre (pour la tessiture). Il tape ensuite sur la touche *Espace* du clavier pour lancer l'expérience. Une fenêtre s'affiche comme illustrée à la figure 5.2, avec la partition à jouer ainsi que le son MIDI (Instrument *choir Aahs 2* du logiciel MIDI SimpleSynth [2]), un métronome sonore et visuel sur lequel le sujet doit se synchroniser, la modalité d'imitation (*Voix*, *Tablette Seule*, ou *Tablette + Audio*), et le numéro de l'essai en cours. La partition écrite présente aussi les noms des notes pour les sujets ne sachant pas lire cette notation, décalés verticalement pour indiquer le sens de l'intervalle. A la fin de chaque stimulus, une nouvelle fenêtre s'affiche et demande d'appuyer sur la touche *Espace* pour passer au stimulus suivant quand le sujet le décide.

Dans l'expérience 1 et 3, le stimulus prend fin automatiquement au bout de 3 essais. Dans l'expérience 2, le stimulus prend fin dès que le sujet appuie sur la touche *Entrée* pour

1. http://www.jmc.blueyeti.fr/download.html, lien consulté le 30 juin 2013
2. http://simplesynth.sourceforge.net/, lien consulté le 30 juin 2013

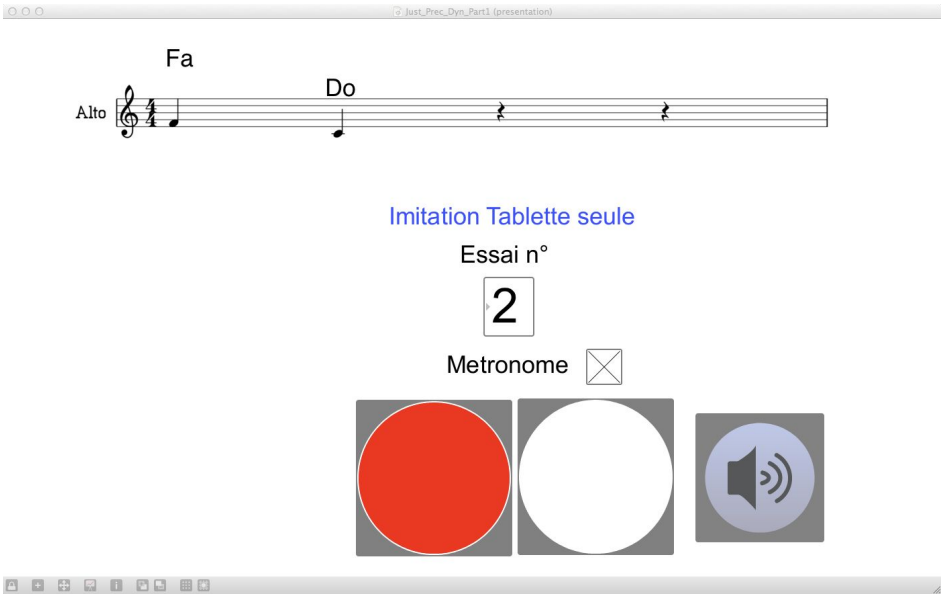


Figure 5.2 – *Capture d'écran de l'interface logiciel*

un nombre d'essais supérieur ou égal à 3, et un bouton est disponible pour rejouer le stimulus audio à imiter.

5.2.3 Protocole expérimental

Trois expériences sont menées à la suite l'une de l'autre, précédées par une séance d'entraînement. L'expérimentateur intervient pour guider le sujet entre chaque expérience. L'ordre des 3 expériences est le même pour tous les sujets : expérience 1, 3 puis 2. Tous les stimulus sont donnés dans un ordre aléatoire.

Quelques règles imposent la manière de jouer notamment pour faciliter l'exploitation des résultats : les vibratos ne sont pas autorisés, ni pour la voix, ni sur la tablette, afin de simplifier l'analyse ; pour la modalité *Voix*, on demande que le stylet soit en contact avec la tablette pendant l'imitation vocale et qu'il soit relevé seulement entre chaque essai ; pour la modalité *Tablette Seule* et *Tablette + Audio*, le stylet doit rester en contact pour toute la séquence de notes d'un essai, c'est à dire que les notes sont liées sans silence intercalé entre elles. Le nombre d'essais diffère pour chaque expérience mais seul le meilleur des essais est sélectionné. Celui-ci est défini comme ayant la plus petite valeur de la moyenne des valeurs absolues de la justesse de note et d'intervalle calculées sur la séquence de notes de l'essai considéré.

Les trois expériences diffèrent suivant les conditions suivantes : séquence de notes à imiter, modalité d'imitation (*Voix*, *Tablette Seule*, ou *Tablette + Audio*), tempo, nombre autorisé d'écoutes du stimulus et nombre d'essais possibles. Dans chaque expérience, la justesse (moyenne des biais entre la note cible et les notes atteintes) et la précision (écart-type de la distribution par rapport à la justesse) sont calculées pour les conditions exposées ci-dessous pour chacune des 3 expériences. Ces définitions peuvent porter à confusion car leur sens commun est opposé aux définitions proposées : une faible justesse correspond à une valeur élevée de la justesse ainsi définie, alors qu'une faible précision correspond à une valeur élevée de la précision ainsi définie. On parlera donc de « valeur de justesse » ou « valeur de précision » quand on se réfère à nos mesures. Ces définitions, bien que maladroites, sont conservées afin de pouvoir comparer nos résultats avec d'autres études extérieures utilisant ces mêmes

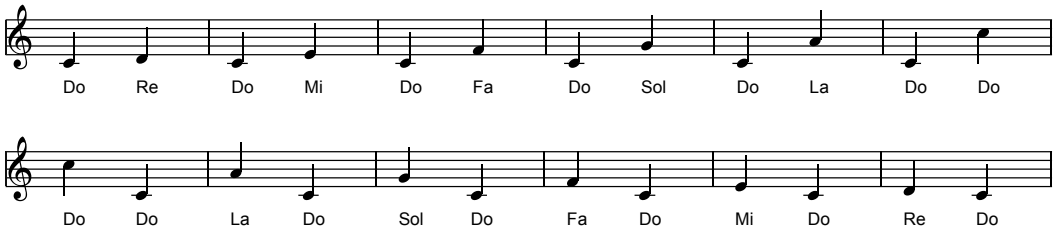


Figure 5.3 – *stimulus de l'expérience 1, composés d'intervalles ascendants et descendants, séparés par des barres de mesures*

définitions.

a) Exp. 1 : effet des 3 modalités d'imitation sur des intervalles

Cette expérience vise à comparer la voix chantée à la chironomie, avec ou sans retour audio, sur les 12 intervalles ascendants (respectivement descendants) issus de la gamme de Do majeur débutant (respectivement terminant) par le Do grave de la tablette, et donnés à la figure 5.3.

Chacun de ces intervalles doit être reproduit suivant les 3 modalités *Voix*, *Tablette Seule* et *Tablette + Audio*, soit 36 stimulus au total. Ils ne peuvent être écoutés qu'une seule fois (pour gagner du temps), et 3 essais sont demandés pour chacun, en suivant un métronome fixé à 120 battements par minute.

b) Exp. 2 : effet des 3 modalités d'imitation sur des mélodies

Cette expérience vise à comparer la voix chantée à la chironomie, avec ou sans retour audio, sur 5 mélodies simples de 6 ou 7 notes, commençant et finissant sur un Do grave, données à la figure 5.4.

Chacun de ces 5 stimulus doit être reproduit suivant les 3 modalités *Voix*, *Tablette Seule* et *Tablette + Audio*, amenant à un nombre de 15 stimulus. Comme la séquence est plus longue que dans la première expérience, les sujets sont autorisés à écouter le stimulus autant de fois qu'ils le désirent, et peuvent effectuer plus de 3 essais d'imitation. Mémoriser la séquence avant de la jouer est une tactique possible. Le métronome est fixé à 120 battements par minute.

c) Exp. 3 : effet du tempo sur des séquences de 3 notes

La troisième et dernière expérience consiste à mesurer l'influence du tempo sur la justesse et la précision dans le cas de la chironomie. Seule la modalité *Tablette + Audio* était demandée. Les sujets doivent imiter 12 séquences de 3 notes, à 3 tempos différents (120, 180 et 240 battements par minute), amenant à 36 stimulus. Les 6 premières séquences débutent et finissent sur un Do grave et la note centrale prend les valeurs de la gamme de Do majeur, de Ré à Do aigu. Les 6 séquences suivantes sont caractérisées par une note centrale fixée au Do grave, et par des notes identiques aux deux extrémités variant dans la gamme de Do Majeur, de Ré à Do aigu. Les 12 séquences sont données à la figure 5.5 séparées par une barre de mesure. Comme dans la première expérience, les sujets ne peuvent écouter le stimulus qu'une seule fois et 3 essais d'imitation sont imposés et enregistrés.

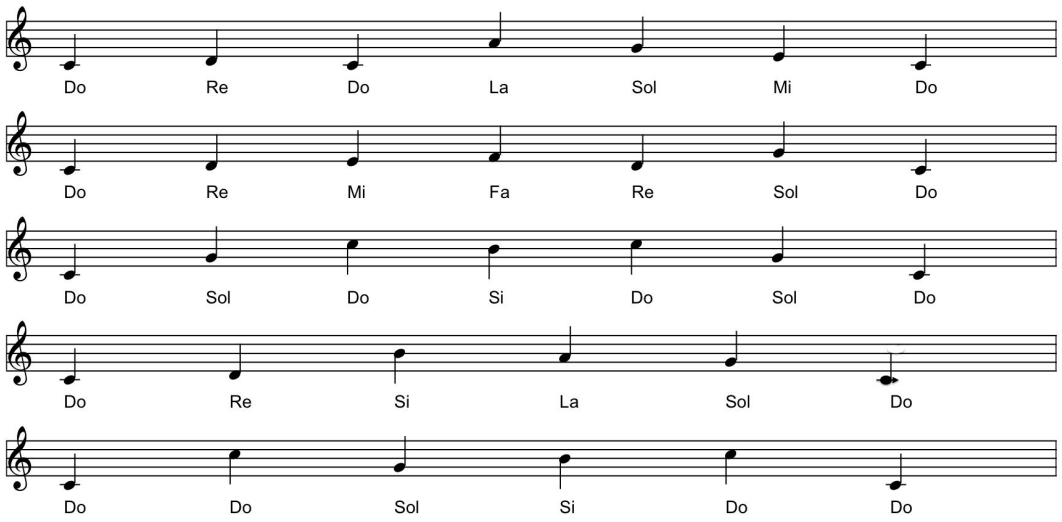


Figure 5.4 – *Stimulus de l'expérience 2, composés de mélodies de 6 ou 7 notes, séparées chacune un retour à la ligne*

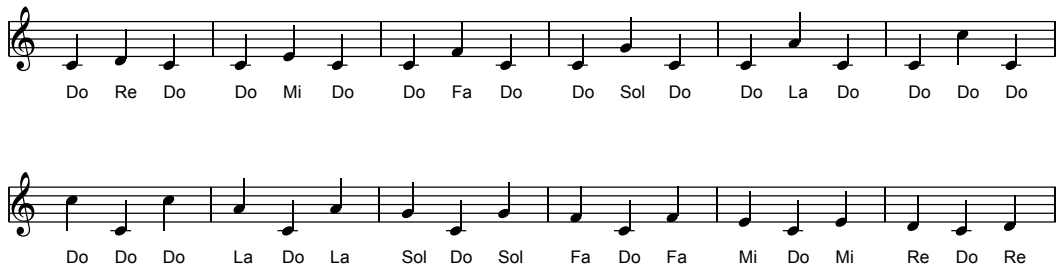


Figure 5.5 – *Stimulus de l'expérience 3, composés de séquences de 3 notes, séparées chacune par une barre de mesure*

d) Session d'entraînement

Une séance d'entraînement précède ces 3 expériences consistant en 18 stimulus de 2 ou 3 notes à imiter chacun suivant l'une des trois modalités, dans un ordre aléatoire et également réparti entre les 3 modalités. Les 18 stimulus sont distincts des stimulus utilisés dans l'expérience.

Cette séance d'entraînement permet de familiariser les sujets à l'usage du stylet et de la tablette, mais surtout permet de bien comprendre le déroulement de l'expérience afin qu'ils soient autonomes et commettent un minimum d'erreurs préjudiciables au bon traitement ultérieur des données.

Les conditions expérimentales des 3 expériences sont résumées dans le tableau 5.1.

Conditions	**Expérience 1**	**Expérience 3**	**Expérience 2**
Protocole	2 notes	3 notes	6 - 7 notes
Modalité d'imitation	Tablette+Audio Tablette Seule Voix	Tablette+Audio	Tablette+Audio Tablette Seule Voix
Tempo(bpm)	120	120 - 180 - 240	120
Nombre d'écoutes	1	1	≥ 1
Nombre d'essais	3	3	≥ 3

Table 5.1 – *Résumé des conditions expérimentales*

5.3 Analyse des données

5.3.1 Extraction des hauteurs de notes atteintes

A la suite des expériences, on dispose pour chaque sujet et chaque stimulus d'un fichier audio WAVE contenant l'enregistrement de l'ensemble des essais si la modalité d'imitation est *Voix*, et d'un fichier Texte contenant à chaque ligne pression et position du stylet sur la tablette ainsi que l'instant temporel associé.

Il convient d'en extraire les hauteurs de notes réalisées par le sujet afin de les comparer avec les notes cibles du stimulus. Les hauteurs de notes du stimulus sont directement accessibles par la partition. Les imitations vocales et chironomiques doivent être récupérées via les données enregistrées, puis les hauteurs de notes atteintes doivent être extraites de ces sons et gestes, qui présentent une variabilité assez grande due au geste naturel, aux hésitations et erreurs. Ce post-traitement est réalisé sous Matlab 2007.

On extrait la fréquence fondamentale F_0 issue de l'imitation vocale de chaque stimulus à l'aide de l'algorithme STRAIGHT [KMKdC99] et on la convertit en demi-ton. La fréquence fondamentale visée par le stylet de la tablette est quant à elle donnée directement par la position du stylet suivant l'axe X de la tablette. On dispose alors d'un signal représentant l'intonation pour chaque modalité d'imitation, chaque sujet et chaque stimulus. Les différents essais que comporte chaque stimulus sont segmentés en utilisant l'information de pression du stylet : selon la consigne exigée, relever le stylet (donc passer d'une pression non nulle à nulle) signifie qu'on passe à l'essai suivant. Un filtre passe-bas est alors appliqué à chacun des essais

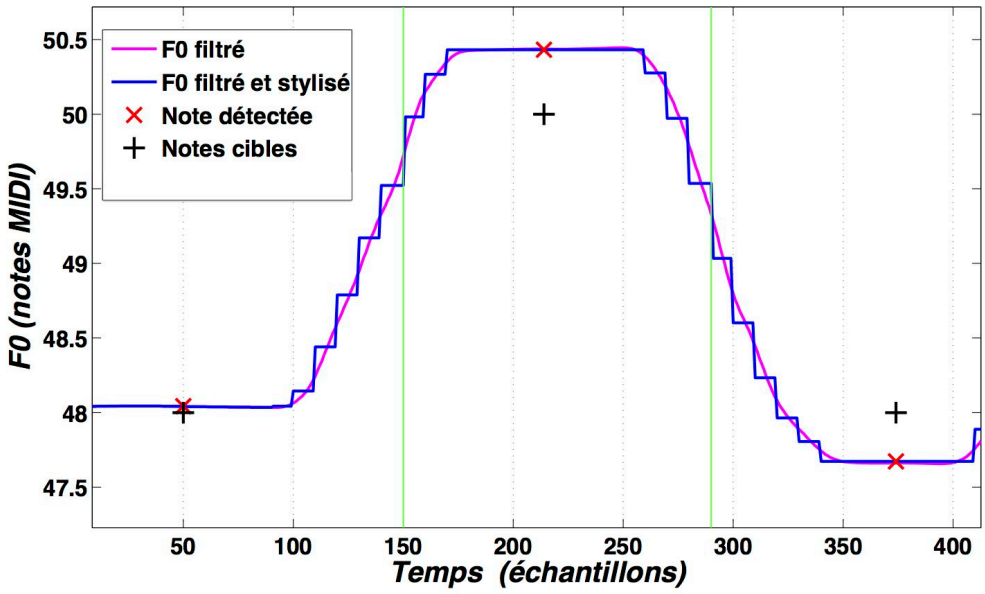


Figure 5.6 – *Courbe stylisée de hauteur de note issue d'une imitation à la tablette*

pour retirer toute variation trop rapide pour être utile dans notre étude. Pour déterminer la note jouée, nous considérons qu'elle correspond à la portion du signal la plus stable dans le temps, étant donné qu'elle s'oppose aux portions transitoires entre deux notes. Pour cela, nous procédons comme il suit :

1. le signal est divisé en segments de 10 ms sur lesquels est calculée la moyenne de F_0 ;
2. si la différence des F_0 ainsi moyennés de deux segments adjacents est inférieure à un seuil (typiquement 50 cents pour la voix et 0.1 demi-ton pour la tablette), on récupère leur moyenne et on les regroupe en un nouveau segment, qui à son tour est comparé au suivant sur l'axe des temps jusqu'à la fin du signal ;
3. l'opération est répétée jusqu'à ce que tous les F_0 moyens des segments soient séparés de leurs voisins d'au moins le seuil ;
4. les N segments les plus longs correspondent alors au N notes de l'essai ;

Une vérification manuelle a été faite pour l'ensemble des segmentations et détections des notes. Un changement des valeurs de paramètres des algorithmes de segmentation et de stylisation était parfois nécessaire pour obtenir un résultat convenable (une note détectée par palier) notamment pour l'extraction des notes de la voix. Tous les essais où la tâche demandée n'était pas réalisée ont été supprimés, c'est à dire ceux dont le nombre de notes total était différent de celui de la séquence cible, ou si un saut d'octave se présentait.

La figure 5.6 est un exemple d'extraction de F_0 à partir d'un signal issu de la tablette pour un stimulus de 7 notes. Les cibles théoriques sont données par des $+$ alors que les $\times$ représentent les notes atteintes ainsi calculées. La courbe rose est le F_0 filtré par un passe-bas, la courbe bleue le F_0 stylisé, et les droites vertes verticales correspondent aux séparations des différentes notes.

Pour chaque essai des stimulus de tous les sujets, nous avons alors une liste des hauteurs de notes atteintes, à comparer avec les hauteurs de notes cibles à l'aide des mesures de justesse

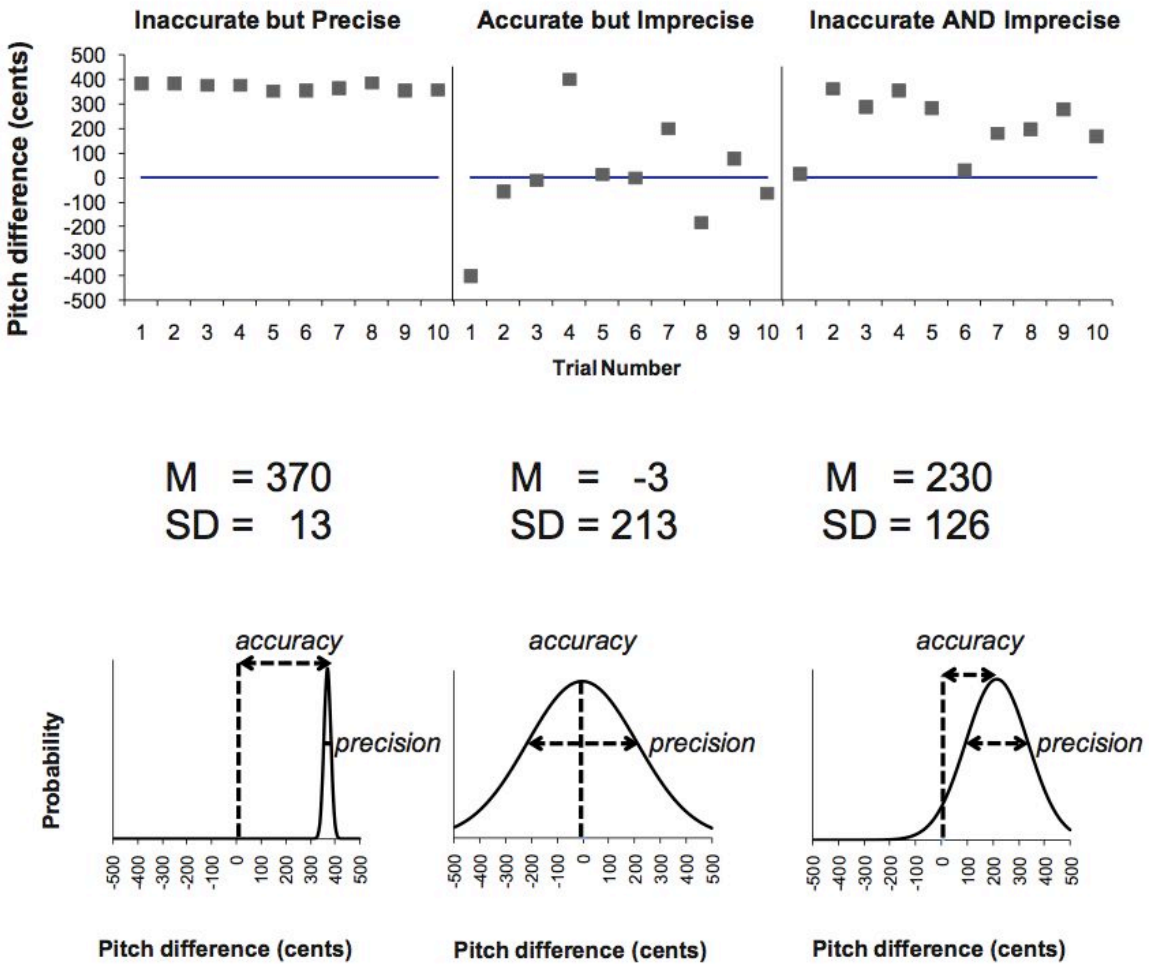


Figure 5.7 – *Justesse et précision selon Pfordresher et al. [PBM+10]. M indique la moyenne et SD l'écart-type des distributions.*

et de précision, comme indiqué dans la partie suivante.

5.3.2 Mesure de la justesse et de la précision : définitions générales

La justesse et la précision sont deux mesures permettant d'évaluer la reproduction d'une ligne mélodique. On peut s'intéresser à la hauteur des notes ou à l'intervalle entre deux notes. La justesse mesure l'écart de la hauteur de note réalisée à celle visée (ou l'intervalle réalisé à celui visé), tandis que la précision mesure la faculté à reproduire avec constance une hauteur de note donnée (ou un intervalle donné). Suivant l'étude qu'on veut entreprendre, l'ensemble des notes sur lequel la justesse et la précision sont calculées peut varier. La figure 5.7 schématise les notions de justesse et précision.

La *justesse de note* correspond à la moyenne des distances entre hauteurs de notes réalisées et visées. Si N_E est le nombre de notes indexées par i dans l'ensemble E considéré, S_i est le F_0 atteint par la note et T_i le F_0 cible, la justesse $Jnote_E$ est donnée par l'expression :

$$Jnote_E = \frac{\sum_i^{N_E}(S_i - T_i)}{N_E} \tag{5.1}$$

Ainsi, chanter ou jouer en moyenne trop haut impliquera que $Jnote_E$ soit plus grande que le seuil de perception, tandis que chanter ou jouer en moyenne trop bas donnera $|Jnote_E|$ inférieure au seuil de perception, le seuil de perception de la différence entre deux F_0 de voyelles chantées se situant en dessous de 10 cents [FS58]. Chanter ou jouer avec la bonne hauteur en moyenne, équivaut à $|Jnote_E|$ inférieure au seuil de perception.

La *justesse d'intervalle* est la moyenne des distances entre intervalles réalisés et visés, dans l'ensemble des intervalles formé par l'ensemble E de notes considérées :

$$Jint_E = \frac{\sum_i^{N_E-1}(|S_{i+1} - S_i| - |T_{i+1} - T_i|)}{N_E - 1} \tag{5.2}$$

Une valeur absolue supérieure au seuil de perception sera l'indication d'un intervalle trop grand entre les deux notes (en moyenne), tandis qu'une valeur absolue inférieure signifiera en moyenne un intervalle trop petit par rapport à l'intervalle cible. Un intervalle parfaitement juste vérifie $|Jint_E|$ inférieure au seuil de perception.

La justesse ainsi définie correspond à un biais et doit s'accompagner d'une autre mesure sur la variance des notes obtenues, ce qu'on appellera ici *précision*.

La *précision de note* correspond à l'écart-type des erreurs entre les hauteurs de notes réalisées et ciblées. Si N_E est le nombre de notes indexées par i dans l'ensemble E considéré, S_i est le F_0 atteint par la note et T_i le F_0 cible, la précision est donnée par l'expression :

$$Pnote_E = \sqrt{\frac{\sum_i^{N_E}(S_i - T_i - Jnote_E)^2}{N_E}} \tag{5.3}$$

La précision est donc toujours positive. Une valeur inférieure au seuil de perception indique qu'on reproduit les notes de l'ensemble avec une justesse parfaitement similaire, et une valeur supérieure sera la preuve de réalisations à justesse d'autant plus irrégulière qu'elle sera élevée.

La *précision d'intervalle* correspond à l'écart-type des erreurs entre les intervalles réalisés et visés. Si N_E est le nombre de notes indexées par i dans l'ensemble E considéré, S_i est le F_0 atteint par la note et T_i le F_0 cible, la précision est donnée par l'expression :

$$Pint_E = \sqrt{\frac{\sum_i^{N_E-1}\left(|S_{i+1} - S_i| - |T_{i+1} - T_i| - Jint_E\right)^2}{N_E - 1}} \tag{5.4}$$

Une valeur inférieure au seuil de perception équivaut à la bonne réalisation des intervalles.

Notons cependant que tout ensemble de notes aléatoires centrées autour de 0 donnera une justesse d'autant meilleure que le nombre de notes est important, puisque la moyenne tendra vers 0. Obtenir une justesse proche de 0 ne nous permet pas de tirer de conclusion sur la qualité de la reproduction de la mélodie. C'est plutôt une justesse qui a tendance à avoir une valeur éloignée du seuil de perception qui donnerait une information sur une certaine qualité de l'imitation mélodique. La justesse doit donc toujours être accompagnée de la mesure de la précision pour juger de la qualité de reproduction des hauteurs de note d'une mélodie. La précision a quand à elle une valeur intrinsèque, car elle signifie qu'on joue juste relativement à une hauteur fixe, même si celle-ci n'est pas la bonne, ce qui a du sens en musique, surtout quand on joue seul.

Types d'ensembles de notes

Pour calculer la justesse et la précision, Pfordresher *et al.* [PBM+10] utilisent des ensembles de notes distincts pour chacune des deux mesures. Pour chaque participant, la justesse est calculée sur toutes les notes d'une séquence mélodique, tandis que la précision est obtenue en prenant, comme Ternström et Sundberg [TS88], l'ensemble des notes de même hauteur à travers toutes les réalisations d'un sujet.

Dans notre étude, la précision est obtenue en prenant pour référence la justesse comme indiquée aux équations 5.3 et 5.2, elle-même définie sur des notes appartenant à divers ensembles. Ces divers types d'ensembles (sujets, stimulus, distances à la note précédente), nous permettent d'étudier l'influence de certains paramètres d'une manière plus complète. Ils sont donnés au tableau 5.2.

Type d'ensembles	**Nombre d'ensembles**	**Nombre de mesures pour le calcul des justesses/précisions**	
		de note	**d'intervalle**
Sujets	20 par modalité	57	40
Stimulus expériences 1, 2	17 par modalité	40, 120, ou 140 selon le stimulus	20, 100, ou 120 selon le stimulus
Distances	16 par modalité	20 à 120 selon la distance, 340 pour ens. des notes sans préc.	20 à 120 selon la distance
Stimulus expérience 3 et sujets	336 par tempo	3	2

Table 5.2 – *Résumé du nombre d'ensembles et de mesures pour chaque type d'ensembles utilisés*

Le premier type d'ensembles permet d'étudier la distribution des justesses et précisions des différents sujets selon les trois modalités d'imitation (*Voix*, *Tablette Seule*, et *Tablette + Audio*). On forme un ensemble par sujet et par modalité d'imitation, soit 20 ensembles pour chaque modalité (car on a 20 sujets). Chaque stimulus représente une séquence de 2, 6 ou 7 notes. On calcule les justesses et précisions pour chaque ensemble à partir d'un total de 57 notes pour les justesses et précisions *de note* (12 stimulus de 2 notes + 2 stimulus de 6 notes + 3 stimulus de 7 notes) et 40 intervalles pour les justesses et précisions *d'intervalles* (12 stimulus de 1 intervalle + 2 stimulus de 5 intervalles + 3 stimulus de 6 intervalles). Les notes (resp. les intervalles) de chaque ensemble correspondent à toutes les notes (resp. tous les intervalles) jouées par un sujet.

Le deuxième type d'ensembles nous permet d'obtenir la distribution des justesses et des précisions calculées pour chacun des stimulus pour une même modalité, et ainsi de regarder si la longueur de la séquence à imiter a une influence sur la justesse et la précision, suivant les 3 modalités d'imitation. Ce type d'ensembles est celui des stimulus (i.e. séquences de notes à reproduire) de l'expérience 1 et 2, tous sujets confondus. On forme 17 ensembles (12 stimulus de 2 notes de l'expérience 1 + 5 stimulus de 6 ou 7 notes de l'expérience 2) pour chacune

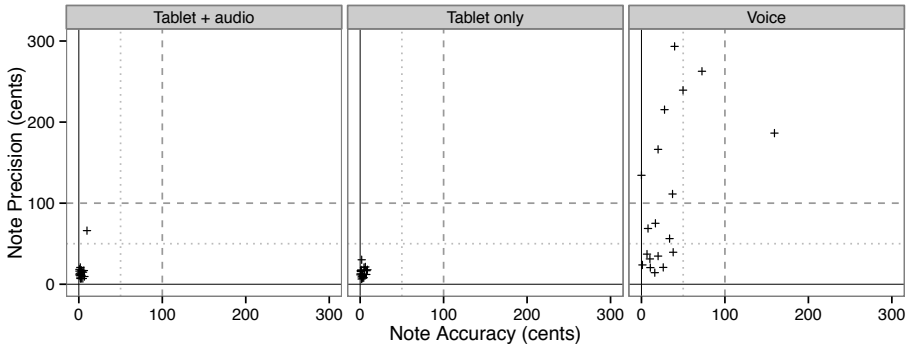
des 3 modalités. On dispose alors de 40 notes et 20 intervalles pour chacun des 3 ensembles de stimulus de 2 notes (2 notes × 20 sujets, 1 intervalle × 20 sujets), de 120 notes et de 100 intervalles pour chacun des 3 ensembles de stimulus de 6 notes (6 notes × 20 sujets, 5 intervalles × 20 sujets), et de 140 notes et de 120 intervalles pour chacun des 3 ensembles de stimulus de 7 notes (7 notes × 20 sujets, 6 intervalles × 20 sujets).

Troisièmement, on peut comparer les distributions des 3 modalités d'imitation en prenant comme type d'ensembles les distances de la note cible à la précédente, s'étalant de -12 à 12 demi-tons. En effet, la taille et le sens du mouvement pour produire un intervalle pourraient influer sur la justesse et la précision de la note qui suit. On dispose alors de 16 ensembles par modalité (car on a 15 distances différentes de la note cible précédente + l'ensemble des notes sans note précédente). Les 15 distances à notre disposition à travers les stimulus de 2, 6 et 7 notes issues des expériences 1 et 2 sont, avec entre parenthèses le nombre de notes associées par sujet et par modalité : 12 (2), 9 (3), 7 (2), 5 (4), 4 (2), 2 (5), 1 (3), -1 (1), -2 (5), -3 (2), -4 (2), -5 (3), -7 (3), -9 (1), -12 (2) demi-ton(s). Nous avons donc de 1 à 6 notes par sujet et par modalité pour chaque distance (sauf pour l'ensemble des notes sans précédentes qui se compte à 17), soit, en multipliant par le nombre de sujets, de 20 à 120 notes par ensemble selon la distance à la note cible précédente (340 pour l'ensemble des sans note précédente). Pour le calcul de la justesse et précision d'*intervalle*, nous obtenons le même nombre d'intervalles par ensemble, mais nous avons l'ensemble des notes sans précédentes en moins puisque aucun intervalle ne peut être calculé.

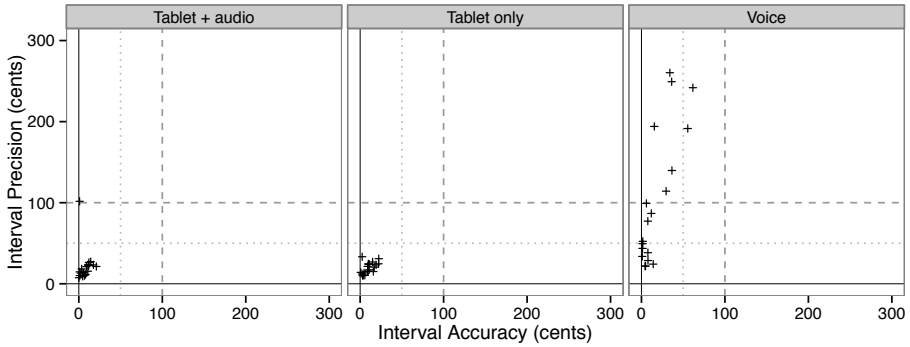
Enfin, l'expérience 3 permet d'étudier l'influence du tempo sur la justesse et la précision en ce qui concerne la modalité *Tablette + Audio*. Le type d'ensembles étudié est celui des stimulus de l'expérience 3 et des sujets. Nous disposons de 336 ensembles par tempo (12 stimulus × 28 sujets, pour chacun des tempos à 120, 180 ou 240 battements par minute), chacun composé de 3 notes, soit de 2 intervalles.

5.4 Résultats

La précision et la justesse mesurant bien à eux deux la qualité de la reproduction de hauteur de notes, et de façon à mettre en évidence nos résultats, nous traçons 3 types de graphiques. Le premier comprend les diagrammes composés des dimensions de la précision et de la valeur absolue de la justesse pour chaque modalité d'imitation, et qu'on donne pour différents types d'ensembles de calculs. Plus la distance de chaque mesure est proche de l'origine, plus l'imitation mélodique est bien reproduite. Ils permettent de comparer les résultats de voix avec l'étude de Pfordresher *et al.* [PBM+10]. Le deuxième type de graphique est la justesse ou la précision pour chaque modalité d'imitation, avec la légende des points de mesure pour chacun des types d'ensembles de calcul. Ils font intervenir une justesse pouvant être négative, et permettent une analyse plus fine que les représentations de Pfordresher *et al.* Le dernier type de représentation graphique est la boîte à moustaches des distributions des justesses et précisions pour chaque modalité d'imitation. Ils permettent une vision d'ensemble des résultats, et illustrent l'étude statistique que nous menons pour confronter les tendances de résultat à leur signification statistique. On commence par étudier principalement les effets des modalités d'imitation sur la justesse et la précision, puis ceux des tempos imposés au sujet pendant l'imitation.



(a) *Justesses et précision de note*



(b) *Justesses et précision d'intervalle*

Figure 5.8 – *Diagrammes justesse-précision des sujets (1 sujet = 1 point du graphe) des expériences 1 et 2 mélangées, pour chaque modalité d'imitation*

5.4.1 Effets de la modalité d'imitation (expérience 1 et 2)

Pour rappel, 3 modalités d'imitation des séquences de notes sont proposées dans l'expérience 1 et 2 : par sa propre voix (*Voix*), par la tablette seule munie de ses repères de notes (*Tablette Seule*), et par la tablette munie de ses repères associée à un retour audio (*Tablette + Audio*). Pour chaque modalité, on calcule la justesse / précision de note / intervalle sur chaque ensemble statistique.

a) Analyse par sujet

Les sujets présentaient des profils variés par diverses compétences musicales ou vocales, et nous proposons de comparer la distribution de leur justesse et précision pour chaque modalité d'imitation. La figure 5.8 présente en haut la distribution de chacun des 20 sujets dans l'espace des précision et justesse de note, tandis qu'en bas, sont données les précision et justesse d'intervalle. Les lignes en pointillés allongés indiquent le seuil de 1 demi-ton (100 cents) de justesse et de précision utilisé par Pfordresher *et al.* [PBM+10], tandis que celles en pointillés resserrés correspondent au seuil plus musical de 1/2 de demi-ton (50 cents).

Le meilleur sujet à la voix tend à dégager une précision du même ordre de grandeur que celle des sujets avec les modalités *Tablette Seule* et *Tablette + Audio*. Environ la moitié des sujets à la voix réussissent moins bien en justesse et précision que les plus mauvais sujets à la tablette. Quand on regarde la différence des distributions entre justesse de note

et justesse d'intervalle pour chaque modalité d'imitation, il est intéressant de noter que la voix est meilleure en intervalle qu'en note, tandis que les tablettes sont meilleures en note qu'en intervalle. On peut sans doute expliquer ce résultat par le fait que la tablette dispose de marqueurs visuels, favorisant la visée de la cible indépendamment de l'intervalle, alors qu'avec la voix, on prête plus attention au retour audio (voix naturelle ici), donc privilégiant la reproduction de l'intervalle, a priori plus facile à mémoriser que des hauteurs de notes absolues.

Le tableau 5.3 catégorise les sujets suivant leur justesse / précision ou non-justessse / imprécision de note, en considérant le seuil d'un demi-ton (100 cents) ou le seuil musical d'un quart de ton (50 cents). En considérant le seuil à 50 cents, hormis un sujet dont la précision de note est comprise entre 50 et 100 cents, tous les sujets jouent juste et précis en note (et intervalle) avec les modalités utilisant la tablette, contre seulement 40% des sujets dans le cas de la voix.

Voix	Précis	Imprécis	
Juste	11 (55)	8 (40)	19 (95)
Non juste	0 (0)	1 (5)	1 (5)
	11 (55)	9 (45)	
Tablette + Audio	Précis	Imprécis	
Juste	20 (100)	0 (0)	20 (100)
Non juste	0 (0)	0 (0)	0 (0)
	20 (100)	0 (0)	
Tablette Seule	Précis	Imprécis	
Juste	20 (100)	0 (0)	20 (100)
Non juste	0 (0)	0 (0)	0 (0)
	20 (100)	0 (0)	

(a) *Seuil à 1 demi-ton*

Voix	Précis	Imprécis	
Juste	8 (40)	9 (45)	17 (85)
Non juste	0	3 (15)	3 (15)
	8 (40)	12 (60)	
Tablette + Audio	Précis	Imprécis	
Juste	20 (100)	0 (0)	20 (100)
Non juste	0 (0)	0 (0)	0 (0)
	20 (100)	0 (0)	
Tablette Seule	Précis	Imprécis	
Juste	19 (95)	1 (5)	20 (100)
Non juste	0 (0)	0 (0)	0 (0)
	19 (95)	1 (5)	

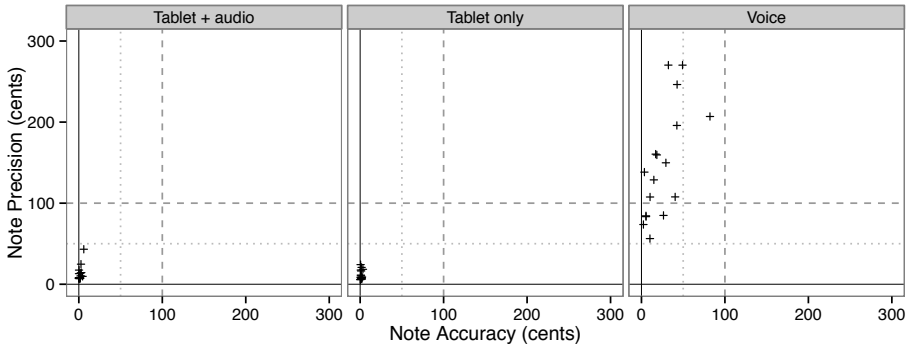
(b) *Seuil à 1/4 de ton*

Table 5.3 – *Nombre (en % entre parenthèses) de sujets (sur 20) dans les catégories de justesse et précision de note pour chaque modalité d'imitation, selon 2 différents seuils*

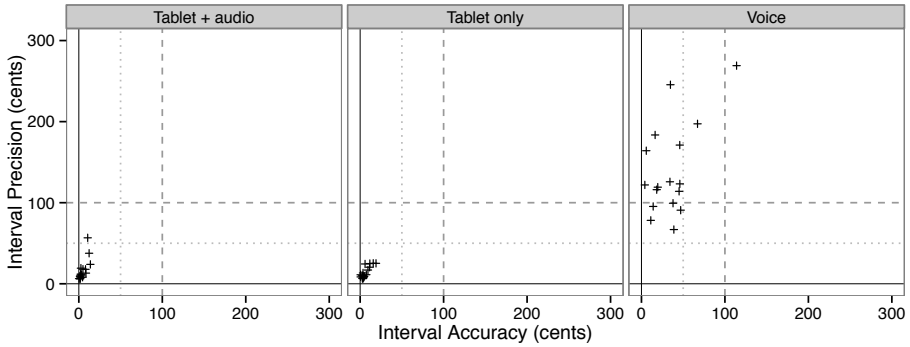
Dans le cas de la modalité *Voix*, nous obtenons le même ordre de grandeur de chanteurs justes que Pfordresher *et al.* [PBM+10] et [DBGP07] (seuil à 1 demi-ton) : 95% contre respectivement 87% et 88%. Le nombre de sujets chantant précisément diffère légèrement avec Pfordresher *et al.* : 55% Précis vs 44%, ce qui peut être dû au fait que la moitié de nos sujets sont musiciens contre aucun dans les participants de Pfordresher *et al.* D'autre part, nous obtenons une médiane de la justesse d'intervalle valant 3 cents, plus petite que celle de la justesse de note valant 20 cents. Cela va dans le sens de l'étude de Pfordresher et Brown [PB07] et dans le sens contraire à celle de Pfordresher *et al.* [PBM+10] où les justesses de note sont meilleures que les justesses d'intervalle.

b) Analyse par stimulus

La figure 5.9 donne en haut le diagramme justesse-précision de note pour les 3 modalités, tandis qu'elle donne en bas celui d'intervalle. Chaque point représente un stimulus (i.e. une séquence identique de notes) tous sujets confondus. Alors que les modalités utilisant les tablettes forment des résultats similaires à l'ensemble des sujets, la performance de la voix est dégradée par rapport à l'ensemble des sujets : avec le seuil à 50 cents, la voix ne donne aucune mesure à la fois juste et précise, contre 40% avec les sujets comme ensemble de calcul. Le tableau 5.4 donne les répartitions des stimulus pour les justesse et précision de note. Enfin,



(a) *Justesses et précision de note*



(b) *Justesses et précision d'intervalle*

Figure 5.9 – *Diagrammes justesse-précision pour des stimulus (1 stimulus = 1 point du graphe) des expériences 1 et 2, pour chaque modalité d'imitation*

on retrouve comme avec les ensembles des sujets et la modalité utilisant les tablettes, que la justesse de note est meilleure que la justesse d'intervalle. Ce résultat est illustré par un zoom de la figure 5.9 à la figure 5.10. On y remarque la très bonne valeur de précision en utilisant des tablettes, de médianes 8-9 cents, valeurs à la limite des possibilités de discrimination de l'oreille humaine.

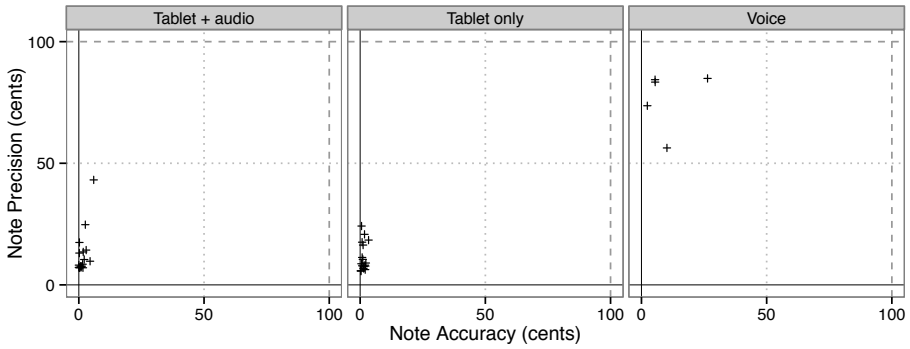
Cet ensemble peut permettre de voir notamment si les séquences de 2 notes sont mieux reproduites que celles de 6 et 7 notes, car a priori plus difficiles à reproduire vue leur longueur, suivant chaque modalité d'imitation. Pour cela nous traçons la justesse et la précision en annotant les mesures de leur légende aux figures 5.11 et 5.12. Concernant la voix, il n'apparaît pas de tendance claire. Par contre, si on regarde les tablettes, la quasi-totalité des mélodies de 6-7 notes constitue les plus mauvaises performances (séparées de 10-15 cents des séquences de 2 notes), à l'exception de la justesse de note où les mesures sont toutes quasi nulles (i.e. notes statistiquement centrées autour de zéro).

c) Analyse par distance à la note cible précédente

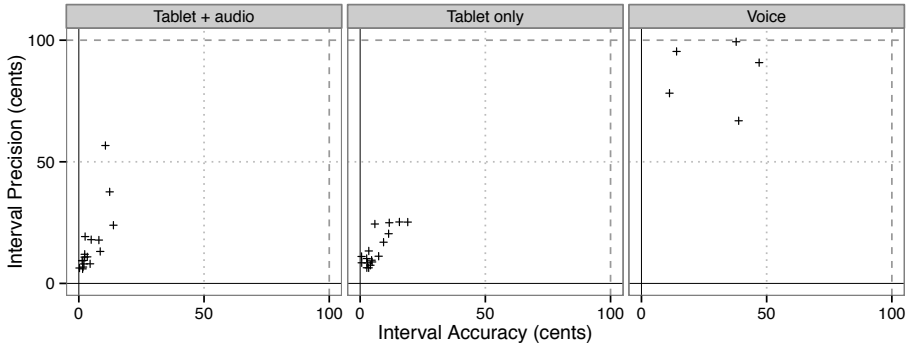
En prenant les intervalles comme type d'ensembles, c'est à dire la distance entre la note précédente et la note mesurée, on obtient la même tendance qu'avec les ensembles de stimulus. La figure 5.13 présente les diagrammes précision-justesse de note et d'intervalle. Pour les notes, chaque point correspond aux distances identiques à la note précédente. Pour la justesse/précision d'intervalle, chaque point correspond aux intervalles identiques. Dans le

Voix	Précis	Imprécis	
Juste	0 (0)	16 (94)	16 (94)
Non juste	0 (0)	1 (6)	1 (6)
	0 (0)	17 (100)	
Tablette + Audio	Précis	Imprécis	
Juste	17 (100)	0 (0)	17 (100)
Non juste	0 (0)	0 (0)	0 (0)
	17 (100)	0 (0)	
Tablette Seule	Précis	Imprécis	
Juste	17 (100)	0 (0)	17 (100)
Non juste	0 (0)	0 (0)	0 (0)
	17 (100)	0 (0)	

Table 5.4 – *Nombre (en % entre parenthèses) de stimulus (sur 17) dans les catégories de justesse et précision de note pour chaque modalité d'imitation, en se basant sur le seuil à 1/2 demi-ton*



(a) *Justesses et précision de note*



(b) *Justesses et précision d'intervalle*

Figure 5.10 – *Diagrammes justesse-précision des stimulus (1 stimulus = 1 point du graphe) des expériences 1 et 2, pour chaque modalité d'imitation, réduits à l'intervalle [0-100 cents]*

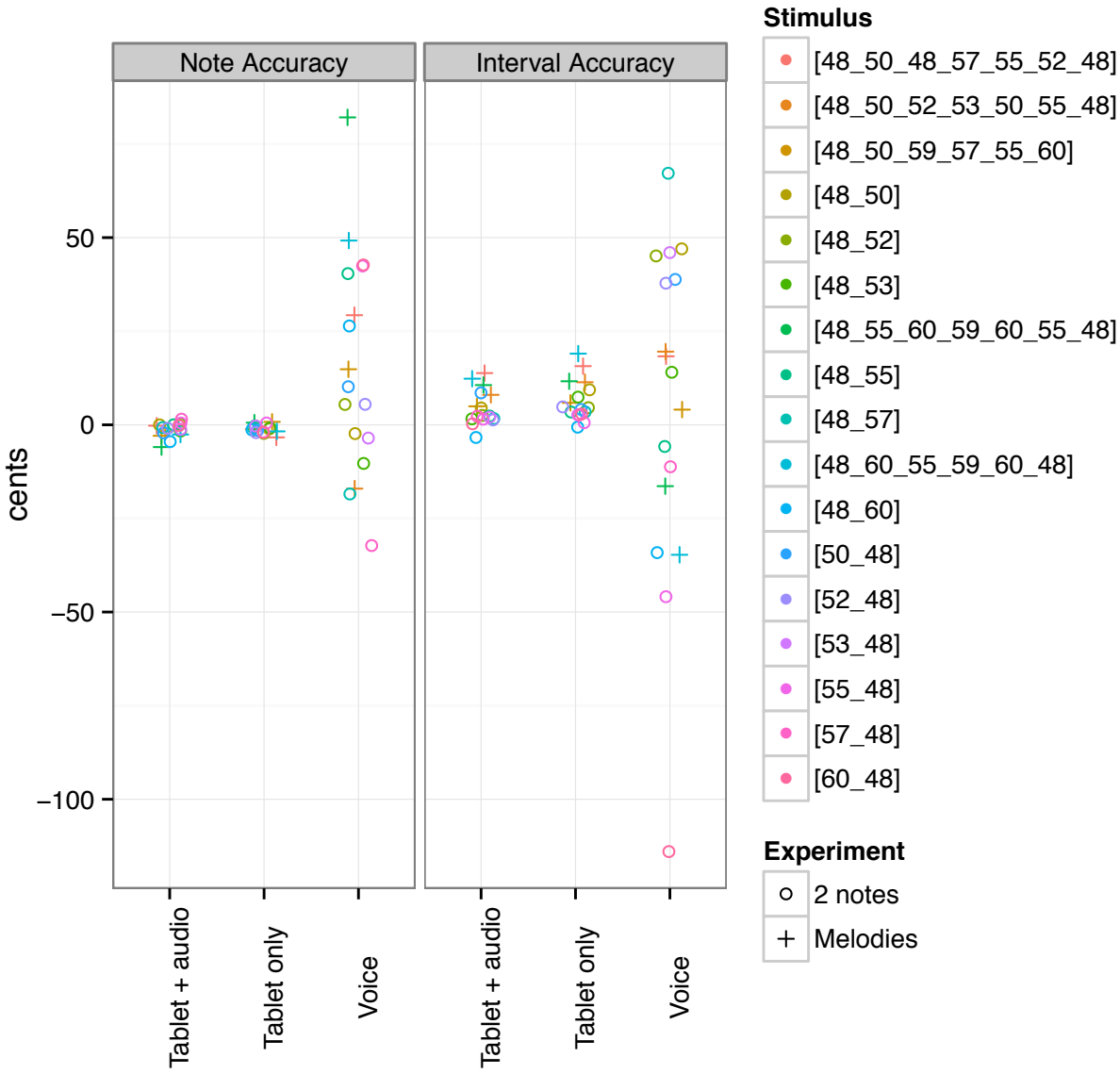


Figure 5.11 – *Justesses de note et d'intervalle, pour chaque mélodie des expériences 1 et 2, pour chaque modalité d'imitation. Les séquences de nombres indiqués pour chaque stimulus sont les notes MIDI du stimulus.*

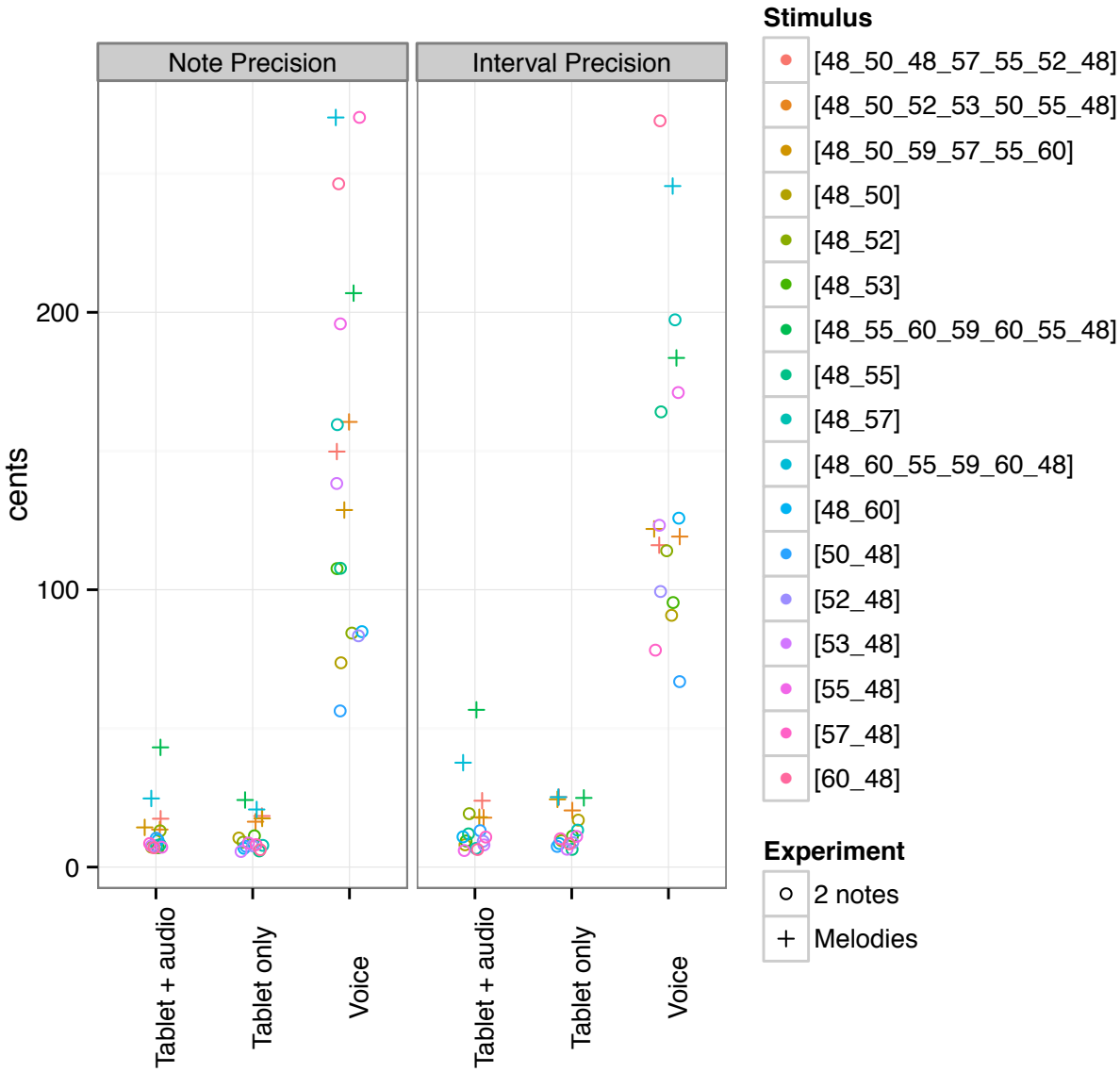
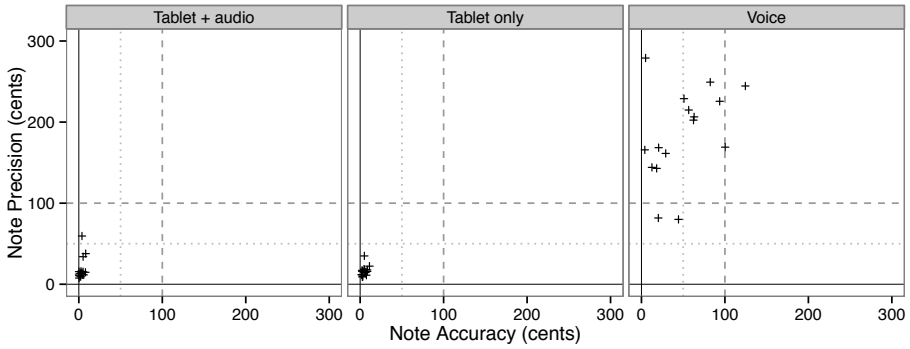
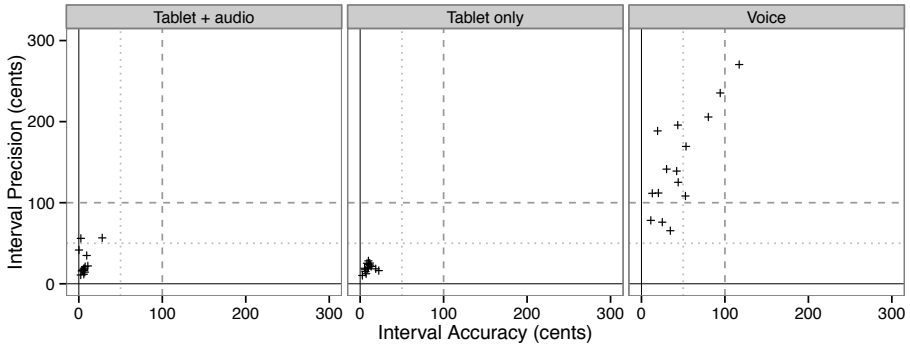


Figure 5.12 – *Précisions de note et d'intervalle, pour chaque mélodie des expériences 1 et 2, pour chaque modalité d'imitation. Les séquences de nombres indiqués pour chaque stimulus sont les notes MIDI du stimulus.*



(a) *Justesse et précision de note*



(b) *Justesse et précision d'intervalle*

Figure 5.13 – *Diagrammes justesse-précision pour chaque distance à la note cible précédente (des expériences 1 et 2) et pour chaque modalité d'imitation*

cas des justesses et précisions d'intervalle, chaque point correspond aux mêmes intervalles. En particulier on n'observe aucune mesure juste et précise pour la voix avec le seuil de 50 cents. Le tableau 5.5 donne les résultats sous forme catégorique avec la mesure de justesse et de précision de note.

En différenciant graphiquement chacune des mesures, on n'observe pas de corrélation entre bonne précision et faible distance à la note précédente, ce à quoi on aurait pu s'attendre concernant les tablettes où la distance à la note précédente se traduit par la distance du mouvement précédent la note à parcourir avec le stylet. Par contre, une légère tendance de corrélation apparaît entre justesse de note et sens de l'intervalle, indiqué par un + pour les intervalles montants et un ∇ pour les intervalles descendants. Ceux-là apparaissent séparés d'environ 5 cents de part et d'autre de la justesse de valeur nulle (on s'approche de l'incertitude de perception de l'oreille humaine). Les intervalles descendants ont une justesse négative, tandis que les intervalles montants une justesse positive. Cela signifie que le parcours du stylet a tendance à aller très légèrement au-delà de la cible visée, quel que soit le sens du mouvement. La raison réside sans doute dans l'inertie du mouvement que les sujets anticiperaient mal.

d) Étude statistique pour la signification des résultats

Afin d'évaluer la signification statistique des résultats de précisions et de justesses obtenues pour chaque modalité d'imitation, les différences entre paires de distribution sont testées par un test U de Mann-Whitney. La procédure *wilcox.test()* du logiciel R a été utilisée. Une

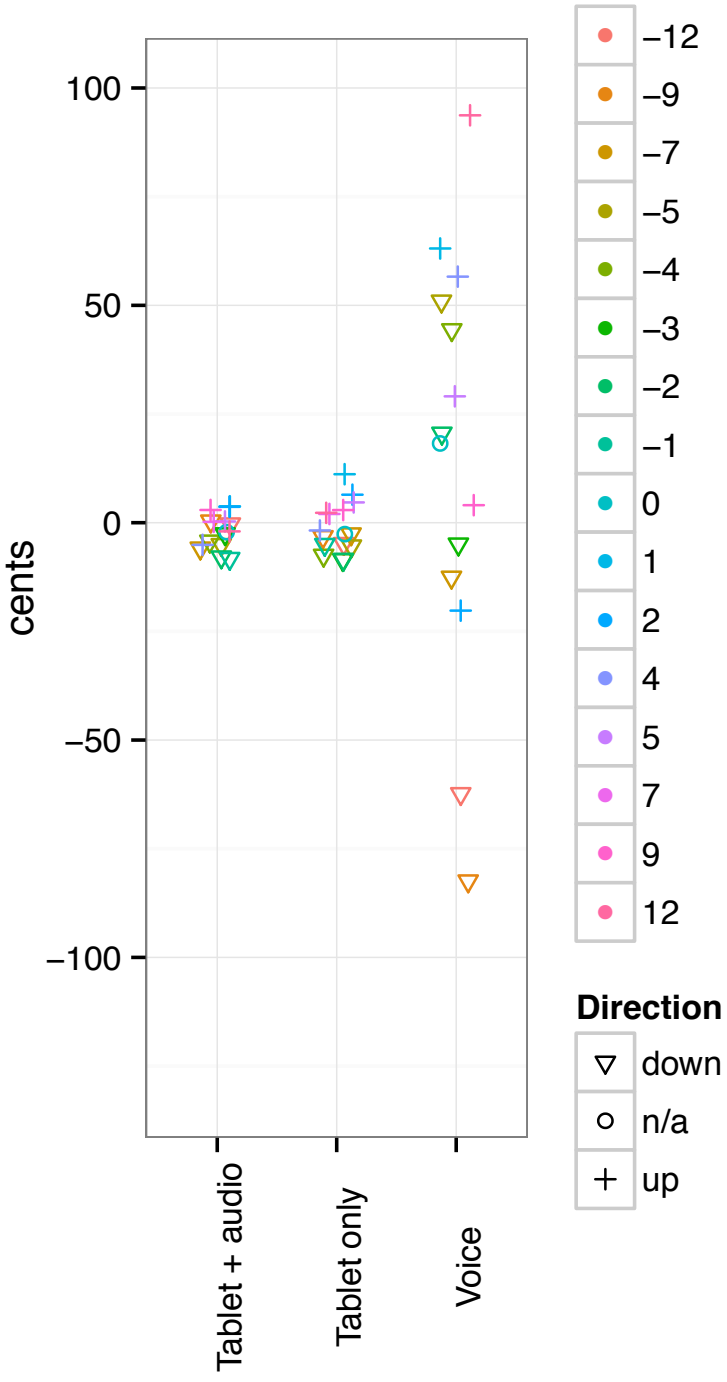


Figure 5.14 – *Justesse de note pour chaque distance à la note cible précédente (en demi-ton), pour chaque modalité d'imitation.*

Voix	Précis	Imprécis	
Juste	0 (0)	8 (50)	8 (50)
Non juste	0 (0)	8 (50)	8 (50)
	0 (0)	16 (100)	
Tablette + Audio	Précis	Imprécis	Somme
Juste	15 (94)	1 (6)	16 (100)
Non juste	0 (0)	0 (0)	0 (0)
Somme	15 (94)	1 (6)	
Tablette Seule	Précis	Imprécis	
Juste	16 (100)	0 (0)	16 (100)
Non juste	0 (0)	0 (0)	0 (0)
	16 (100)	0 (0)	

Table 5.5 – *Nombre (en % entre parenthèses) de distances à la note cible précédente (sur 16) dans les catégories de justesse et précision de note pour chaque modalité d'imitation, en se basant sur le seuil à 1/2 demi-ton*

différence significative se traduit par un $p < 0.05$ et une différence faiblement significative par un $0.05 < p < 0.1$. Les distributions des justesses et précisions sont tracées sur les figures 5.15 et 5.16. Sur ces graphiques sont également indiquées la ou les signification(s) statistique(s) associée(s) : forte en bleue, faible en orange, et nulle en rouge. Quand un seul p est indiqué sur un graphe, il correspond à la signification statistique entre la modalité d'imitation à la voix d'une part et celle aux tablettes avec ou sans audio d'autre part. Si deux p sont indiqués, alors ils désignent la signification statistique pour chacun des couples « Voix - Tablette Seule » et « Voix - Tablette + Audio ».

Le premier résultat est la non différenciation des distributions des justesses et précisions de la modalité *Tablette + Audio* et *Tablette Seule* ($p > 0.1$). On ne peut donc rien déduire sur la potentielle aide du retour audio sur la justesse et précision de jeu. On peut avancer plusieurs explications. Tout d'abord, les résultats d'imitation avec la tablette seule sont très bons : la médiane des justesses s'étale de -3 à 10 cents, et celle des précisions de 8 à 19 cents, sachant que le seuil de perception de l'oreille humaine est d'environ 5-10 cents et que 100 cents représentent le plus petit intervalle dans une gamme chromatique. Ainsi l'audio est presque inutile pour cette tâche, la tablette seule permettant déjà d'atteindre de très bons scores. Une deuxième explication est la possible plus forte focalisation des sujets sur la vue plutôt que sur l'audio. En effet, la vue est un sens dominant chez l'Homme et l'audio dans notre expérience nécessite sans doute une connaissance musicale relativement poussée pour être en mesure de l'exploiter, d'autant plus que la tâche demandée est contrainte par un métronome de fréquence supérieure ou égale à 120 battements par minute, ce qui accroît la difficulté.

Le deuxième résultat est la différence significative entre les distributions des précisions de la modalité *Voix* et les modalités utilisant la tablette, *Tablette Seule* et *Tablette + Audio*. Sur la figure 5.16 sont présentées les précisions de notes et d'intervalles pour les 3 modalités d'imitation et chacun des ensembles de calculs, ainsi que l'indication $p < 0.05$ en bleu si la distribution de la modalité *Voix* est significativement différente des deux autres prises individuellement. La présence visuelle des marqueurs de hauteurs de notes sur la tablette pourrait être une source de biais dans la comparaison entre les modalités utilisant la tablette

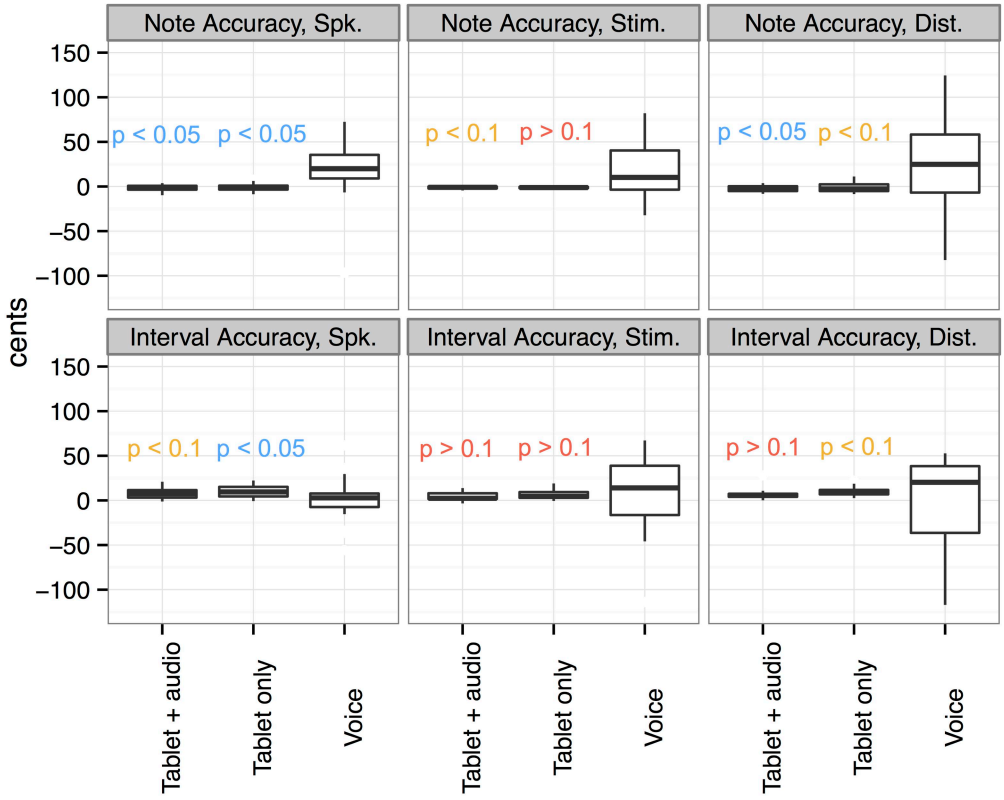


Figure 5.15 – *Distribution de la justesse de note (haut) et justesse d'intervalle (bas) pour chaque ensemble de mesures (sujet, stimulus, distance). Une signification statistique p inférieure à 0.05 indique ici une distribution significativement différente entre la modalité d'imitation à la voix et celles aux tablettes avec ou sans audio, en fonction de la position du p. Pour les valeurs de U associées, se reporter au tableau 5.6.*

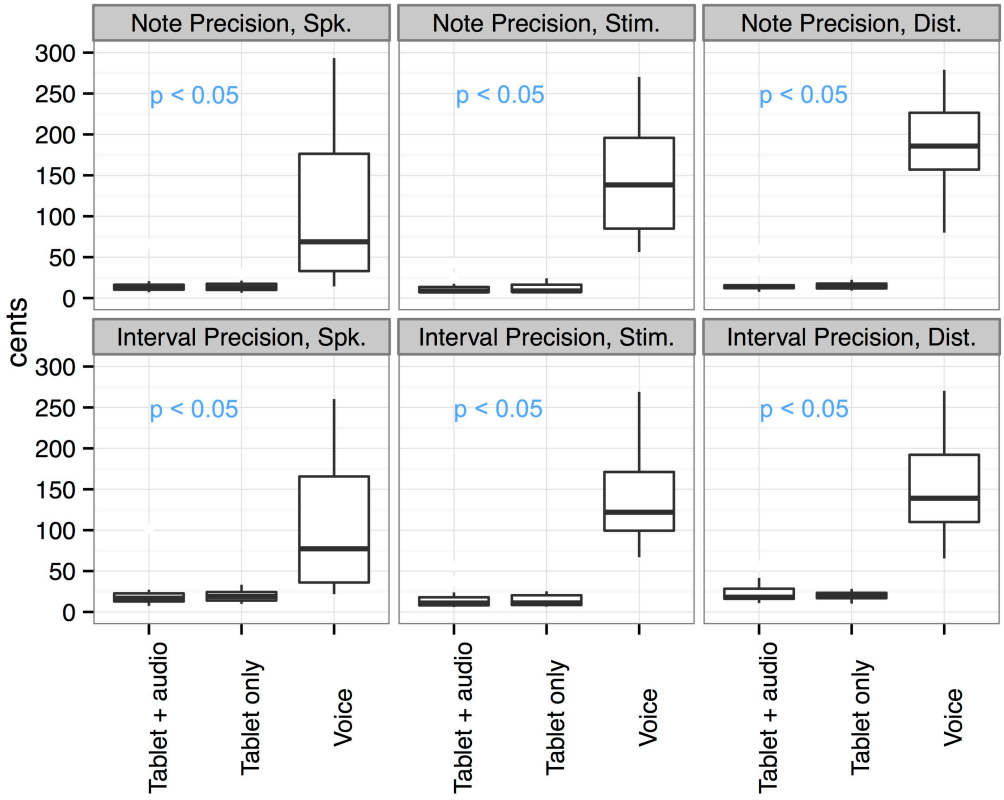


Figure 5.16 – *Distribution de la précision de note (haut) et précision d'intervalle (bas) pour chaque ensemble de mesures (sujet, stimulus, distance). Une signification statistique p inférieure à 0.05 indique ici une distribution significativement différente entre la modalité d'imitation à la voix d'une part et celle aux tablettes avec ou sans audio d'autre part. Pour les valeurs de U associées, se reporter au tableau 5.6.*

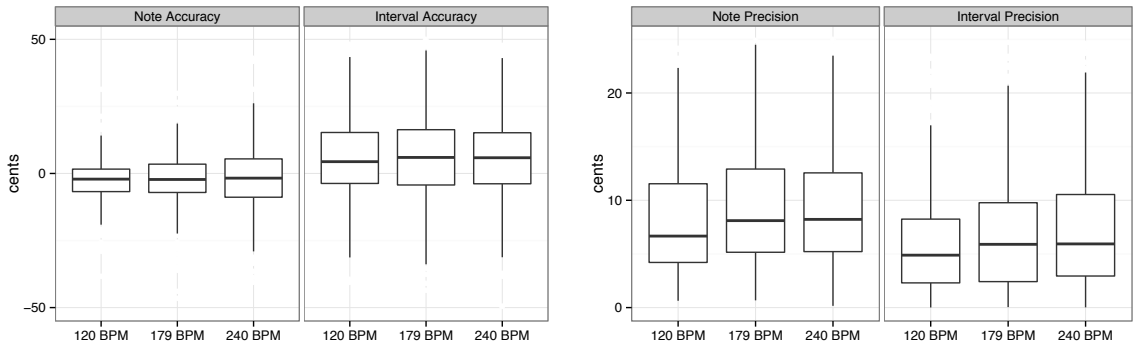


Figure 5.17 – *Distribution de (de gauche à droite) la justesse de note puis d'intervalle et la précision de note puis d'intervalle pour chaque tempo de l'expérience 3.*

et celle utilisant la voix. La tâche d'imitation sur la tablette pourrait alors faire intervenir le sens de la vue de manière prépondérante et aider la réalisation sur la tablette. Pour le détail des statistiques, voir le tableau 5.6 donnant les résultats du test U de Mann-Whitney.

Le troisième résultat se rapporte aux justesses de note et d'intervalle, illustré par la figure 5.15. Sont indiqués les significations statistiques des différences de distribution entre la modalité *Voix* et les modalités des tablettes comme détaillé ci-dessus. Les distributions des justesses de note (première ligne de la figure 5.15) de la modalité *Voix* et des modalités utilisant les tablettes sont significativement différentes avec l'ensemble des sujets. Le reste est faiblement significatif, ou non significatif dans un cas (ensemble des mélodies, entre tablette seule et la voix). Pour les cas significatifs, les résultats sont en défaveur de la voix en ce qui concerne la justesse de note. Quant à la justesse d'intervalle (deuxième ligne de la figure 5.15), la seule différence significative de distribution des justesses d'intervalle apparaît entre la modalité *Tablette Seule* et la modalité *Voix* dans le cas des ensembles de sujet, révélant une meilleure médiane pour la voix que pour la tablette : 3 cents pour la voix contre 10 cents pour la tablette seule. Mais vu le seuil de perception à 5-10 cents, le résultat n'est pas significatif perceptivement.

5.4.2 Effet du tempo (expérience 3)

Lors de l'imitation dans l'expérience 3, le sujet doit reproduire le stimulus entendu en jouant de manière synchronisée avec un métronome audio et visuel. Ce métronome peut battre à 3 fréquences différentes : 120, 180 et 240 battements par minute, mais elle est stable pour chaque stimulus. Les différents tempos nous permettent de mesurer l'influence de la vitesse d'exécution sur la justesse et la précision.

La figure 5.17 représente les distributions de justesses et de précisions de note et d'intervalle. Chaque mesure a été établie pour l'imitation d'un stimulus par un sujet à un tempo, et pour toutes notes/intervalles confondus. On n'observe pas ou très peu de différence de dispersion avec le tempo. La différence des médianes des distributions est de l'ordre du cent. L'effet de la vitesse du geste sur la difficulté à atteindre la cible est donc négligeable. Aussi, on n'a pas atteint de seuil critique de tempo où la précision et/ou la justesse s'effondrerait. Une étude statistique est donc inutile dans ce cas du point de vue musical, puisque l'amplitude des variations est négligeable perceptivement.

Variable	Cond A	Cond B	U	p
Just. de note	Tab.+Aud.	Tab.	165	0.6650
Just. de note	Tab.+Aud.	Voice	53	**0.0000**
Just. de note	Tab.	Voice	55	**0.0001**
Préc. de note	Tab.+Aud.	Tab.	163	0.6235
Préc. de note	Tab.+Aud.	Voix	30	**0.0000**
Préc. de note	Tab.	Voix	35	**0.0000**
Just. d'int.	Tab.+Aud.	Tab.	150	0.3854
Just. d'int.	Tab.+Aud.	Voix	245	0.0611
Just. d'int.	Tab.	Voix	266	**0.0118**
Préc. d'int.	Tab.+Aud.	Tab.	149	0.3697
Préc. d'int.	Tab.+Aud.	Voix	29	**0.0000**
Préc. d'int.	Tab.	Voix	29	**0.0000**

(a) *Ensembles des sujets*

Variable	Cond A	Cond B	U	p
Just. de note	Tab.+Aud.	Tab.	159	0.6339
Just. de note	Tab.+Aud.	Voix	85	**0.0410**
Just. de note	Tab.	Voix	86	**0.0447**
Préc. de note	Tab.+Aud.	Tab.	142	0.9458
Préc. de note	Tab.+Aud.	Voix	7	**0.0000**
Préc. de note	Tab.	Voix	14	**0.0000**
Just. d'int.	Tab.+Aud.	Tab.	103	0.1598
Just. d'int.	Tab.+Aud.	Voix	143	0.9729
Just. d'int.	Tab.	Voix	149	0.8919
Préc. d'int.	Tab.+Aud.	Tab.	133	0.7084
Préc. d'int.	Tab.+Aud.	Voix	4	**0.0000**
Préc. d'int.	Tab.	Voix	0	**0.0000**

(b) *Ensembles des stimulus*

Variable	Cond A	Cond B	U	p
Just. de note	Tab.+Aud.	Tab.	133	0.8672
Just. de note	Tab.+Aud.	Voix	88	0.1381
Just. de note	Tab.	Voix	88	0.1381
Préc. de note	Tab.+Aud.	Tab.	102	0.3414
Préc. de note	Tab.+Aud.	Voix	6	**0.0000**
Préc. de note	Tab.	Voix	6	**0.0000**
Just. d'int.	Tab.+Aud.	Tab.	105	0.4016
Just. d'int.	Tab.+Aud.	Voix	72	0.1697
Just. d'int.	Tab.	Voix	69	0.1318
Préc. d'int.	Tab.+Aud.	Tab.	121	0.8091
Préc. d'int.	Tab.+Aud.	Voix	2	**0.0000**
Préc. d'int.	Tab.	Voix	14	**0.0000**

(c) *Ensembles de distances de notes cibles précédentes*

Table 5.6 – *Test U de Mann-Whitney pour les différents ensembles étudiés*

5.5 Discussion et conclusions

5.5.1 Résumé des résultats

Dans cette étude nous avons d'abord comparé les justesse et précision d'un chant chironomique à celles du chant naturel. Le chant chironomique est réalisé à l'aide de l'instrument de voyelles chantées synthétiques *Cantor Digitalis*, pour lequel l'interface, une tablette graphique munie de repères musicaux, a été légèrement modifiée. La justesse et la précision de note et d'intervalle, deux mesures complémentaires pour l'étude de l'imitation de hauteurs musicales, ont été calculées sur plusieurs ensembles statistiques : les sujets, les stimulus, et les distances à la note cible précédente.

Nous avons montré que le jeu à la tablette permet d'obtenir des distributions de précisions de note et d'intervalle significativement inférieures de 60 à 170 cents par rapport à la voix naturelle, en considérant la médiane des distributions (figure 5.16). Les distributions des justesses de note des modalités aux tablettes sont quant à elles significativement inférieures de celles à la voix de 9 à 27 cents (médiane) pour 2 ensembles sur 3 (figure 5.15).

Il n'y a pas de différence significative entre les 2 modalités d'imitation utilisant la tablette, à savoir la tablette seule et la tablette avec retour audio. La totalité des médianes des distributions des justesses et précisions à la tablette seule ou avec retour audio sont inférieures à 20 cents (moins d'un huitième de ton), donc il est assez difficile de faire mieux, d'autant plus pour des sujets pour la grande majorité non habitués à cet exercice.

Les trois types d'ensembles sur lesquels sont calculés les justesses/précisions de note/intervalle (à savoir les ensembles de sujets, de stimulus et de distance à la note cible précédente) ne créent pas de différence marquée en ce qui concerne la justesse. Concernant la précision, les résultats sont meilleurs pour les ensembles de sujets en terme de médiane (72-82 cents contre 122-186 cents), mais avec une plus grande dispersion. Cela se traduit par quelques sujets mauvais qui se détachent des autres provoquant une grande dispersion mais une médiane de petite valeur. Sur les autres types d'ensembles de calcul, les sujets sont moyennés sur chaque mesure, donc l'effet des mauvais sujets se fait sentir plus sur la médiane que sur la dispersion.

Puis nous avons étudié l'influence du tempo imposé lors de l'imitation. Les distributions des précisions de note et d'intervalle entre un tempo à 240 battements par minute et un battement égal ou inférieur à 180 battements par minute sont en dessous de ce que l'oreille peut détecter. L'étendue des tempos ou la longueur des séquences de notes pourraient être augmentées de façon à rendre la tâche plus difficile et parvenir au point critique où les résultats chuteraient.

Cette étude démontre que le Cantor Digitalis et le Digitartic utilisent un paradigme de contrôle de l'intonation musicale très efficace comparé à la voix chantée naturelle, du moins chez des sujets non chanteurs professionnels. Cela nous pousse à persévérer dans le contrôle gestuel de l'intonation musicale à l'aide d'une tablette graphique munie d'un stylet.

5.5.2 Discussion

Le son utilisé pour synthétiser les hauteurs des stimulus est une voix de synthèse de mauvaise qualité utilisant le protocole de contrôle MIDI. Granot *et al.* [GIKGK13] ont montré que la qualité et le type de son utilisé comme stimulus ont une influence sur le résultat de l'imitation vocale. En particulier, la justesse est meilleure avec un stimulus de voix qu'avec un stimulus de voix de synthèse, et meilleure avec un stimulus de voix de synthèse qu'avec un son de piano, pour des raisons reliant audition et motricité lors de l'imitation. Dans notre cas, le stimulus est une voix de synthèse MIDI sans glissando entre chaque note, assez médiocre

et différente de celle utilisée en retour audio dans le cas de la modalité d'imitation *Tablette + Audio*. La modalité d'imitation *Voix* peut alors être défavorisée par rapport aux deux autres, amplifiant la tendance générale de nos résultats. Le résultat d'imitation par la voix naturelle pourrait être meilleure si le stimulus était une vraie voix.

Pour les expériences 1 et 2, peu d'ensembles par modalité sont à disposition (20, 17 ou 13), donc la signification statistique des résultats est possible seulement si les distributions sont clairement différentes. Pour l'expérience 3, beaucoup d'ensembles sont utilisés pour la distribution (336), donc la signification statistique est plus facilement atteignable si les distributions proviennent de processus différents. Ainsi, bien que nos résultats donnent des médianes et des dispersions trop proches les unes des autres pour ne pas être des différences négligeables (de l'ordre du cent), nous parvenons à des différences mathématiquement significatives concernant notre étude musicale. Dans les expériences 1 et 2, des distributions ne sont pas significativement différentes les unes des autres alors que leur allure pourrait le suggérer.

On remarque que la justesse de note de la modalité *Voix* est toujours un peu trop haute de 3 à 25 cents (médiane), c'est à dire moins de 1/8 de ton, ce qui n'apparaît pas pour les modalités d'imitation avec la tablette (voir figure 5.15). Après vérification, ce décalage n'a pas sa source dans un biais de l'extraction du F_0 stylisé. En effet, en envoyant le stimulus de voix de synthèse en entrée de l'algorithme utilisé pour l'expérience, ce dernier détecte bien la bonne justesse (en l'occurrence, à moins de 10 cents près autour de 0).

5.5.3 Travaux futurs

Il serait intéressant de refaire une expérience similaire avec des chanteurs professionnels, et de la comparer avec la présente étude, autant en ce qui concerne la modalité *Voix* que la modalité *Tablette Seule*. Les résultats avec la voix naturelle devraient être bien meilleure, la question restante étant : à quel point seraient-ils plus proches de ceux utilisant la modalité d'imitation *Tablette Seule* ? Les résultats de l'utilisation de la tablette par des musiciens experts pourraient aussi s'améliorer, mais dans une moindre mesure puisqu'ils ne sont pas experts de la tablette graphique. La différence entre la modalité *Tablette Seule* et *Tablette + Audio* serait alors intéressante dans le sens où, en tant que musiciens experts, leur attention porterait peut être plus sur le retour audio de la voix de synthèse qu'avec des sujets moins ou pas musiciens.

L'effet des marqueurs visuels semble prépondérant dans les meilleurs résultats de la tablette. Si on concevait une étude pour comparer la justesse/précision de sujets jouant d'un violoncelle standard et d'un violoncelle muni de repère mélodique sur son manche, il est certain que les sujets obtiendraient de meilleurs résultats avec le violoncelle muni de repère mélodique (sauf s'ils sont violoncellistes bien sûr). C'est un peu ce qui se passe dans notre étude où on compare la voix qui n'a pas de repère visuel. Qu'en serait-il si les marqueurs de la tablette étaient retirés ? L'utilisateur devrait alors se concentrer sur le retour audio, et la modalité *Tablette Seule* n'aurait plus de sens puisque aucun repère de notes auditif ou visuel ne serait disponible. Les modalités *Voix* et *Tablette + Audio* seraient alors plus comparables si l'on veut comparer des tâches faisant intervenir les mêmes sens, en l'occurrence seulement l'audition. On comparerait alors le geste interne des cordes vocales au geste manuel avec moins de biais que dans notre étude.

L'inverse serait aussi possible : munir la voix de repères visuels afin de pouvoir mieux la comparer au jeu à la tablette. En analysant la hauteur de la voix en temps-réel, on pourrait tracer un curseur évoluant sur un axe muni de repères de note comme c'est la cas avec le calque sur la tablette.

Chapitre 6

Étude préliminaire sur la justesse chironomique inter-musiciens

Sommaire

Après avoir étudié rigoureusement la justesse de musiciens jouant du *Cantor Digitalis* par des tâches individuelles, nous proposons d'étudier la justesse inter-musiciens dans un quatuor de *Cantor Digitalis* de l'ensemble *Chorus Digitalis*.

Nous introduirons notre étude par une discussion sur la difficulté de mesurer la justesse dans les chorales de voix naturelles, quand le groupe n'a pas de référence fixe donnée par un instrument à hauteur discrète comme le piano. Puis nous donnerons le protocole expérimental et ses limites. Enfin, les résultats seront discutés.

6.1 Justesse dans les chorales

Dans une chorale, un chanteur doit ajuster sa hauteur mélodique en se basant principalement sur deux signaux : sa propre voix (retour) et le son des autres voix (référence).

Deux facteurs identifiés par Ternström et Sundberg [TS88] peuvent influencer la justesse d'un chanteur. Tout d'abord le niveau sonore de la référence va empêcher le chanteur d'entendre suffisamment sa voix s'il est trop élevé. La différence entre ce niveau et celui de sa propre voix sera d'autant plus important que l'écart spatial entre les choristes sera élevé (moins de son direct des voisins) et que la réverbération de la salle sera faible (moins de son réfléchi en provenance des voisins). Deuxièmement, la justesse de F0 semble facilitée par le contenu spectral. Des harmoniques en commun avec les autres membres de la chorale, des harmoniques de fréquence élevée et l'absence de vibrato va permettre au chanteur d'accroître sa justesse. Les auteurs expliquent ce comportement par l'utilisation des battements entre leurs harmoniques respectives pour ajuster leur fréquence fondamentale, et par l'utilisation de vibrato brouillant cet effet. Dans un groupe de 6 chanteurs amateurs et pour des chants considérés comme bien réussis par la chorale, l'écart-type de la mesure de F0 est d'environ 10-15 cents avec des déviations pouvant aller occasionnellement à 45 cents.

Il a été remarqué que certains intervalles sont joués différemment, au moins par certaines chorales. Lottermoser et Meyer [LM60] (en allemand, cité en anglais dans [Ter03]) ont analysé la justesse de 3 chorales « réputées ». Ils ont montré que bien que les octaves et les quintes soient très près de la justesse théorique, les tierces majeures - 400 cents - étaient en moyenne plus grandes de 16 cents et que les tierces mineures - 300 cents - étaient en moyenne plus basse de 24 cents, en considérant la gamme comme tempérée. Ils interprètent cela comme un moyen d'accentuer les différences entre ces deux types d'intervalles de tierce.

Dès qu'il est question d'intervalles autres que l'unisson ou les octaves (voire la quinte), il faut définir dans quel type de gamme l'intervalle est considéré. On a d'un coté la gamme tempérée (l'octave est divisé en 12 intervalles égaux) et de l'autre coté les gammes dîtes *naturelles* (divisant l'octave suivant des fractions) comme la gamme Pythagoricienne ou la gamme de Zarlino. Dans la plupart des cas de nos jours, la gamme tempérée est utilisée. Les instruments à hauteurs discrètes sont construits pour un type de gamme donné, donc l'ambiguïté de type de gamme utilisé ne se présente pas pour ces instruments. On peut supposer qu'il en est de même pour les instruments à hauteur continue (dont la voix) qui seraient joués avec ces instruments, afin de s'accorder avec eux. En ce qui concerne une chorale jouant sans accompagnement, la question du type de tempérament utilisé se pose si on veut mesurer sa justesse en la comparant à une référence.

Nordmark et Ternström [NT96] ont demandé à 16 sujets présentant une expérience d'ensemble chorale d'ajuster manuellement des intervalles de tierces majeures formés par des sons synthétiques, en se basant sur leur préférence personnelle. Les résultats s'étalent de 388 à 407 cents (moyenne de 395 $\pm$ 7 cents), à comparer aux 400 cents de la gamme tempérée ou aux 386 cents de la gamme de Zarlino.

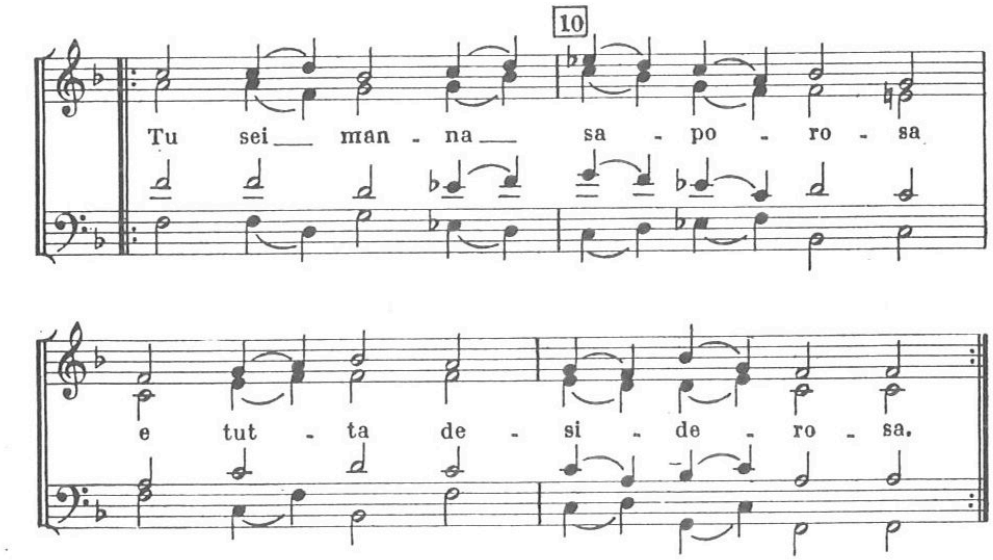


FIGURE 6.1 – *Partition de l'extrait du morceau Alta Trinita Beata*

Enfin, Ternström [Ter93] a étudié jusqu'à quel point un chanteur pouvait considérer qu'une section de 5 chanteurs à l'unisson (même fréquence fondamentale) était juste. Les résultats donnent une tolérance maximale d'écart-type de 14 cents, et une préférence d'écart-type de 0 ou 5 cents.

On considère que les musiciens du *Chorus Digitalis* jouent en gamme tempérée, notre instrument favorisant cette gamme donnée par les repères de son interface. D'autre part, ils ne sont pas considérés comme professionnels de l'instrument, mais avec une pratique récente ayant débuté en quartet en décembre 2010 et en sextet à partir de janvier 2011.

6.2 Protocole expérimental

On choisit deux types de références pour calculer la justesse d'une voix donnée de notre chorale :

- une référence fixée avant le début du morceau, correspondant aux notes cibles théoriques, et qui apparaissent sur les repères de l'interface ;
- une référence dépendant de l'ensemble de la chorale au temps i donné au cours du morceau, de la note cible du musicien k, et calculée comme il suit :

$$Ref_{i,k} = T_{i,k} + \frac{\sum_{k_0 \neq k}(S_{i,k_0} - T_{i,k_0})}{3} \tag{6.1}$$

où $T_{i,k}$ est la hauteur de la partition au temps i et pour la voix de synthèse k, et $S_{i,k}$ est la hauteur atteinte au temps i et pour la voix de synthèse k. La moyenne est réalisée sur les 3 autres voix.

Un extrait du morceau Alta Trinita Beata, morceau anonyme du XV$^{\text{ème}}$ siècle (extrait de partition donné à la figure 6.1) a été joué en quartet avec le Chorus Digitalis dont les sorties audios correspondant à chaque voix ont été enregistrées.

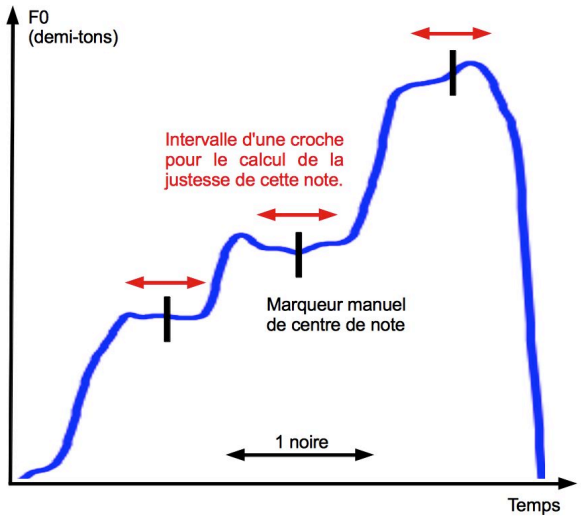


Figure 6.2 – *Méthode du calcul de la justesse*

L'enregistrement a été fait en mars 2011 avec comme interface pour chacun une tablette graphique permettant de contrôler avec le stylet, la hauteur musicale suivant la position suivant l'axe X, et la force vocale avec la pression. La voyelle était fixée à l'avance pour tout le morceau.

Pour pouvoir analyser le morceau, on procède d'abord de la façon suivante pour chacun des enregistrements correspondants aux 4 voix :

- la hauteur musicale est détectée sous Matlab grâce à l'algorithme YIN [CK02].
- le signal résultant est filtré pour retirer les variations trop rapides ;
- manuellement, on repère le centre temporel des notes atteintes. La longue durée de chacune d'elles (noire ou blanche, soit environ 0.6 ou 1.2 seconde chacune) permet assez facilement de le repérer, la hauteur musicale détectée formant des paliers assez nets. L'incertitude de mesure de ces centres temporels de notes est estimée à $\pm$ 0.05 seconde, soit $\pm$ 8% pour une noire.

Ensuite, on calcule chacune des notes atteintes, définie comme la moyenne de la hauteur musicale sur la durée d'une croche autour du centre de la note. On a donc au début et à la fin des paliers une durée d'au moins une double-croche où la justesse n'est pas calculée, permettant d'éviter les effets de bords (voir figure 6.2).

On calcule une note tous les temps pour chacune des voix, même si la note se poursuit en réalité sur 2 temps. On peut alors calculer la justesse pour chacune des voix k par rapport à deux références explicitées ci-dessus.

La justesse du musicien k, calculée à partir de la référence de la partition (soit des repères de l'interface), est la suivante :

$$J_{k,Ref=partition} = \frac{1}{N} \cdot \sum_i (S_{i,k} - T_{i,k}) \tag{6.2}$$

La justesse du musicien k, calculée à partir de la référence des autres voix donnée à l'équation 6.1, est la suivante :

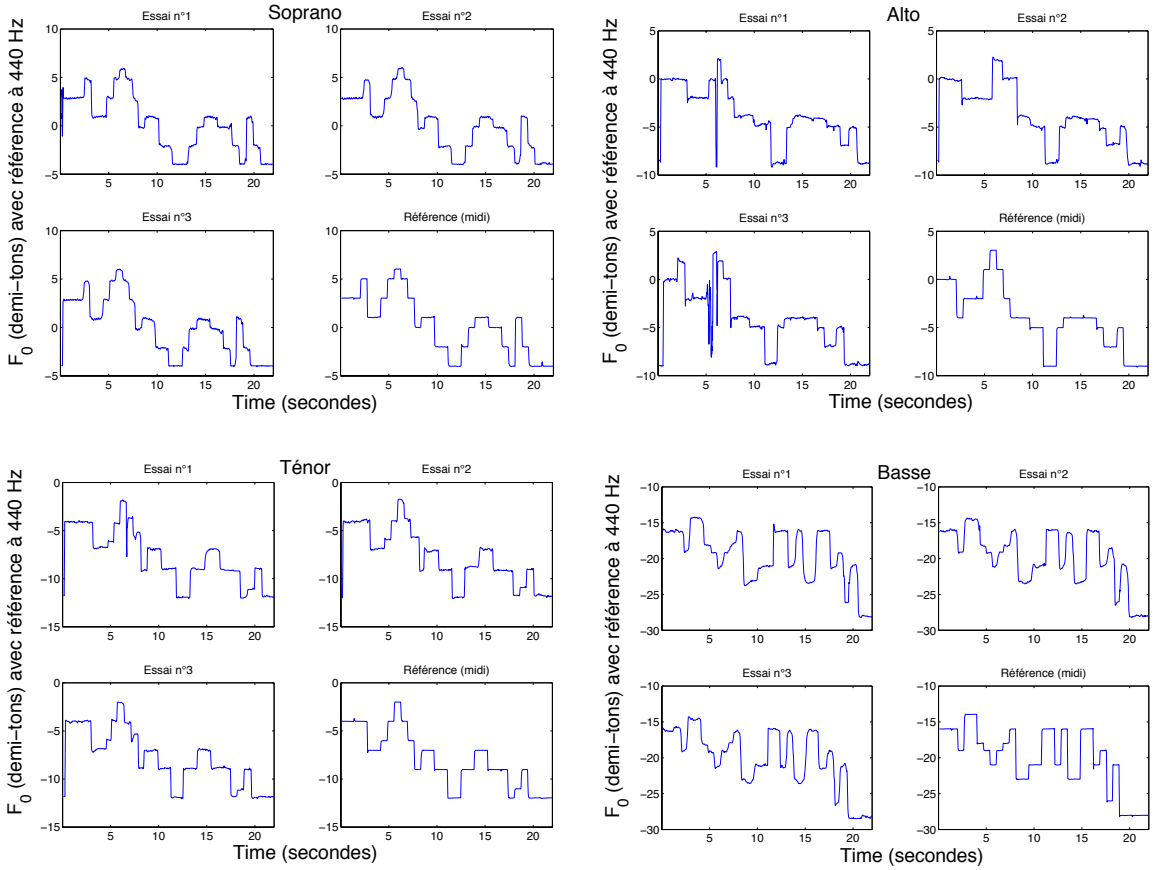


FIGURE 6.3 – *Mesure de F_0 des voix de soprano, alto, ténor et basse sur un extrait du morceau Alta Trinita Beata*

$$
\begin{aligned}
J_{k,Ref=ensemble} &= \frac{1}{N} \cdot \sum_i (S_{i,k} - Ref_{i,k}) \\
&= \frac{1}{N} \sum_i (S_{i,k} - T_{i,k} - \frac{1}{3} \cdot \sum_{k_0 \neq k} (S_{i,k_0} - T_{i,k_0}))
\end{aligned}
\tag{6.3}
$$

où $T_{i,k}$ est la hauteur de la partition au temps i et pour la voix de synthèse k, $S_{i,k}$ est la hauteur atteinte au temps i et pour la voix de synthèse k, et N est le nombre de temps à la noire de l'extrait (i.e. 32).

La même procédure a été réalisée pour les enregistrements issus de fichiers MIDI.

Trois essais ont été réalisés, dont la fréquence fondamentale est tracée pour chacune des voix avec sa référence MIDI à la figure 6.3. Seul l'essai 2 a été gardé, le mieux réussi. Les quatre voix de cet essai sont tracées à la figure 6.4, à comparer avec la référence MIDI de la figure 6.5. Les notes sont toutes atteintes, excepté pour l'alto qui en omet 3 et se trompe de note à 2 reprises.

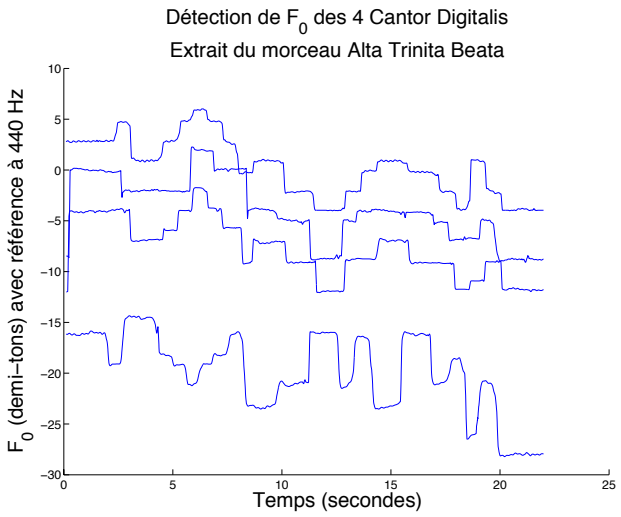


FIGURE 6.4 – *Mesure de F_0 des 4 voix du Chorus Digitalis, sur un extrait du morceau Alta Trinita Beata*

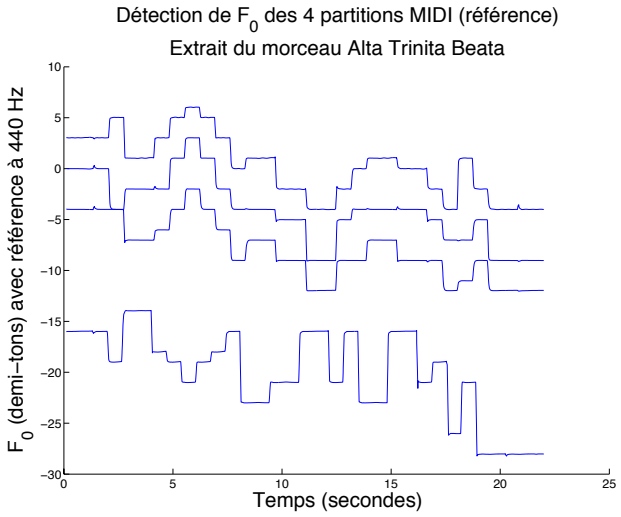


FIGURE 6.5 – *Mesure de F_0 des 4 voix de référence (MIDI), sur un extrait du morceau Alta Trinita Beata*

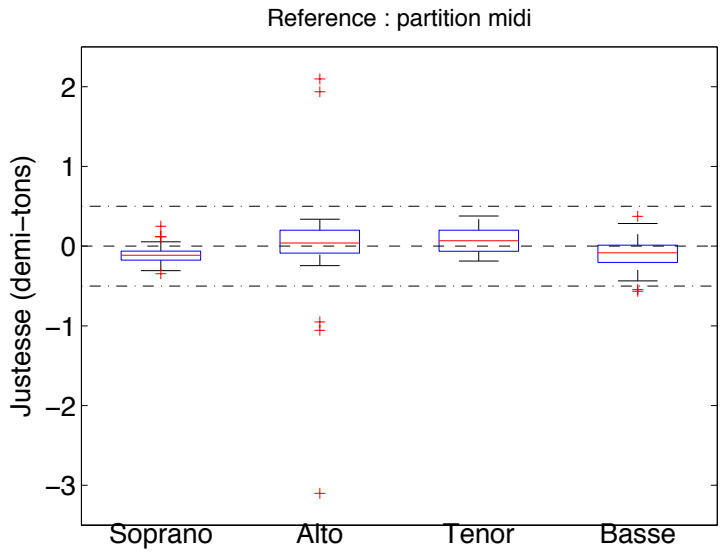


FIGURE 6.6 – *Justesse de chacun des musicien, avec pour référence la partition MIDI*

6.3 Résultats et discussions

Tout d'abord, il convient de noter les limites de notre méthode d'analyse. Le centre des notes est relevé à des instants temporels indépendamment des autres voix. Il en résulte plusieurs biais :

- les notes sont supposées être bien synchronisées temporellement entre les 4 musiciens ;
- les décalages temporels peuvent perturber la visée des hauteurs mélodiques de joueurs s'ils tentent de la corriger en fonction de ce qu'ils entendent des autres joueurs ;
- on minimise l'erreur de justesse en ne prenant en compte que la durée d'une croche au centre de la note, alors que cette dernière dure une noire (2 fois plus longtemps). Mais cela permet d'être plus sûr que cette note soit synchronisée avec les autres voix en enlevant les effets de bord dus aux phases transitoires où il est plus probable que tous les musiciens n'aient pas atteint leur note cible.

D'autre part, comme on l'a vu dans la partie 6.1, des chorales peuvent agrandir ou réduire certains intervalles pour des raisons esthétiques. Ainsi l'étude de la justesse comme critère de réussite doit être interprétée avec prudence.

On considère une note juste si la différence de sa hauteur et de celle de la référence est inférieure 0.5 demi-ton, étant donné que dans la musique considérée, l'intervalle minimal possible entre deux notes est de 1 demi-ton. Ainsi, on considère pour une note S et sa cible T :

- Si $|S - T| = \delta < 0.5$ demi-ton, alors on considère que le musicien a bien voulu viser la note T
- Si $|S - T| = \delta > 0.5$ demi-ton, alors il y a eu erreur dans la note T visée, i.e. la tâche demandée n'a pas été réalisée

En prenant pour référence la partition MIDI de la figure 6.5, on observe d'assez bons scores, avec une moyenne d'écart absolue aux notes cibles variant de 1 à 10 centièmes de demi-ton (4 à 11 centièmes pour la valeur absolue de la médiane). Ils sont à comparer avec l'intervalle minimum perceptible de hauteurs de sons purs, d'environ 5 cents. La distribution

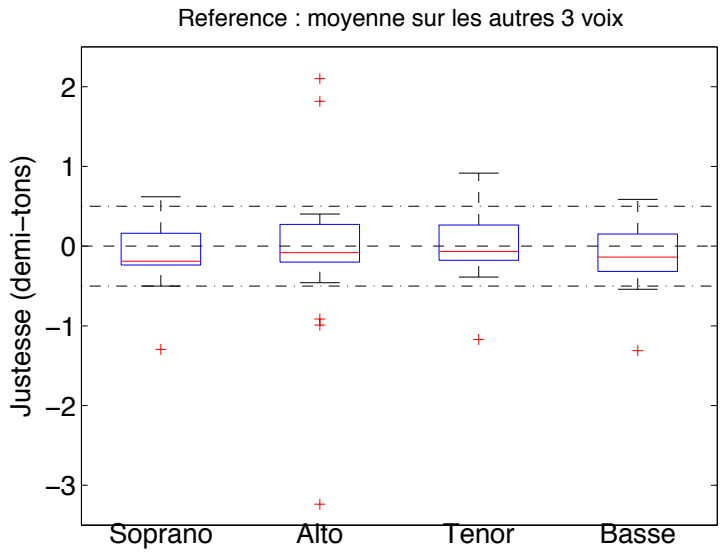


Figure 6.7 – *Justesse de chacun des musiciens, avec pour référence les autres musiciens, indépendamment de la partition*

des résultats en boite à moustaches avec pour référence la partition MIDI est présentée à la figure 6.6. 50% des notes sont comprises dans un intervalle de 11 à 29 centièmes de largeur selon les musiciens. On remarque des erreurs de notes pour l'alto déplacé de 2, -1 ou -3 demi-ton, comme on peut le vérifier aisément sur le tracé de la fréquence fondamentale.

En prenant comme référence non pas la partition MIDI mais les notes réalisées par les 4 musiciens, on obtient des résultats assez similaires mais avec une dispersion presque deux fois plus élevée (voir figure 6.7). Cependant, certaines estimations de la justesse se voient biaisées par une référence incohérente : c'est le cas des références des notes où l'alto s'est trompé ou a oublié de jouer une note. La figure 6.8 ne prend pas en compte les 5 notes de l'alto dont la justesse par rapport à la partition MIDI est supérieure à 0.5 centièmes de demi-ton, afin d'avoir des références plus plausibles.

Le tableau 6.1 donne les résultats sous la forme des valeurs de la moyenne, de la médiane et de l'écart-type.

	Soprano	**Alto**	**Ténor**	**Basse**
Référence : partition MIDI				
Moyenne	-10	1	7	-10
Médiane	-12	4	7	-8
Écart-type	13	81	16	21
Référence : les 3 autres voix (moins les notes aberrantes)				
Moyenne	-9	3	1	-10
Médiane	-18	8	6	-13
Écart-type	22	84	23	27

Table 6.1 – *Récapitulatif des statistiques de la justesse inter-musicien (en cents)*

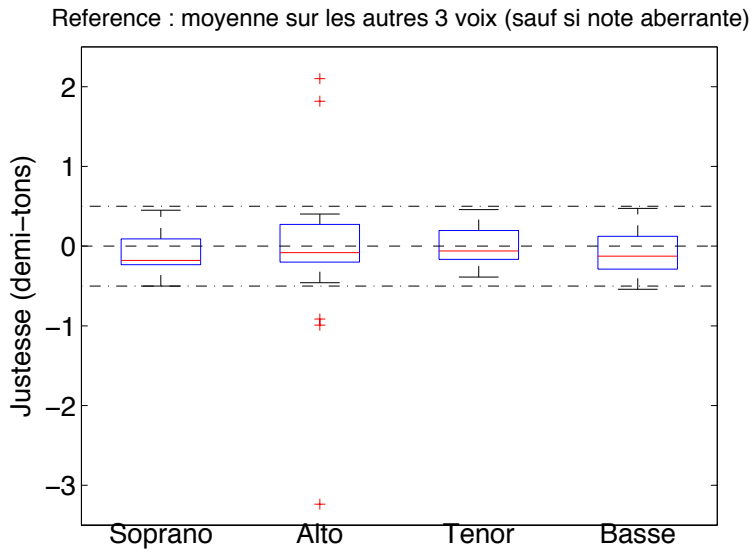


FIGURE 6.8 – *Justesse de chacun des musiciens, avec pour référence les autres musiciens, indépendamment de la partition*

Avec l'interface du Cantor Digitalis et concernant la hauteur mélodique, le musicien est attiré par deux références : celle fixe indiquée par les repères de hauteur de notes sur la tablette et celle mouvante formée par le son des autres joueurs à un instant. Le poids respectif de ces deux références dépend en grande partie de deux facteurs. Premièrement, l'expertise dans l'utilisation de la tablette permettra de viser plus justement les hauteurs de notes indiquées sur la tablette. L'effet sera l'augmentation du poids de la référence de la partition. Deuxièmement, l'expertise musicale permettra d'entendre une déviation de son propre son par rapport aux autres et de le rectifier. Le poids de la référence constituée des autres musiciens sera alors renforcé.

C'est le cas chez d'autres instruments plus traditionnels mais avec quelques différences :

- les instruments à cordes sans frette. Ici, les repères autres qu'auditifs ne sont pas visuels mais plutôt liés à la mémoire des positions de la main.
- le jeu au ruban des ondes Martenot. Hormis les repères auditifs, l'instrumentiste a un retour tactile (creux et bosses) pour viser les notes avec l'index.
- la voix où le repère est lié à la mémoire de la tension de muscles pour faire vibrer les plis vocaux aux fréquences désirées.

Le cas le plus proche de notre instrument serait celui des guitares basses sans frette auxquelles on ajoute souvent des repères visuels peints sur le manche représentant les frettes (donc les demi-tons).

6.4 Conclusion

La justesse des musiciens dans l'ensemble a été mesurée suivant une référence fixe (la partition / les repères musicaux de la tablette) et une référence mouvante construite par l'ensemble des musiciens. Les résultats montrent la capacité des musiciens à jouer juste en groupe, tout en laissant supposer qu'ils se basent plus sur les repères de leur instrument respectif plutôt qu'en s'écoutant les uns les autres. Une étude plus complète devrait être

entreprise pour en avoir la certitude.

Le contexte musical doit être aussi pris en compte. Quand les notes s'enchaînent rapidement, les joueurs n'ont surement pas le temps d'ajuster leur visée en fonction de ce qu'ils entendent. Par contre, sur des notes tenues, il est probable que les joueurs s'écoutent plus entre eux dans ces conditions.

Enfin, notons que même si l'écoute intervient sans doute peu dans la justesse inter-musiciens à travers les résultats de ce protocole expérimental, elle pèse fortement sur la synchronisation temporelle inter-musiciens. Mais si les joueurs s'écoutaient moins, ils seraient moins synchronisés temporellement et par conséquence, leur décalage temporelle ne pourrait qu'amener à une situation confuse harmoniquement. Mais notre protocole expérimental ne prend pas en compte ce décalage en considérant par hypothèse les notes inter-musiciens bien calées temporellement, amenant à un biais potentiellement important.

Conclusion générale et perspectives

Contributions de la thèse

Développement d'instruments de synthèse vocale

Nous avons présenté deux instruments de voix chantée, le *Cantor Digitalis* et le *Digitartic*. Ils se présentent sous la forme de deux modes de jeu d'une même application écrite en Max [Max]. Les interfaces utilisées sont deux tablettes graphiques dont on a ajouté des repères de hauteurs de notes et d'articulation.

Basée sur un modèle source-filtre, la source glottique est simulée par le modèle de débit glottique RT-CALM [ddLBD06], et les résonances du conduit vocal sont reproduites en utilisant la synthèse par formant. Des règles améliorent la qualité du modèle, comme l'ajout d'un seuil de phonation, des perturbations automatiques de la source glottique à différentes échelles temporelles, ainsi que des dépendances entre fréquence fondamentale, position et amplitude des résonances, et effort vocal.

Les voix sont personnalisables en termes de longueur de conduit vocal, tension et taux de souffle vocaux, tessiture, et apériodicités de la source. Une diversité de timbres est alors disponible, accessible par des réglages continus de paramètres ou sous forme de pré-configurations. Les voix les plus réussies sont : 4 types de voix classiques européennes, à savoir basse, ténor, alto et soprano ; une voix d'enfant et de bébé ; une voix animale proche du rugissement de fauve. La voix de ténor a été choisie comme voix de référence pour le réglage des consonnes.

Pour contrôler la hauteur musicale, la position d'un stylet sur une tablette graphique est utilisée. Cette interface répond à nos besoins dans la mesure où elle allie haute résolution spatiale et fréquentielle, et permet de réutiliser notre faculté à écrire pour produire des gestes très précis. L'autre contrôle sur la source glottique est l'effort vocal. Relié à plusieurs paramètres temporels et spectraux du modèle de source glottique, on l'associe à la pression du stylet sur la tablette, représentant elle aussi un effort.

Le conduit vocal est contrôlable différemment selon le *Cantor Digitalis* et le *Digitartic*, mais pour tous les deux avec la main secondaire dans la plupart des configurations. On résume ici les modèles de contrôle les plus efficaces. Pour changer la couleur vocalique avec le *Cantor Digitalis*, on utilise les capteurs tactiles de la même tablette que celle contrôlant la source glottique. Ainsi, la position de l'index de la main secondaire permet de se déplacer dans un espace 2D correspondant aux voyelles orales du français. Avec le *Digitartic*, une deuxième tablette graphique est utilisée par la main secondaire permettant de contrôler continûment la phase et le lieu d'articulation parmi quatre modes d'articulation. La couleur vocalique est gérée par la dimension Y de la tablette de la main préférée s'occupant de la source glottique.

Les deux instruments sont construits de telle sorte qu'il soit possible de contrôler précisément la hauteur musicale et l'articulation de consonnes et de voyelles à l'aide de gestes manuels. La mélodie, la couleur vocalique, le lieu et la phase d'articulation peuvent être contrôlés continûment en temps réel sans latence perceptible. On peut alors jouer des syl-

labes articulées expressives et réactives dans un contexte musical de voix chantée, de scat, ou de récitation d'onomatopées.

Contrôle chironomique de l'intonation musicale

Nous avons étudié et comparé la faculté de participants à imiter des séquences de 2 à 7 hauteurs de notes avec leur propre voix et avec le *Cantor Digitalis*, à différents tempos d'imitation. L'interface utilisée pour contrôler la hauteur musicale est une tablette graphique proche de celle utilisée par le *Cantor Digitalis* et le *Digitartic*, à la différence de la plus grande distance spatiale entre chaque note. Une tâche témoin de l'imitation avec la tablette sans le retour audio du synthétiseur a également été réalisée.

La justesse et la précision, deux mesures complémentaires pour l'étude de l'imitation de hauteurs musicales, ont été calculées sur plusieurs ensembles statistiques : les sujets, les stimulus, et les distances à la note cible précédente.

Le jeu à la tablette présente des justesses proches de zéro (i.e. réalisations centrées autour de la cible) et des précisions inférieures à 25-50 cents de demi-ton (médiane). Les résultats sur la voix présentent quant à eux un écart de justesse d'environ 20-30 cents (médiane), et moins précis de 30-75 cents (médiane), selon les ensembles de calcul. La tablette est significativement meilleure que la voix en ce qui concerne la précision de note et d'intervalle. Pour la justesse, les résultats sont moins évidents, avec une non signification statistique pour la justesse d'intervalle. La justesse de note est significativement meilleure pour un jeu à la tablette qu'à la voix, sauf pour l'ensemble des distances à la note cible.

La différence (statistiquement significative) entre la modalité d'imitation à la tablette avec et sans retour audio est de l'ordre de quelques cents, mais elle est en-dessous du seuil de perception d'environ 5-10 cents. L'audio n'a donc pas d'influence dans cette tâche. La raison principale est avant tout que les résultats de la tablette seule ne pourraient être améliorés puisqu'ils atteignent des scores proches de la justesse et précision parfaites.

Concernant le choix des ensembles statistiques pour le calcul des justesses et précisions, il y a peu de différence pour les deux modalités utilisant la tablette. Avec la voix, les différences sont de 20 à 40 cents, avec la plus grande différence entre les ensembles de sujets et ceux de distances à la note cible.

Cette étude démontre l'efficacité du paradigme de contrôle de l'intonation musicale du *Cantor Digitalis* et du *Digitartic* comparé à la voix chantée naturelle, du moins chez des sujets non chanteurs experts, d'autant plus que le retour audio n'est pas nécessaire pour aboutir à de très bons résultats.

Gestes instrumentaux et jeux collectifs

Nos deux instruments, surtout le *Cantor Digitalis*, ont démontré aussi leur aptitude à être joués dans des situations musicales en dehors des conditions de laboratoire. Avec cinq concerts à son actif et deux prévus prochainement, l'ensemble *Chorus Digitalis* composé de 1 à 6 voix de synthèse et d'autant de musiciens, a élaboré un répertoire varié. Des morceaux traditionnels ont permis de mettre en évidence que notre chorale de voix de synthèse est capable de reproduire des pièces vocales similaires à des chorales ou à des chants individuels de voix naturelle, comme la polyphonie baroque ou le chant Khayal d'Inde du Nord (annexe D).

Les gestes de nos instruments nécessaires à la production vocale de synthèse ont été décrits suivant la topologie de Cadoz [Cad88]. Les gestes nécessaires à la réalisation de certaines tâches musicales comme le portamento, le vibrato ou encore les attaques ont été décrits et

analysés à partir des enregistrements de répétition du *Chorus Digitalis*. Des tendances à développer des styles de jeu personnels ont été identifiées parmi les usagers.

Une analyse de la justesse inter-musicien dans l'ensemble a été entreprise en mesurant la justesse de leurs notes suivant une référence fixe (la partition / les repères musicaux de la tablette) et une référence mouvante construite par l'ensemble des musiciens. Les résultats montrent la capacité des musiciens à jouer juste en groupe, tout en laissant supposer qu'ils se basent plus sur les repères de leur instrument respectifs que sur l'écoute mutuelle.

Application pédagogique

Enfin, une application à but pédagogique a été construite à partir du *Cantor Digitalis* (annexe B). Tous les modules de l'instrument sont présentés au chapitre 2. L'application permet par exemple d'écouter un formant isolé tout en modifiant le paramètre de couleur vocalique, afin d'entendre l'effet de la modification du conduit vocal sur ce formant en particulier. Ainsi, on a à disposition une sorte d'atelier de luthier numérique qui permet de construire brique par brique l'appareil vocal, d'écouter l'effet de chacune d'elle, puis d'affiner la voix et le type de contrôle pour en faire un chanteur doté d'une expressivité contrôlée par l'utilisateur. Cette application a été utilisée à la fête de la science, lors d'une conférence scientifique grand public et en cours de licence à l'UPMC.

Perspectives

Vers de nouveaux instruments de voix numériques

Pour les contraintes musicales spécifiques qu'on s'est fixées dans cette thèse, et indépendamment des problèmes de puissance de calcul, la méthode de synthèse idéale est le modèle physique. C'est en effet celle qui se rapproche le plus de la réalité, car l'onde acoustique rayonnée et calculée résulte comme dans le cas naturel de sa propagation dans le conduit vocal et de son interaction avec les articulateurs. Avec l'augmentation des puissances de calculs des ordinateurs, on peut espérer un jour pouvoir piloter un modèle physique complexe en temps-réel avec le même type de modèle de contrôle que celui proposé dans cette thèse. En fonctionnant par interpolation des positions articulatoires correspondant à des cibles phonétiques, on pourrait agir sur des paramètres de haut niveau comme la phase d'articulation ou le lieu d'articulation pour contrôler gestuellement un tel modèle. Théoriquement, tous les sons possibles issus de la voix naturelle pourraient être produits. Cependant, les autres méthodes de synthèse ont de beaux jours devant elles en attendant ce jour prochain où la puissance de calcul des ordinateurs sera suffisante. D'autre part, si on se dirige vers de la parole, ce type de modèle de contrôle sera insuffisant vue la vitesse élevée des changements articulatoires et les combinaisons multiples des phonèmes possibles.

D'autres méthodes de synthèse pourraient être employées, en essayant de garder à la fois réactivité et qualité. Les modèles possédant la meilleure qualité de synthèse aujourd'hui sont ceux utilisant des bases de données [KO07], mais leur organisation intrinsèque les empêchent d'interagir continûment sur le phonème en cours de synthèse. Mais cela peut être suffisant selon les applications [SBVB06] [Bel11]. De plus, pour pouvoir travailler en temps réel, les segments synthétisés doivent présenter un nombre moindre de segments passés en dépendance, impliquant alors une diminution de qualité [AdP+12]. Au contraire, les modèles utilisant la synthèse par formant ou les fonctions d'ondes formantiques peuvent réagir sans latence en temps réel, mais leur qualité est moindre et la synthèse de parole est plus difficile. Il s'agit de faire des compromis en fonction des exigences des applications envisagées. En écoutant

la qualité des synthétiseurs par formants en temps différé (voir exemples sonores donnés dans [Sun06]), il est possible d'améliorer la qualité du *Digitartic* en agissant d'abord sur le modèle de production des syllabes avant de modifier éventuellement le modèle de contrôle.

Comparaison de tâches articulatoires vocales et chironomiques

L'intonation a été étudiée en terme de comparaison entre voix naturelle et geste chiromique, à la fois dans le cadre de la voix parlée avec la prosodie [dRLB11] que dans celui de la musique (voir chapitre 5). Nous pourrions poursuivre la comparaison des gestes chironomiques et de la voix par l'étude de tâche d'imitation dans le domaine temporel de l'articulation à l'aide du *Digitartic*.

En contrôlant continûment la phase d'articulation, on peut imaginer plusieurs expériences. La première serait d'articuler des séquences de type "VCVCVCV..." en synchronisant l'attaque musicale des syllabes avec un rythme donné par le stimulus. La tâche consisterait à contrôler avec sa propre voix ou avec le stylet d'une tablette le long d'un axe 1D la phase articulatoire de manière à synchroniser la phase CV avec une grille temporelle. On comparerait alors la justesse et précision rythmique par rapport au métronome. La seconde expérience pourrait demander d'imiter le degré d'articulation de séquence VCV avec sa voix et avec le *Digitartic*, toujours suivant la même dimension de la tablette, où le geste consisterait en un mouvement de différentes amplitudes suivant le degré d'articulation à reproduire, comme expliqué dans la partie 3.3.6. Par contre, l'analyse du degré d'articulation de la voix naturelle peut être difficile à identifier et quantifier, surtout à cause de la diversité articulatoire des participants.

Si la première expérience proposée ci-dessus donne de bons résultats pour la chironomie, nous pouvons alors combiner celle-ci à celle réalisée dans le chapitre 5 sur l'imitation de hauteurs de notes, en demandant de synchroniser temporellement une séquence de type "VCVCVC..." tout en jouant une mélodie dont les notes changent à chaque syllabe. On évaluerait alors à la fois la capacité à jouer en rythme des syllabes et celle à jouer juste et précis.

Étude du Chorus Digitalis

L'ensemble Chorus Digitalis est un bon moyen pour étudier l'évolution des pratiques et des stratégies de jeu des musiciens. Il serait intéressant d'introduire de vrais chanteurs dans l'ensemble et d'examiner jusqu'à quel point les chanteurs de voix naturelle et les chanteurs chironomiques peuvent interagir. L'interaction peut se dérouler entre deux personnes différentes ou avec une même personne qui chanterait et jouerait de la voix synthétique de manière superposée ou en alternance. Le musicien disposerait alors de deux voix a priori indépendantes mais de nature différente, que ce soit par rapport au modèle de production ou au modèle de contrôle. On peut aussi imaginer qu'un chanteur contrôle un instrument vocal à l'aide de sa propre voix par reconnaissance automatique. La correspondance pourra être directe (e.g. mêmes hauteurs mélodiques, voyelles et efforts vocaux entre voix naturelle et synthétique) ou indirecte en réfléchissant à un mapping avec un sens musical intéressant.

Le Chorus Digitalis a deux buts, un musical et un scientifique, qui se nourrissent l'un de l'autre. L'aspect musical permet de démontrer la viabilité instrumentale de nos choix de contrôle et de méthode de synthèse, tandis que le côté scientifique permet la réalisation technique de concepts originaux sur l'usage de la voix chantée, donnant au Chorus Digiatis sa force du point de vue artistique.

ANNEXES

Annexe A

Le modèle de source RT-CALM : aspects mathématiques

Le modèle de source RT-CALM se compose de deux filtres en cascades, l'un correspondant à la position et l'amplitude du « formant glottique », l'autre à la pente spectrale de l'ODGD dans les hautes fréquences. Sur la figure A.1, on représente le « formant glottique » par sa fréquence centrale F_c et son amplitude A_c, et le début de la pente spectrale par sa position fréquentielle F_c et son amplitude A_c [DdH03].

Le filtre correspondant au « formant glottique » est un filtre avec deux pôles anti-causaux et un pôle causal qui a pour fonction de transfert :

$$H(z) = \frac{b_1 z}{1 + a_1 z + a_2 z^2} \tag{A.1}$$

où, si $T_e = \frac{1}{F_e}$ est la période d'échantillonnage :

$$\begin{aligned} a_1 &= -2e^{-a_p T_e} \cos(b_p T_e) \\ a_2 &= e - 2a_p T_e \\ b_1 &= E \frac{\pi^2}{b_p^3} e^{-a_p T_e} \sin(b_p T_e) \end{aligned} \tag{A.2}$$

et a_p et $\pm b_p$ sont la partie réelle et imaginaire des deux pôles anti-causaux p du filtre (avec $T_0 = F_0$) :

$$\begin{aligned} p &= a_p \pm j b_p \\ a_p &= -\frac{\pi}{O_q T_0 \tan(\pi \alpha_m)} \\ b_p &= \frac{\pi}{O_q T_0} \end{aligned} \tag{A.3}$$

Le paramètre O_q est le quotient ouvert de l'ODGD, α_m représente son coefficient d'asymétrie, et $-E$ son minimum. La figure A.2, représentant la forme temporelle de l'ODGD sur une période fondamentale T_0, donne la définition graphique de O_q, E et α_m.

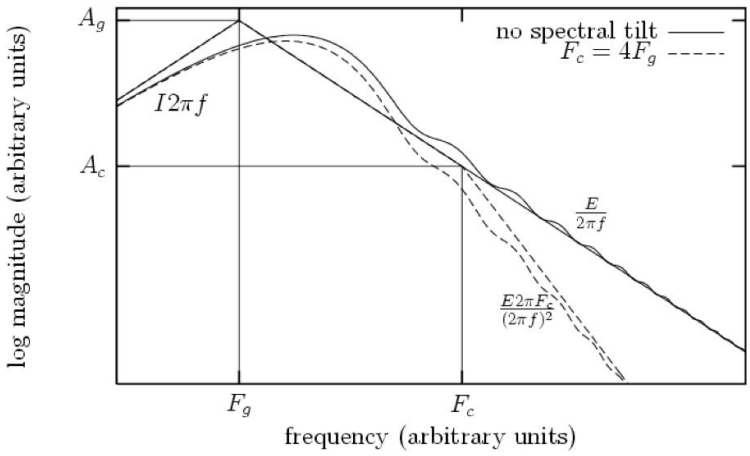


FIGURE A.1 – *Spectre de l'ODGD avec formant glottique (Fg, Ag) pour deux pentes spectrales (Fc, Ac), d'après Doval et al. [DdH03]*

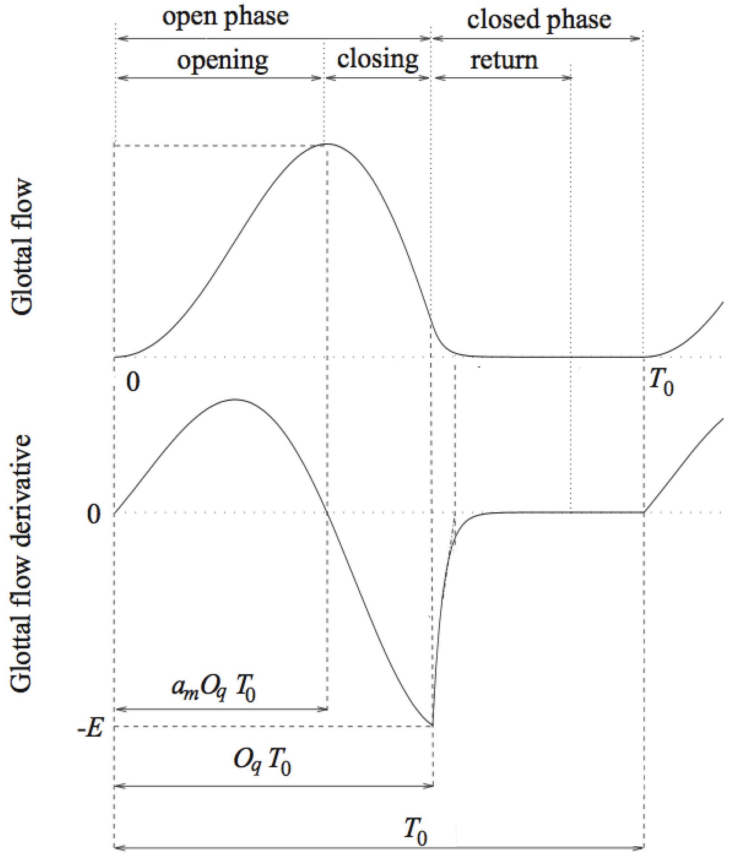


FIGURE A.2 – *L'ODG et l'ODGD, et ses paramètres E, O_q et α_m sur une période fondamentale T_0, d'après Doval et al. [DdH06]*

Le filtre correspondant à la pente spectrale de l'ODGD dans les hautes fréquences est un filtre à un pôle, de fonction de transfert :

$$H(z) = \frac{b_{TL}}{1 - a_{TL} z^{-1}} \tag{A.4}$$

a_{TL} et b_{TL} sont calculés en fonction de la pente spectrale désirée TL (amplitude du spectre à 3000 Hz en dB) par :

$$\begin{aligned} a_{TL} &= \nu - \sqrt{\nu^2 - 1} \\ b_{TL} &= 1 - a_{TL} \\ \nu &= 1 - \frac{1}{\eta} \\ \eta &= \frac{10^{TL/10} - 1}{\cos(2\pi \frac{3000}{F_e}) - 1} \end{aligned} \tag{A.5}$$

Annexe B

Application ludo-éducative en phonétique acoustique et voix chantée

B.1 Introduction

Comme avec toute connaissance à transmettre, le processus d'apprentissage est plus aisé et amusant en utilisant des supports d'enseignement variés. De plus, si l'étudiant peut interagir avec le support en temps réel, le succès est presque garanti. Le fonctionnement de la voix peut être une de ces connaissances à enseigner.

Malgré un usage quotidien, le fonctionnement de la voix reste obscur pour le grand public car la majeure partie des organes impliqués est cachée de l'extérieur et la production de la voix met en jeu des concepts physiques abstraits difficiles à appréhender. Séparer un système en sous-parties peut aider à sa compréhension, mais une contrainte s'impose : notre objet d'étude est une partie du corps humain et il serait dangereux de la modifier plus largement que ce qu'on peut faire naturellement.

Nous utilisons alors le modèle de production de voix du *Cantor Digitalis*. Son approche de type signal, conjuguée une correspondance judicieuse entre les paramètres du modèle de production et les paramètres de contrôle permettent la modification de paramètres de haut-niveau ayant un sens physique. La synthèse par formant et le modèle source-filtre nous laisse la possibilité de *déconstruire* le modèle de production de voix pour écouter des phénomènes acoustiques abstraits tels un formant isolé ou encore le bruit isolé formé lors de la vibration des plis vocaux.

Dans cette partie, une application interactive et temps-réel implémentée en Max 6 [Max] est présentée, basée sur l'instrument *Cantor Digitalis*.

Quelques systèmes de synthèse vocale sont disponibles pour des applications pédagogiques, tels que le Vocal Tract Lab [Bir] ou le Benoit Project [Fuc]. Ils ne permettent souvent pas d'écouter des phénomènes acoustiques abstraits cités ci-dessus à cause de la méthode de synthèse utilisée, basée sur des modèles physiques de l'appareil vocal, et donc dépendant de son fonctionnement global. Mais la plus grande originalité de notre application est avant tout sa capacité à pouvoir jouer avec le modèle en temps réel et donc d'écouter activement la transformation du conduit vocal et de la source glottique. Le but initial du *Cantor Digitalis* est la musique, donc nous pouvons l'utiliser facilement dans un contexte ludo-éducatif en utilisant son interface de contrôle pour modifier les paramètres du modèle.

Nous traiterons d'abord de la « décomposition » du modèle source-filtre pour une voix

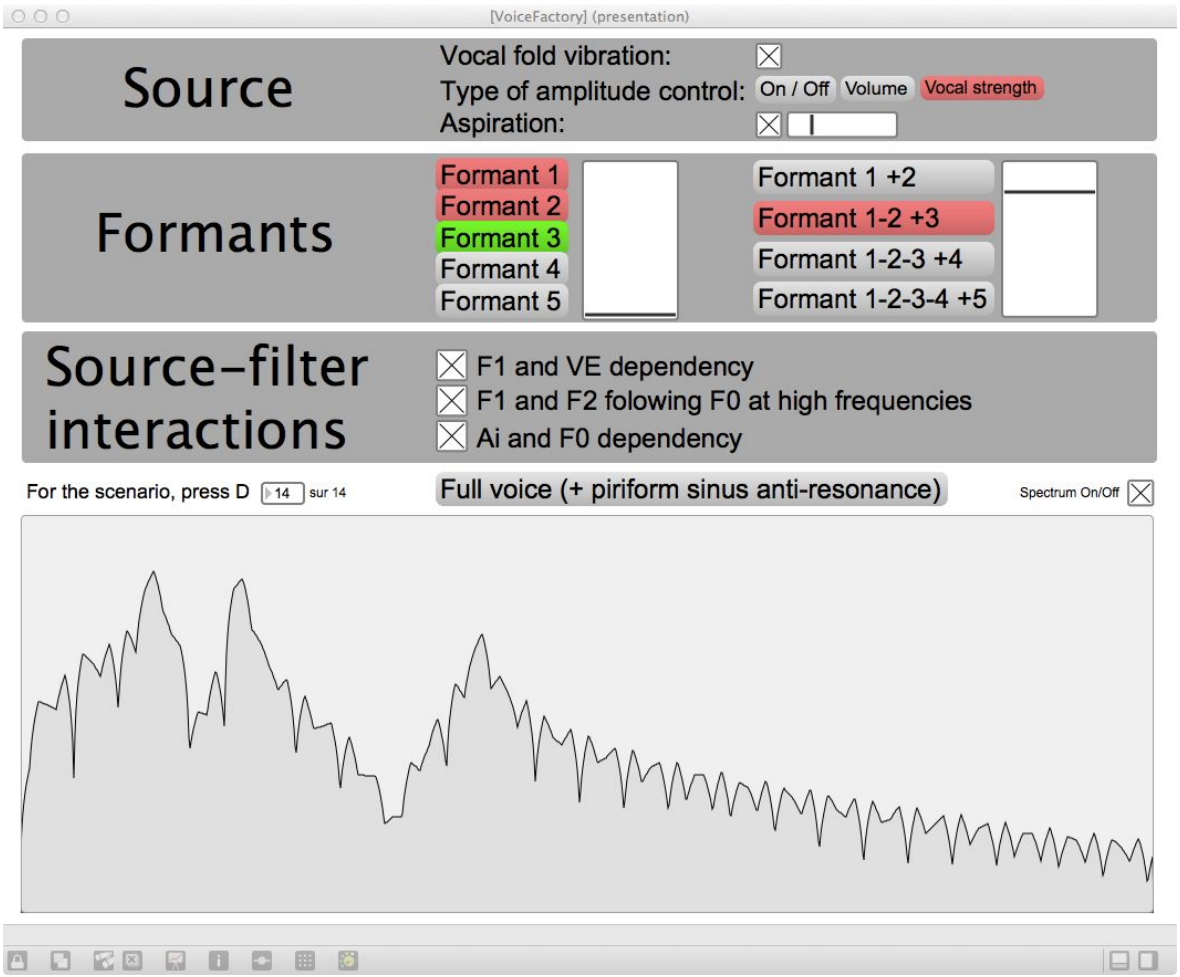


FIGURE B.1 – *Capture d'écran de l'interface 1 (décomposition du modèle source-filtre)*

donnée et expliquerons comment l'utiliser dans une perspective pédagogique. Ensuite, nous étudierons sa transformation vers un autre profil vocal à travers un processus continu et réactif.

B.2 Enseignement par le jeu : le modèle source-filtre de la production vocale

La démarche utilisée ici est à l'opposé de celle employée lors de la modélisation de la production vocale : on part d'un modèle établi et on le déconstruit pour en écouter ses éléments ou des combinaisons de ces éléments, tout en utilisant la même interface gestuelle que celle du modèle fini.

A cela s'ajoute une interface logicielle (capture d'écran reproduite à la figure B.1), séparée en une zone cliquable pour choisir les modules à enlever ou ajouter et en une zone de visualisation montrant le spectre temps-réel de la sortie audio du modèle recomposé. Ci-dessous sont détaillés les différents usages du logiciel pour l'enseignement.

Jouer avec la source glottique

Le son harmonique de la source est un effet de la vibration des plis vocaux. Cependant, ce son n'est jamais entendu séparément du conduit vocal, car le son de la voix est la convolution de la source avec le conduit vocal. De plus, il est impossible de retirer le conduit vocal. Ce n'est pas le cas d'autres instruments, comme par exemple avec le saxophone, où l'embouchure peut être séparée du reste de l'instrument.

Avec un modèle de production de la voix utilisant la théorie source-filtre, il est possible d'écouter la source isolément comme si le conduit vocal avait été retiré. Ainsi, en utilisant la même interface qu'avec le *Cantor Digitalis*, le joueur peut écouter comment la source glottique réagit. Le contrôle de la source glottique est limité à la hauteur mélodique et à l'effort vocal. Donc changer la configuration du conduit vocal (voyelle) n'affecte pas le son en sortie.

L'interface logicielle permet de choisir la correspondance entre la pression du stylet de la tablette et l'intensité vocale, et ainsi de tester et écouter différentes correspondances : de type booléen en fonction du contact du stylet avec la tablette ; de type « bouton de volume », en contrôlant linéairement l'amplitude du signal ; et de type « effort vocal », plus naturel, qui agit sur l'amplitude et la pente spectrale dans les hautes fréquences.

Enfin, le bruit de la source, dû aux turbulences de l'air autour des plis vocaux, peut être manuellement ajouté au son harmonique de la source, ou être écouté indépendamment. Cet effet n'est pas réalisable avec une vraie voix, ou approximativement sur une voix enregistrée après traitement informatique.

Écouter les mouvements des formants suivant les trajectoires articulatoires

Le mot « formant » est commun en sciences du langage, mais il peut être difficile à comprendre, surtout pour les étudiants non familiers avec le traitement du signal ou la physique. De plus, un formant pris isolément n'existe pas dans la voix, l'ensemble des formants résultant de la forme globale du conduit vocal. Il est alors difficile d'identifier chacun d'entre d'eux dans le son d'une voix naturelle.

A partir du son de synthèse de la source glottique et de curseurs de l'interface, on peut écouter l'émergence d'un formant individuel, c'est à dire son effet de filtrage sur la source glottique.

Tout en écoutant un des 5 formants (numéroté de celui ayant la plus petite à celui ayant la plus haute fréquence centrale), le conduit vocal peut être modifié en se déplaçant dans l'espace des voyelles et ainsi la contribution de ce formant à la voyelle choisie peut être identifiée. Le contrôle temps-réel de la fréquence centrale du formant permet d'identifier grossièrement les mouvements articulatoires : ouverture de la mâchoire au formant 1, position de la langue au formant 2, ouverture des lèvres au formant 3.

Combiner les formants pour faire émerger l'identification de la voyelle

Après avoir écouté chaque formant indépendamment des autres, on peut s'intéresser aux effets audios de l'ajout de chacun des formants un par un, de celui ayant la plus petite fréquence centrale à celui ayant la plus grande. La bande passante de chacun des filtres formantiques est continument contrôlée, grâce à un curseur, pour le faire intervenir petit à petit jusqu'à la bande passante associée à la voyelle.

L'ajout du $2^{\text{ème}}$ formant au 1^{er} permet d'identifier presque toutes les voyelles. Additionner le $3^{\text{ème}}$ formant supprime des ambiguïtés entre /i/ et /u/. Enfin, le $4^{\text{ème}}$ et le $5^{\text{ème}}$ formant améliorent le naturel de la voix, mais sans changer l'identification de la voyelle.

Écouter l'émergence de la perception de la voix humaine ou de voyelles tout en observant les changements du spectre de la sortie audio, permet de comprendre intuitivement le lien entre perception sonore et spectre à travers le concept de résonances/formants.

La contribution harmonique de la source glottique peut être retirée, permettant d'écouter les effets des formants sur le bruit de l'écoulement turbulent de la source, en se déplaçant dans l'espace vocalique à l'aide de l'interface.

Synchroniser les mouvements de la source glottique avec ceux du conduit vocal

Ce logiciel a déjà été utilisé plusieurs fois avec le grand public (Journée Science et Musique à Rennes, fête de la science à Orsay) et en classe avec des étudiants scientifiques (UPMC).

En plus des applications possibles présentées plus haut, il a aussi été utilisé pour illustrer la tâche de coordination entre le source glottique et le conduit vocal pour produire des mots avec une expression donnée. Une personne de l'audience devait contrôler la source glottique (hauteur mélodique et effort vocal), une autre le conduit vocal (voyelles). Il leur a été demandé d'imiter de courts mots expressifs tels *Oh yeah* ou *Oui-oui*. La tâche se faisait en 4 étapes : tout d'abord, un mot de voix naturelle était prononcé ; ensuite les 2 personnes devaient analyser l'évolution de la hauteur mélodique et des voyelles du mot prononcé ; puis chacun devait s'entrainer à reproduire le geste approprié, la hauteur mélodique pour l'un, la couleur vocalique pour l'autre ; enfin les deux joueurs devaient synchroniser leurs gestes afin de reproduire le mot expressif prononcé au début. Cet exercice ludo-éducatif permet de comprendre en partie comment sa propre voix est produite en contrôlant une voix de synthèse. L'expérience a produit de l'enthousiasme et les utilisateurs ont rapidement compris comment améliorer leur premier essai et à généraliser la méthode pour créer d'autres mots courts avec des expressions différentes.

Les interactions source-filtre, décrites dans la section 2.4, ne sont pas jouables, mais peuvent être activées ou désactivées afin d'apprécier leur effet sur le son final.

B.3 Enseignement par le jeu : réglages des voix et individualisation

Nous avons décrit la manière de construire et déconstruire une voix donnée pour enseigner le fonctionnement de la production vocale. Maintenant, nous traitons de la façon d'utiliser l'individualisation des voix comme moyen d'enseignement par le jeu.

Pour cela, nous utilisons une interface graphique de l'instrument *Cantor Digitalis*, destinée à l'individualisation des voix. L'interface graphique, dont la capture d'écran est donnée à la figure B.2 est divisée en différentes parties de pré-configurations et de réglages manuels. La partie supérieure concerne les types de voix et la partie inférieure les valeurs des formants des voyelles. Tous les réglages s'appliquent immédiatement au son de voix de synthèse.

Ajustement des formants

Une base de données de formants contenant 6 voyelles est disponible. Mais par interpolation de ces voyelles, des voyelles intermédiaires et une infinité de couleurs vocaliques peuvent être obtenues, contrôlées gestuellement par la position du doigt ou d'un stylet dans le plan à 2-dimensions de la tablette.

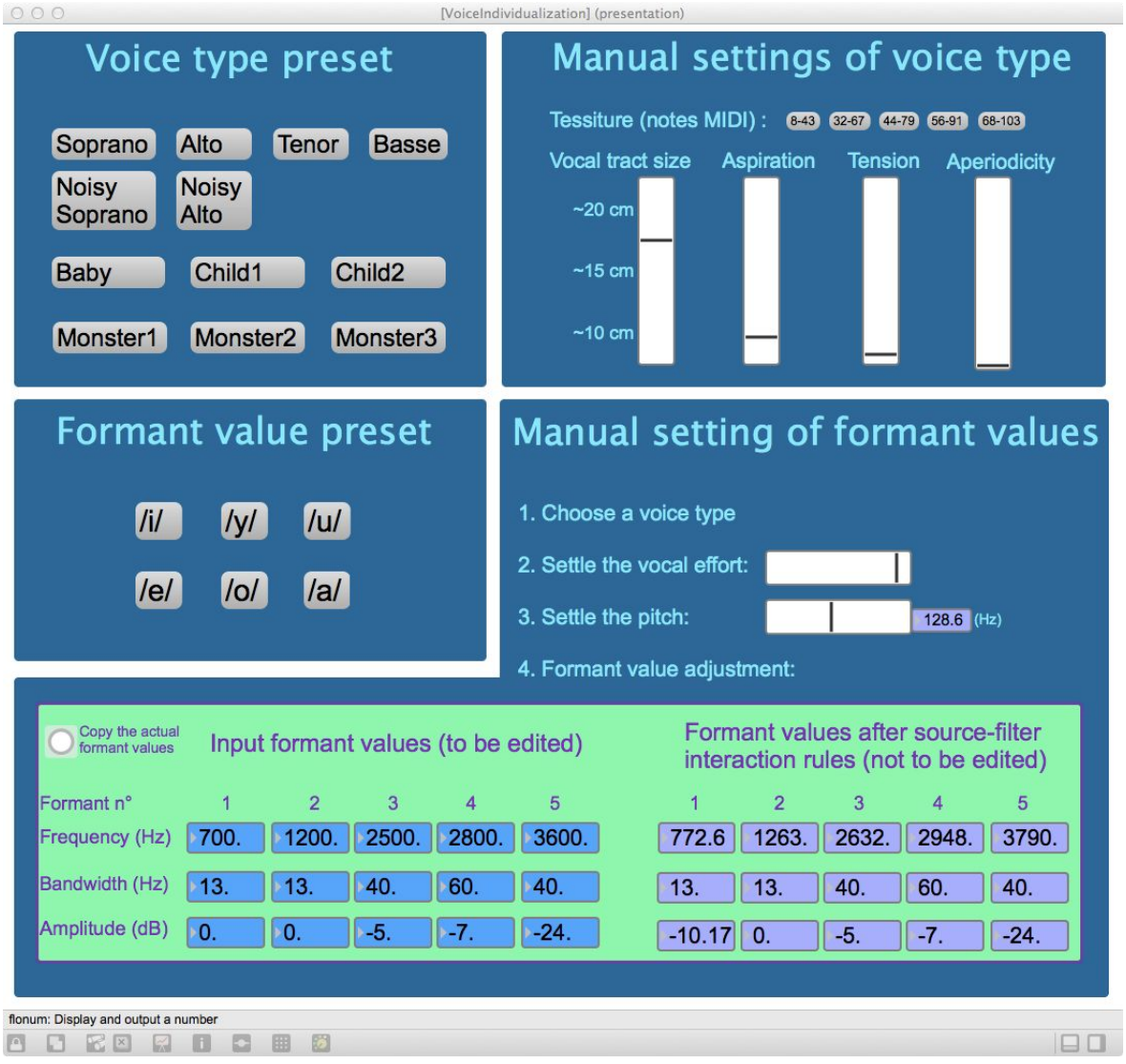


Figure B.2 – *Capture d'écran de l'interface graphique pour l'individualisation des voix*

Chacun des filtres passe-bandes, représentant un formant, peut être manuellement ajusté par sa fréquence centrale, sa bande-passante et son amplitude. Ainsi, les 3 paramètres de contrôle de chacun des 5 filtres formantiques peuvent être réglés indépendamment, en bas à gauche de l'interface graphique. Les valeurs effectives, après les interactions source-filtres, sont affichées en bas à droite. La modification est faite en temps-réel mais il n'est pas possible pour l'utilisateur de jouer ces nouvelles voyelles avec la tablette (la hauteur et l'effort vocal sont donnés par un curseur graphique). Ce module peut être utilisé pour démontrer que différentes couleurs vocaliques (i.e. différentes configurations vocaliques) peuvent être perçues comme une même voyelle, un effet illustrant la perception catégorielle des voyelles.

Taille du conduit vocal

Plus le conduit vocal est petit, plus grandes sont les fréquences de résonance, et donc plus grandes sont les fréquences centrales des formants du conduit vocal. Ainsi, nous pouvons changer la taille apparente du conduit vocal de notre voix synthétique par un facteur commun aux fréquences centrales des filtres formantiques. On peut alors passer en temps-réel d'une voix de bébé (petit conduit vocal) à une voix d'adulte (grand conduit vocal). L'effet est encore plus efficace si la tessiture décroit d'une manière consistante avec l'augmentation de la taille du conduit vocal.

Qualité de voix

Plusieurs paramètres de qualité de voix sont disponibles et opèrent sur le modèle de source glottique, comme le taux de bruit d'aspiration, la tension de voix (inversement proportionnelle à la fréquence du maximum du spectre de l'ODGD), ou l'apériodicité de la vibration de la source (jitter/shimmer).

Au-delà de la voix humaine

En modifiant l'ensemble de ces paramètres en temps-réel, il est facile, par essais et erreurs, de donner une individualité aux voix synthétiques. On peut alors illustrer les similarités entre appareil vocal humain et celui de certains animaux, en modifiant la taille du conduit vocal et la tessiture, le bruit d'aspiration et la tension vocale. Il est important de modifier les paramètres les uns après les autres tout en écoutant et en indiquant quel paramètre est concerné. Un exemple emblématique de cette application réside dans la transformation d'une voix humaine en un rugissement de fauve, et ce en multipliant par 2 la taille du conduit vocal, en baissant de 3 ou 4 octaves la tessiture, et en augmentant le bruit d'aspiration et de la tension vocale.

B.4 Conclusion

Nous avons présenté une application ludo-pédagogique dérivée de l'instrument *Cantor Digitalis* et destinée à l'enseignement de la phonétique acoustique. Elle permet d'utiliser un contrôle gestuel pour interagir avec la construction temps-réel et l'individualisation du modèle de voix. Deux principales directions peuvent être suivies :

1. Comment la production vocale fonctionne en se basant sur un modèle source-filtre ?
2. Quelles sont les différences entre les voix humaines (voir avec des voix animales) ?

Des phénomènes abstraits comme les formants, l'identification des voyelles, la décomposition de la voix en une partie *source* et une partie *filtre* peuvent être comprises par le son et une interaction gestuelle.

Cette application pédagogique a déjà été utilisée en différents contextes comme des festivals scientifiques pour le grand public ou pendant des cours à l'université.

B.5 Perspectives

Étant tout d'abord destiné aux enseignants de la production vocale et dans le but d'aider un maximum d'enseignants et d'étudiants, la prochaine étape serait de rendre ce logiciel disponible gratuitement ou avec ses sources de codes informatiques.

Des nouveautés pourraient être ajoutées au présent prototype, selon les besoins exprimés par ses utilisateurs potentiels.

Enfin, ce type d'instrument de synthèse est idéal pour la création musicale et la construction rapide de nouvelles voix (humaines ou monstrueuses), grâce à la réponse temps-réel du système à n'importe quelle modification de ses paramètres et à son interface de contrôle idéale pour « faire vivre » la voix.

Une caractéristique puissante et spécifique au contrôle gestuel de la synthèse est sa capacité à jouer avec la dynamique vocale, ce qui est souvent la clé pour obtenir un naturel dans la synthèse vocale.

Il conviendrait enfin de l'adapter à l'instrument de synthèse de syllabes *Digitartic*. Afin de mettre en évidence la production des consonnes, on pourrait par exemple visualiser les trajectoires formantiques dans le temps. Le jeu à plusieurs du *Chorus Digitalis* pourrait être quant à lui visualisé en terme de trajectoires de F_0.

Annexe C

Interface logicielle

L'instrument est programmé dans la langage *Max 6*. Il est constitué de deux applications dialoguant entre elles. La première, dont une capture d'écran est donnée à la figure C.1, récupère les données tactiles et des stylets des tablettes. Les données de la tablette issues du stylet sont collectées à l'aide de l'objet *s2m.wacom* et celles issues de la position des doigts à l'aide de l'objet *s2m.wacomtouch*[1]. Les données sont ensuite envoyées par le protocole UDP, précédées par un message permettant de les distinguer les unes des autres (pression, position X, position Y, stylet/tactile). La réception des données de la tablette réalisée indépendamment du modèle de production et exportée en application, permet de réduire la latence d'acquisition des données.

Les différentes options de l'application concernent l'activation de la réception de chaque tablette, le choix du mode d'interaction (stylet ou tactile), et des solutions envisageables si de la latence est présente : limiter le nombre de paramètres envoyés en désactivant ceux non utilisés ; limiter le débit de données en sous-échantillonnant l'envoi des données.

L'application comporte également un enregistreur de gestes décrit dans la partie 4.3.

La deuxième application reçoit les messages envoyés via le protocole UDP par l'application de réception des données de la tablette, et s'en sert pour contrôler l'instrument (figure C.2). Il permet de :

- choisir sa voix parmi un ensemble de pré-configuration ou de l'individualiser. En double-cliquant sur le cadre « VoiceIndividualization », une fenêtre s'ouvre (figure B.2) permettant de régler chacun des paramètres décrits dans la partie 2.5 ;
- choisir les modes de contrôle de l'espace vocalique (mono- ou bi-manuel) ou d'activer le mode pour avoir accès aux consonnes, explicité dans le chapitre 3 ;
- d'accorder la voix en éditant finement sur quelle note commence et finit la tessiture ;
- d'activer/désactiver le seuil de phonation ;
- de régler la réverbération en sortie d'instrument ;
- d'accéder à un module pédagogique « VoiceFactory » sur la voix présenté dans l'annexe B.

La deuxième application correspond au moteur de synthèse, basé sur la théorie source-filtre et la synthèse par formant. La figure C.3 est une capture d'écran du patch Max/MSP principal de l'application. On peut y voir sur la partie en bas à gauche les sous-patchs en

1. Disponibles sur la page http://metason.cnrs-mrs.fr/Resultats/MaxMSP/index.html (laboratoire LMA), lien consulté le 06/12/2012

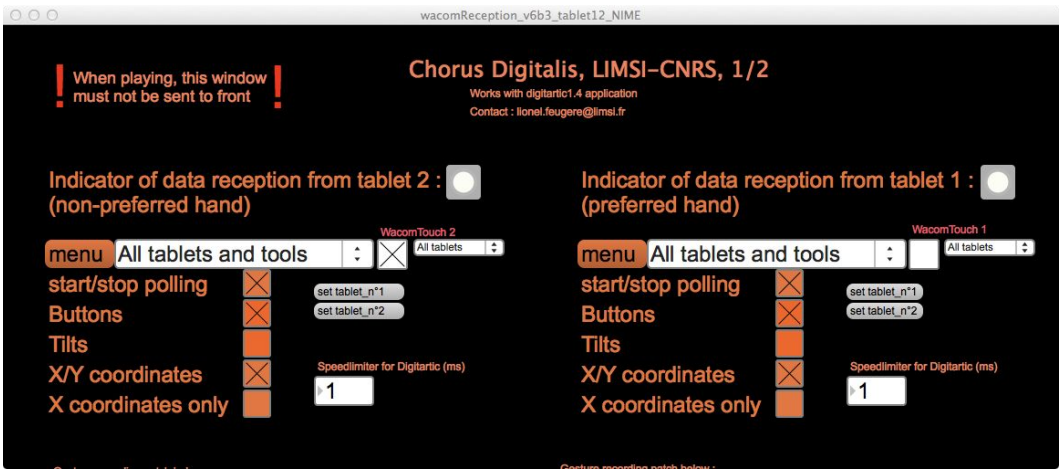


FIGURE C.1 – *Capture d'écran de l'interface utilisateur Max/MSP du Cantor Digitalis*

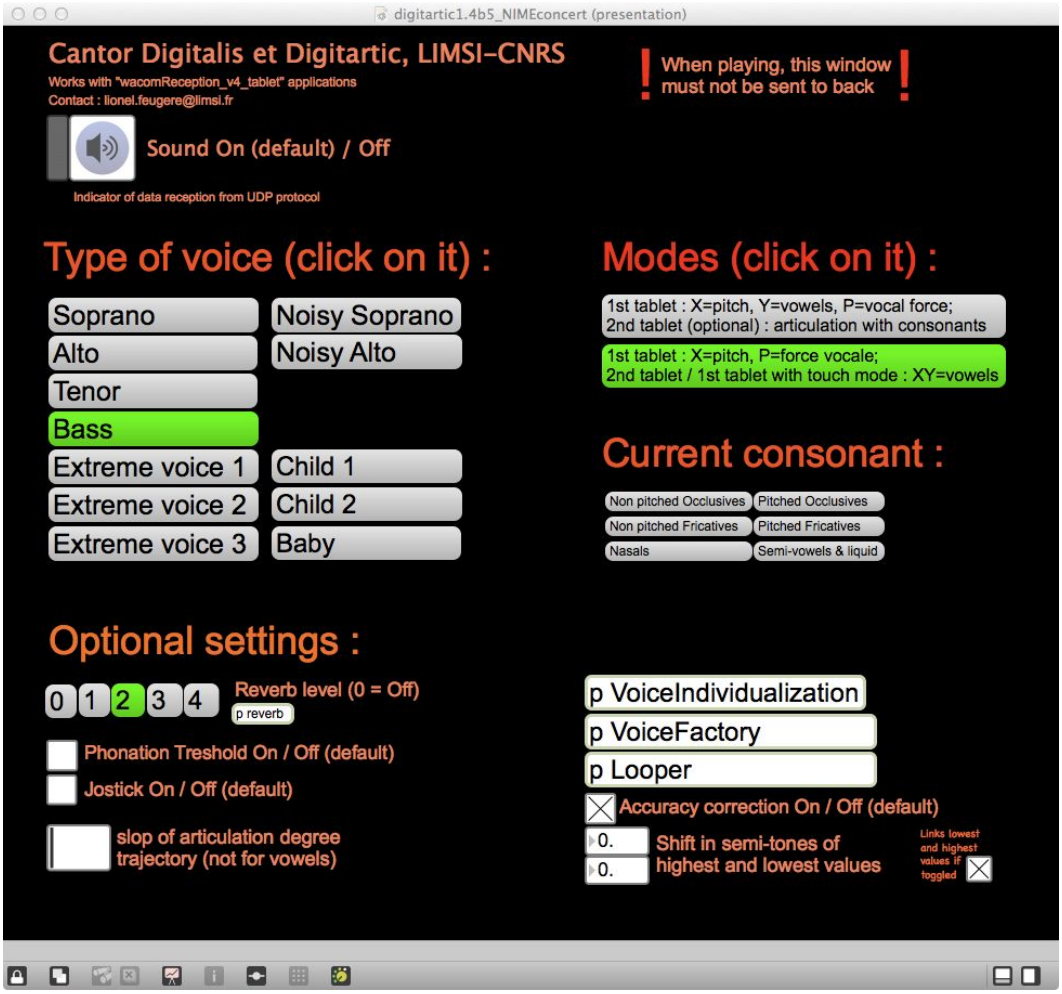


FIGURE C.2 – *Capture d'écran de l'interface utilisateur Max/MSP du Cantor Digitalis*

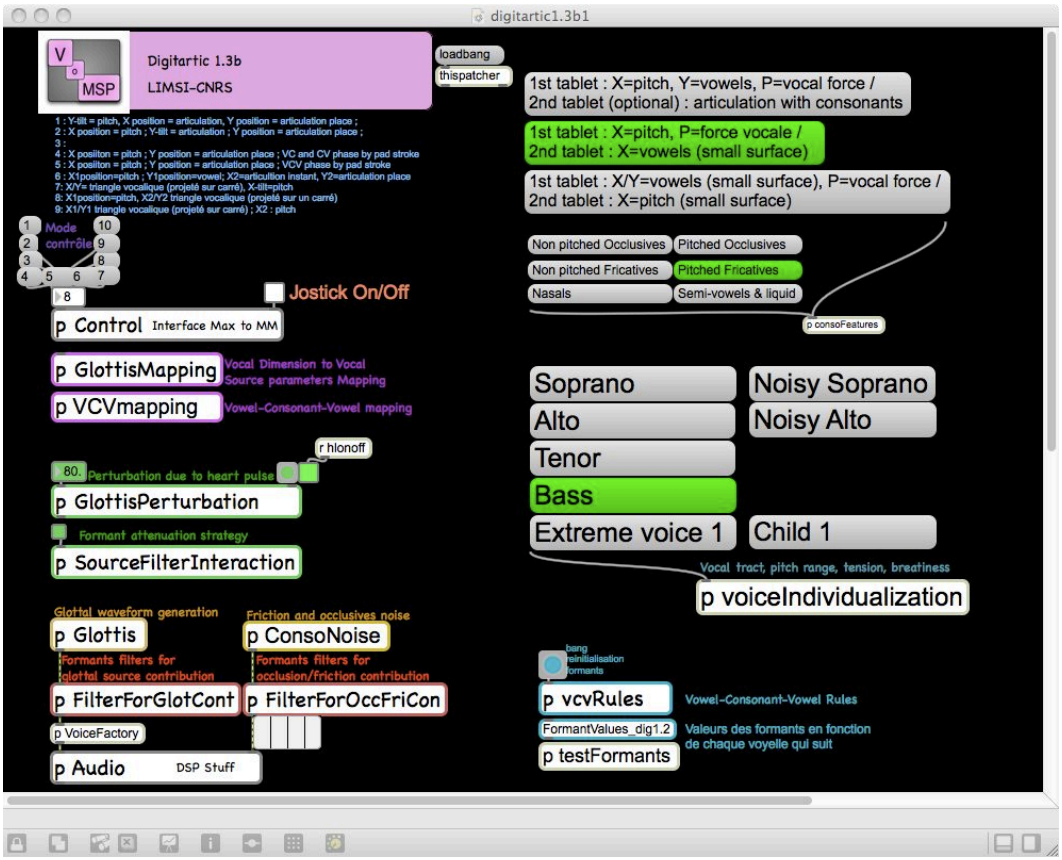


FIGURE C.3 – *Capture d'écran de la fenêtre d'édition des types de voix*

jaune et bleu correspondant à la structure source-filtre : *Glottis* représentant la source harmonique, *ConsoNoise* représentant la source de bruits consonantiques, et les filtres associés *FilterForGlotCont* et *FilterForOccFriCon* correspondant à la modélisation du conduit vocal. Les autres patchs correspondent au mapping entre l'interface et les paramètres de haut-niveau du modèle de contrôle et du mapping interne au modèle de production. En particulier, les règles d'articulation sont implémentées dans le sous-patch *vcvRules*. Pour connaitre la façon dont est implémenté le moteur de synthèse, le lecteur est invité à se reporter aux chapitres correspondant, à savoir le chapitre 2 pour la synthèse de voyelles et le chapitre 3 pour les syllabes.

Annexe D

L'ensemble musical *Chorus Digitalis*

L'utilisation d'instruments numériques issus de recherches dans un contexte musical n'est pas systématique. Tout d'abord, elle nécessite l'achèvement de l'instrument. Il existe des différences entre un instrument utilisé uniquement pour des expériences scientifiques et un instrument utilisé également à des fins musicales. En effet, le développeur et le musicien n'étant pas forcément les mêmes personnes, l'interface doit être relativement simple pour être utilisée par une personne extérieure au développement de l'instrument. De plus, jouer en concert impose une stabilité du programme informatique, un bogue en cours de jeu n'étant pas envisageable. De la même manière, la commercialisation de l'instrument nécessiterait une étape supplémentaire dans l'achèvement de l'instrument en terme de facilité d'utilisation et de documentation. Kessous [Kes02] parle de « prise en main » de l'instrument numérique et considère que le plus important est la « capacité de l'instrument à se révéler par lui-même ». En d'autres termes, un utilisateur quelconque doit rapidement prendre conscience des principes de fonctionnement de l'instrument, même s'il demande une pratique longue pour les maîtriser.

L'expérimentation musicale des instruments développés fait aussi intervenir d'autres compétences qui ne sont pas forcément celles du chercheur : une pratique musicale, l'organisation de répétitions et de concerts. Ici, les chercheurs impliqués dans le projet ont choisi de participer également en tant que musiciens, accompagnés d'autres musiciens extérieurs au projet de recherche.

L'utilisation musicale effective de l'instrument créé est indispensable pour pouvoir mesurer sa « jouabilité ». C'est du moins le critère qui nous a semblé le plus pertinent pour son évaluation. Le concert représente au mieux le résultat de cette pratique individuelle et/ou collective.

Nous traitons dans cette présente annexe de notre chorale, créée fin 2010 et appelée *Chorus Digitalis*[1]. Nous nous attachons dans un premier temps à décrire ses caractéristiques logicielles et logistiques. Nous exposerons ensuite son répertoire musical.

D.1 Description générale

Le projet *Chorus Digitalis* découle d'une idée originale de Christophe d'Alessandro, directeur de cette thèse. Les premières répétitions ont eu lieu fin 2010 en quatuor avec la

1. http://groupeaa.limsi.fr/projets:orjo:chorus_digitalis

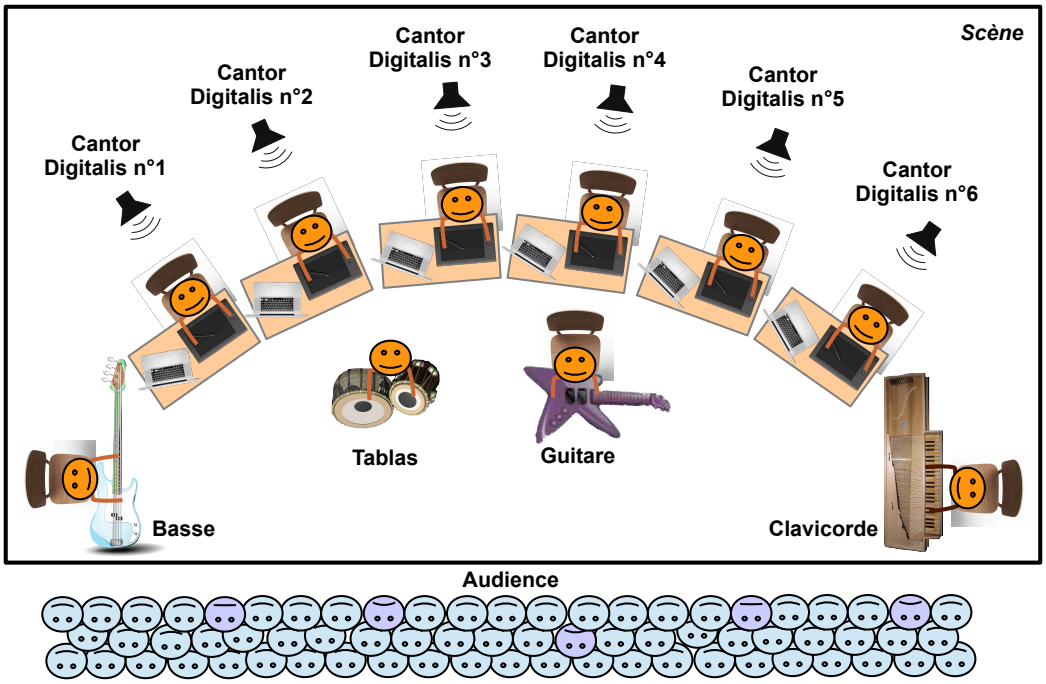


FIGURE D.1 – *Plan de scène du Chorus Digitalis dans sa configuration du concert au Printemps de la Culture 2012*

participation des autres concepteurs de cette chorale, à savoir Sylvain Le Beux et Boris Doval. Nous avons répété assez irrégulièrement jusqu'à notre premier concert en mars 2011 d'une durée de quinze minutes [FLBdD11]. Nous avons par la suite recruté 7 personnes extérieures au projet de recherche, afin de préparer un concert de plus grande envergure dans le cadre des journées Art Science du Printemps de la Culture 2012. Des répétitions régulières ont alors été entreprises dès décembre 2011. Notre plan de scène pour ce concert était un arc de cercle pour les joueurs de voix de synthèse, comme indiqué sur la figure D.1. A ces instruments de voix ont été ajoutés d'autres instruments, suivant la pratique instrumentale des musiciens présents. Ainsi, selon la configuration des morceaux, un clavecin, des tablas, une basse et une guitare électrique ont été ajoutés à l'ensemble.

Chacune des voix de l'ensemble *Chorus Digitalis* fait intervenir les éléments suivants :

- Un musicien, impliqué ou non dans le développement de l'instrument. Dans notre ensemble, environ la moitié de l'effectif est intégrée au projet de recherche, l'autre est constituée de personnes seulement intéressées par la partie musicale.
- Une interface physique, composée d'une à deux tablettes graphiques, comme décrite dans les sections 2.7 et 3.3, contrôlée gestuellement par le musicien.
- Un ordinateur disposé à proximité du musicien qui récupère les données de l'interface physique pour calculer la voix de synthèse en temps réel. Une interface logicielle est également disponible pour communiquer avec le musicien (figure D.3). Elle permet notamment de changer rapidement de mode de contrôle et de type de voix entre les morceaux, voire de régler en temps réel l'individualisation de sa voix (explicité dans section 2.5).
- Un haut-parleur individuel placé derrière le musicien et permettant de transformer

Figure D.2 – *Le Chorus Digitalis en concert au Printemps de la Culture 2012*

le signal audio calculé en onde acoustique, qui est écouté par tous les musiciens et l'auditoire, comme illustré sur la figure D.2.

D.2 Difficultés rencontrées avec l'ensemble

D'une manière générale, *Chorus Digitalis* présente les mêmes caractéristiques qu'une chorale de voix naturelles ou même qu'un groupe d'instruments plus classiques. Par conséquent, les difficultés rencontrées dans le cadre du travail musical en chorale sont valables également ici, à savoir :

- les problèmes humains : disponibilité de tous à un créneau commun et sur la durée ; équilibre entre travail personnel et travail collectif ; différence de goûts musicaux de chacun ; différence de niveau musical ; efficacité des répétitions parfois difficiles quand le nombre de musiciens est élevé ; etc.
- les problèmes logistiques : un ensemble nécessite de grand locaux, où l'on peut faire du bruit ; laisser le matériel encombrant sur place ; etc.
- les problèmes musicaux : justesse inter-musiciens ; synchronisation temporelle ; équilibre des niveaux sonores des voix ; etc.

Mais il existe cependant un certain nombre de particularités spécifiques à la chorale de voix synthétiques qu'on détaille dans les paragraphes suivants.

Pratique musicale dans un cadre de recherche scientifique

Étant donné que les musiciens de la chorale sont quasiment tous membres de notre laboratoire, et que notre ensemble nécessite l'utilisation de matériel du laboratoire (en particulier les hauts-parleurs amplifiés avec leur pied, chers et lourds à transporter), nous avons choisi de répéter dans l'enceinte du laboratoire. Or, ce dernier ne disposant pas de locaux dédiés à ces pratiques musicales de répétition, occasionnant des restrictions dans la durée et le moment de la journée afin de ne pas trop gêner nos collègues. Notre laboratoire dispose de cabines acoustiques pour les expériences acoustiques, mais elles sont trop petites et acoustiquement non adaptées à la pratique d'un groupe musical.

FIGURE D.3 – *Capture d'écran de l'application de l'instrument*

Encombrement des dispositifs informatiques et audio

Le travail de cette thèse a été financé par le projet FEDER *OrJo*[2] (Orchestre de Joysticks). La société *Puce Muse* y développe un logiciel, la *Méta-Mallette* [DLG08], destinée à jouer des instruments audio-visuels à plusieurs et à partir d'un même ordinateur. L'idée est de transformer le *Personnal Computer* en *Collective Computer*. Dans ce projet, Le LIMSI a pour tâche de développer des instruments de voix numériques. Pour des raisons techniques développées dans l'annexe E liées à la latence de notre instrument, nous n'avons pas utilisé cette approche mais plutôt 1 ordinateur par voix de synthèse, amenant à 6 ordinateurs, 6 tablettes, et 6 hauts-parleurs avec leur pieds. Des solutions pour diminuer la latence en Max et/ou en utilisant une tablette graphique sous Max sont proposés dans cette annexe.

Également notifié dans le Plork [TCSW06], nous subissons une perte de temps non négligeable pendant les répétitions et les concerts au lancement des applications informatiques et aux branchements des câbles nécessaires, mais avant tout à résoudre les problèmes techniques de chacun. La plupart du temps, seule une personne dans l'ensemble en avait les compétences, devant passer successivement d'un musicien à l'autre pour corriger ces problèmes. Avec la pratique, ce temps a diminué grâce à la diminution du nombre de bogues informatiques, d'une meilleure interface utilisateur pour le changement de morceau, mais surtout grâce à l'apprentissage des musiciens au fonctionnement du système et à la résolution des problèmes simples. Ici, l'autonomie de chacun des musiciens vis-à-vis des problèmes informatiques est primordiale pour gagner du temps, d'autant plus que nous utilisions un ordinateur/tablette/haut-parleur pour chacun des musiciens. Pour contrer ce problème, les nouveaux musiciens intégrant l'ensemble Plork se voient proposer une petite formation à la

2. http://pucemuse.com/orjo/

résolution des problèmes informatiques simples d'une façon autonome. De plus, une ou deux personnes sont destinées exclusivement à la résolution d'éventuels problèmes informatiques / sonores pendant les répétitions et les concerts. Notre ensemble n'a pas encore connu un renouvellement de ses membres pour entreprendre une telle formation, mais cela reste une bonne idée à conserver pour plus tard.

Inconvénients de notre instrument

Contrairement à la voix naturelle qui est un instrument logé à l'intérieur de notre corps, l'interface de l'instrument nécessite d'avoir presque constamment les yeux concentrés sur la tablette. Cela a des avantages mais aussi des inconvénients. Il est ainsi bien plus difficile qu'en choeur de suivre un chef d'orchestre ou de se faire des signes visuels entre les musiciens. Cela devrait nous pousser à entreprendre à la réalisation d'une interface munie de repères tactiles ou disposés différemment. Par exemple, un simple plan incliné pour poser la tablette permettrait de relever le champs du regard et ainsi d'être visuellement plus ouvert aux autres musiciens.

Un des intérêts d'un instrument vocal externe au corps permet d'utiliser des repères autres que sonores pour surveiller sa justesse. En effet, un des problèmes dans une chorale est d'avoir des difficultés à entendre sa propre voix à travers le mélange des différentes voix. Ici, les repères visuels des hauteurs de notes sur la tablette permettent d'avoir un contrôle visuel sur la justesse, même en perdant temporairement de l'oreille la voix qu'on contrôle.

Le son ne sort pas physiquement de l'instrument, comme tout instrument électrique (clavier, guitare électrique, ...). Il est plutôt rare d'avoir des ensembles d'un même instrument électrique (orchestre de claviers, orchestre de guitares électriques, ...). Ici, nous sommes souvent 6 avec le même instrument de voix de synthèse. Nous avons donc, comme beaucoup d'ensembles d'instruments numériques, installé des hauts-parleurs individuels derrière chaque musicien pour diffuser le son que le musicien produit.

Enfin, une autre différence avec une chorale de voix naturelles repose sur l'indépendance du type de voix et du musicien. Ici, chaque musicien peut choisir d'en changer à tout moment, le modèle de voix étant ajustable comme détaillé dans la section 2.5.

D.3 Le répertoire des concerts

Nous avons montré qu'il est possible de jouer de nombreux styles de musique avec notre instrument *Cantor Digitalis* et son ensemble *Chorus Digitalis*. Nous présentons ici le répertoire du groupe selon trois styles musicaux : musique européenne classique, musique savante d'inde du nord, et musique actuelle.

D.3.1 Liste des concerts

Au 1er juin 2013, les concerts donnés par le *Chorus Digitalis* sont les suivants (ordre anti-chronologique) :

- *29 mai 2013, conférence New Interface for Musical Expression, KAIST Auditorium, Daejon, Corée du Sud.*
  Nous avons soumis une suite de 15 min acceptée dans le cadre des concerts de la conférence[3], en collaboration avec un membre du VoxTactum (figure D.4).

3. http://nime2013.kaist.ac.kr/program/concerts/

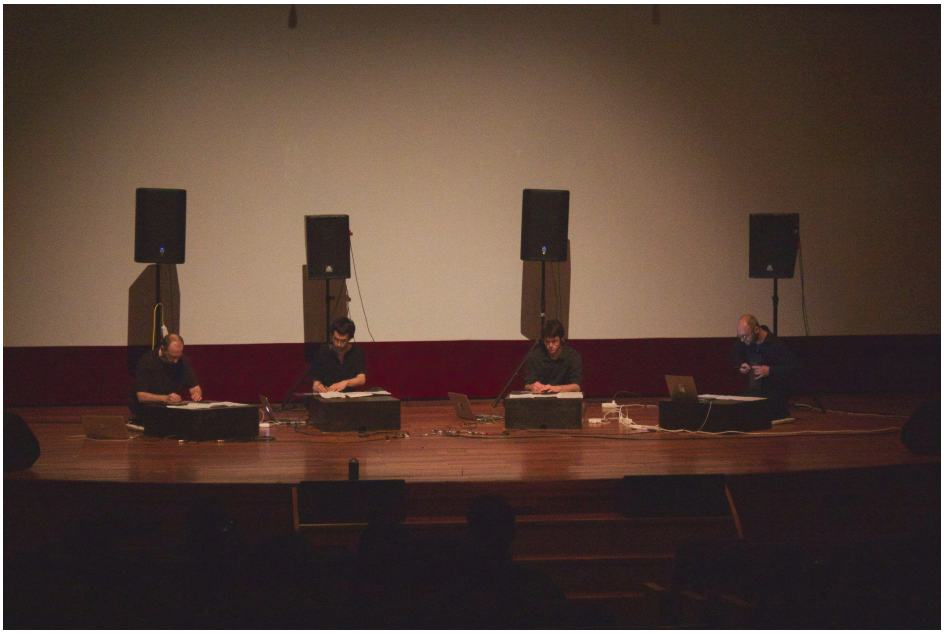


FIGURE D.4 – *Le Chorus Digitalis en concert avec Vox Tactum à la conférence NIME 2013, Daejon*

- *13 octobre 2012, Journée Science & Musique, Diapason, campus Universitaire de Beaulieu, Rennes.*
  Un concert de 15 min est donné à la suite d'une conférence sur la fabrique d'instrument de voix où nous avons été invités comme intervenants pour la journée. La conférence explique sous forme d'un dialogue entre un chanteur et un chercheur-développeur le processus de réflexion pour aboutir à la modélisation numérique et instrumentale de la voix chantée [4] (figure D.5).
- *22 juin 2012, Fête de la musique, Laboratoire « Lutherie-Acoustique-Musique » (LAM) de l'institut Jean Le Rond d'Alembert.*
  Un concert de 15 min est donné, suite à la proposition en interne d'un des musiciens.
- *25 mai 2012, Journées Arts Sciences, Printemps de la culture, Université Paris Sud, Orsay.*
  Le projet de concert de 45 min a été sélectionné par le comité d'organisation du festival [5] (figure D.6).
- *14 mars 2011, First International Workshop on Performative Speech and Singing Synthesis (P3S), University of Bristish Columbia, Vancouver, Canada.*
  Nous avons inauguré le *Chorus Digitalis* lors de ce workshop international qui proposait de faire des démonstrations de nos systèmes de synthèse temps réel. Notre concert a duré moins de 10 minutes [6] (figure D.7).

D.3.2 Chorale classique européenne

Voir fichier audio / vidéo 11

4. http://jsm2012.irisa.fr/toppage.php?page=programme
5. http://www.crea.u-psud.fr/ed-2012/programme.html
6. http://www.magic.ubc.ca/p3s/program.html

Figure D.5 – *Le Chorus Digitalis en concert à la Journée Science Musique 2012, Rennes. Crédit Photos : B. Arnaldi.*

Figure D.6 – *Le Chorus Digitalis en concert au Printemps de la Culture 2012, Orsay*

Figure D.7 – *Le Chorus Digitalis en concert à la conférence P3S 2011, Vancouver*

Les polyphonies baroque et médiévale ont été les premiers morceaux à avoir été travaillés, constituant pour le groupe un bon premier défi à atteindre.

Le premier morceau est *Alta Trinita Beata* (partition sur la figure D.8), une polyphonie médiévale anonyme à quatre voix. Le rythme est simple et les notes qui s'enchaînent offrent de petits intervalles faciles à réaliser avec la tablette (moins vrai pour la ligne de basse). Nous l'avons d'abord joué en quartet pour un concert donné pendant le workshop P3S à Vancouver en mars 2011. A chaque musicien était associé une des quatre voix (basse, ténor, alto, soprano). Les paroles originales, en latin, ont été remplacé par la voyelle /a/. Ainsi, nous pouvions nous concentrer sur la hauteur des notes, les nuances et la synchronisation.

Le deuxième morceau est le choral à 4 voix *Wie Schön leuchtet der Morgenstern* de Johan-Sebastian Bach (partition sur la figure D.9). Il a été joué également à Vancouver, avec la seule voyelle /a/.

Ces deux morceaux ont été rejoués 1 an plus tard en mai 2012 à Orsay lors du festival Printemps de la Culture. Cette fois-ci, les voyelles évoluaient dans *Wie Schön leuchtet der Morgenstern* comme indiqué sur la partition du morceau (figure D.9), et étaient contrôlées par l'axe Y de la tablette graphique de la main principale (voir section 2.7.3). Le morceau *Alta Trinita Beata* était accompagné par un clavecin.

Le troisième morceau est la sarabande de la suite en ré mineur joué au clavecin et l'air *Lascia ch'io pianga* joué au *Cantor Digitalis*, extrait de « Rinaldo » de George Frideric Händel (partition sur la figure D.8). La voix est jouée en solo et le musicien peut alors se concentrer sur l'expressivité de sa voix, soutenue par l'accompagnement au clavecin.

Enfin le quatrième et dernier morceaux travaillé de style classique européen est l'air du *Génie du Froid* de Henry Purcell, joué en quatuor à la conférence NIME en mai 2013 en collaboration avec un membre du *Vox Tactum*. Trois voix tapissent le morceau de croches régulières sur la voyelle /a/ tandis que la voix Basse expose le thème principal en improvisant les voyelles.

D.3.3 Chant vocal d'Inde du nord

Voir fichiers audios / vidéos 12

En voix solo également mais dans un autre style que la sarabande, nous avons interprété un chant vocal de musique savante d'Inde du nord. Il s'agît du raga *Miyan Ki Malhar* en *Teental* dans le style *Khayal*. Ce chant est accompagné traditionnellement par des tablas et une tampura. Le joueur de tablas joue le *Tala*, c'est à dire le rythme de base utilisé, ici le Teental. Tout en marquant le tala, il accompagne le chanteur soliste et peut également devenir soliste à son tour. La tampura est un instrument à cordes accordées sur la tonalité du raga, défini lui-même par les notes autorisées ainsi que des règles sur leur ordre. Elle permet d'asseoir le jeu du chanteur dans la bonne tonalité tout au long du morceau. Le soliste commence par introduire le raga lors de l'*Alap* où le cycle rythmique n'est pas encore défini, et en chantant lentement. Puis vient la partie *Bandish* où le joueur de tablas intervient et le chanteur phrase en respectant le cycle rythmique et le raga, à différents tempos successifs croissants

Le raga a été interprété 3 fois en concert jusqu'à l'écriture de cette thèse. Au printemps de la culture en mai 2012, la tampura a été remplacée par la superposition de plusieurs *Cantor Digitalis* imitant un chant diphonique en gardant la note fondamentale du morceau et en modifiant lentement leur couleur vocalique. Ces voix tapissaient alors le fond sonore avec la tonalité du morceau et remplissait la fonction de la tampura. Pour la fête de la musique en juin 2012, nous avons utilisé une vrai tampura jouée par une troisième personne. Enfin,

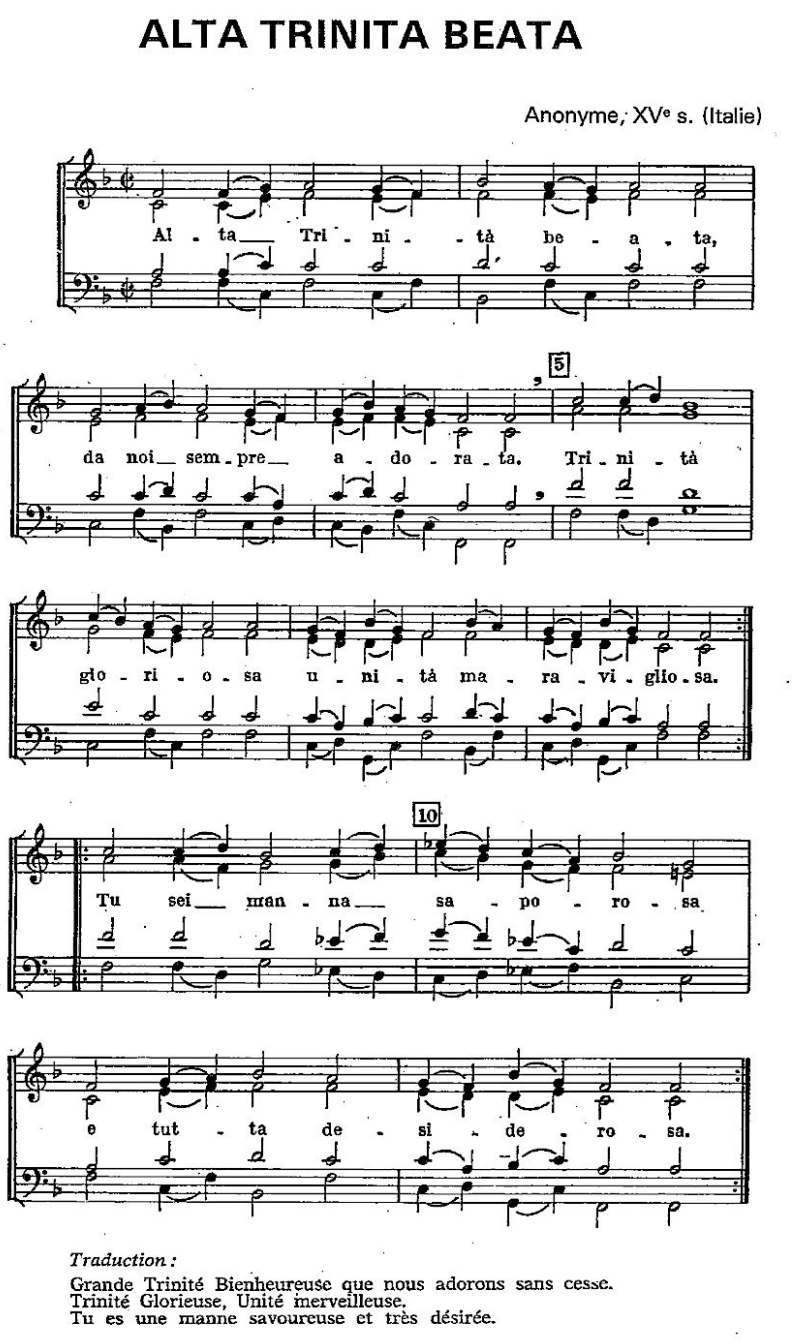


Figure D.8 – *Partition du morceau Alta Trinita Beata (anonyme)*

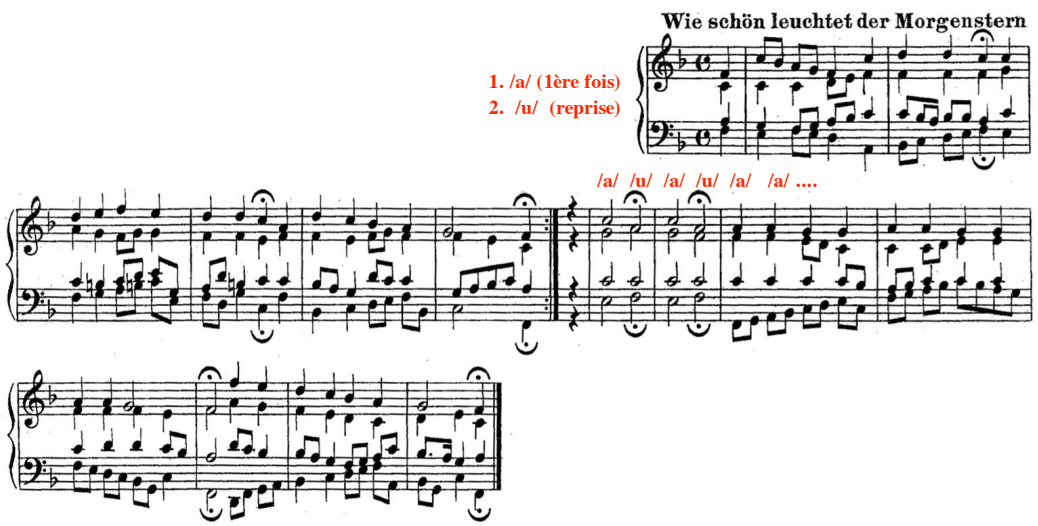


FIGURE D.9 – *Partition du morceau Wie Schön leuchtet der Morgenstern (Johan-Sebastian Bach)*

pour le concert de la journée Science et Musique en octobre 2012, nous avons enregistré une tampura et nous avons lu le fichier audio pendant le concert. A chaque fois, le raga était accompagné par un joueur de tablas et le morceau durait environ 15 minutes.

Pour le raga, le musicien jouant du *Cantor Digitalis* a préféré utiliser un autre calque de repère mélodique que celui présenté dans la section 2.7.3, présentant seulement les repères de notes du raga. Un raga conserve la même gamme de notes tout au long du morceau et le minimum de repères permet d'accroître la vitesse d'exécution en évitant d'être perturbé visuellement par des repères inutiles.

Lors des deux premiers concerts, un contrôle mono-manuel était utilisé comme présenté en 2.7.3, limitant l'articulation des voyelles sur un seul axe, en l'occurence /u,o,a,e,i/. Pour le raga joué en octobre 2012, les voyelles étaient contrôlées dans un espace 2D permettant toute combinaison de ces voyelles à l'aide d'une deuxième tablette graphique.

Le style vocal classique d'Inde du nord est bien adapté à un jeu à la tablette où les nombreux *portamentos* prennent un sens spatial simple (relier deux notes par un glissement du stylet et contrôler sa vitesse) et sont exécutés relativement facilement grâce aux repères de notes. Nous avons rallongé volontairement la partie *Alap* du raga par rapport à ce qui se fait généralement, car elle permet, par sa lenteur, de bien utiliser les capacités de notre instrument. Le musicien jouant le raga a beaucoup travaillé sa technique de tablette pour reproduire ce style particulier de chant. Le résultat est très probant.

D.3.4 Chorale contemporaine

Voir fichier audio / vidéo 13

Six morceaux de style plus actuel ont été interprétés ou composés pour le *Chorus Digitalis*. Ils ont tous été joués une seule fois pour le festival Printemps de la Culture en mai 2012, à l'exception de *Picato*, morceau introductif du concert du Workshop P3S à Vancouver au

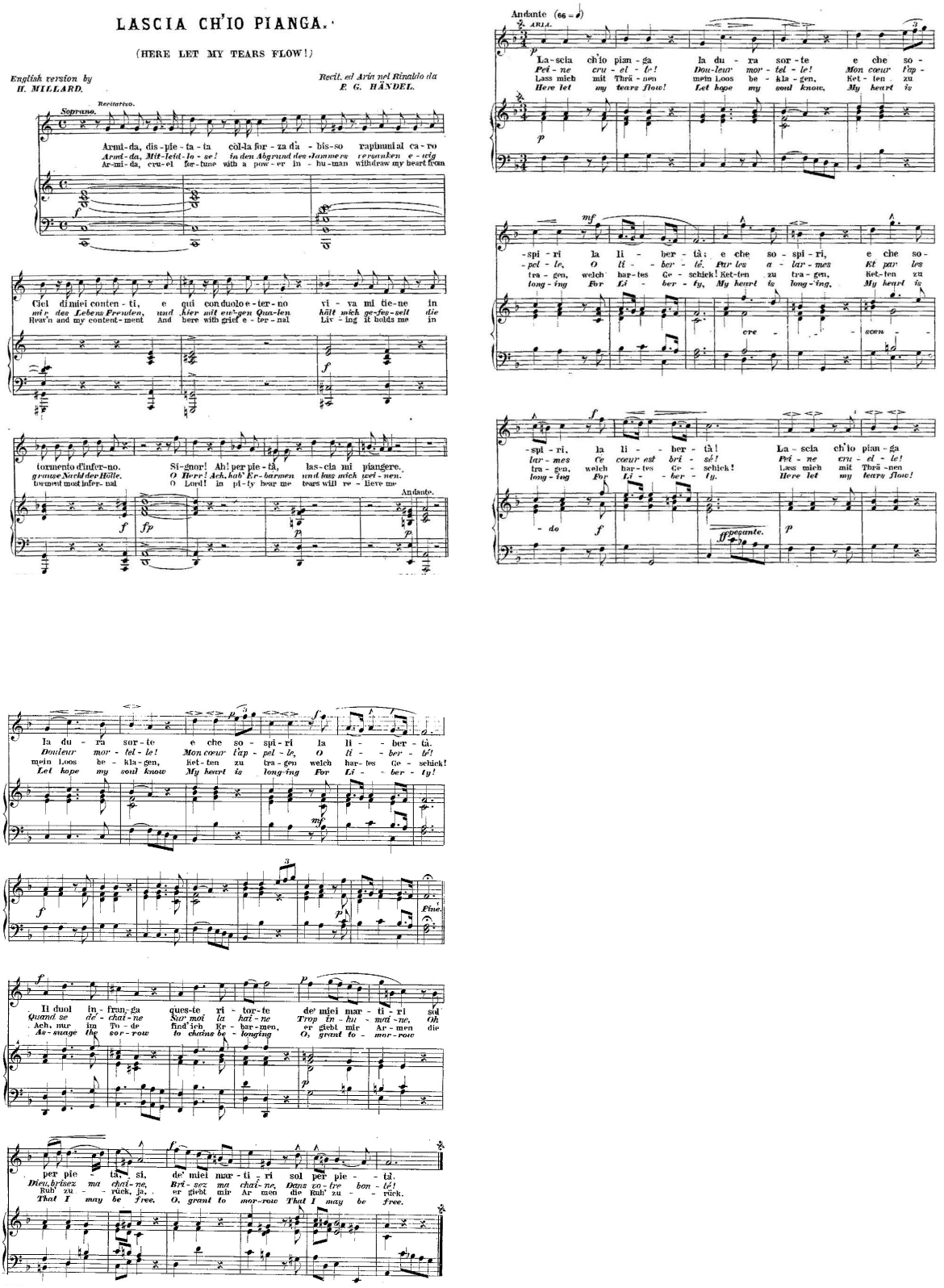


Figure D.10 – *Partition de la Sarabande de la suite en ré mineur, et Air « Lascia ch'io pianga", extrait de « Rinaldo » (George Frideric Händel)*

Canada en mars 2011, et une partie de la suite jouée à la conférence NIME à Daejon en Corée du Sud en mai 2013.

Le premier est une pièce de musique minimaliste de Philip Glass, nommée *Polar Star* (partition à la figure D.11). Elle est constituée de la superposition graduelle des différentes voix jouant des motifs répétitifs rapides. La difficulté réside dans l'équilibre entre rapidité du geste et conservation de la justesse, d'autant plus que les intervalles à jouer sont grands (se traduisant sur notre interface par une longue distance à parcourir). L'ensemble du groupe a été mis à contribution, avec six voix (dont deux altos et deux soprano), le tout accompagné par un clavier (partie « Orgue » sur la partition).

Le deuxième morceau est *Ocean* de Björk et Dirty Projector (partition à la figure D.12). Il s'agît d'une pièce à 4 voix dont une sourdine constante tout au long du morceau. Les 3 autres voix partent d'une même note puis atteignent chacune en même temps une cible différente par un lent portamento sur la voyelle /a/. Le tout est répété sur la voyelle /ɛ/.

On reproduit d'assez près l'original avec le *Chorus Digitalis*. Comme pour le raga, l'instrument est bien adapté et le résultat global est très satisfaisant.

Le troisième morceau est l'interprétation de *Valse*, un choeur moderne du compositeur Bruno Lecossois (partition à la figure D.13). Il présente 5 voix dont une dédoublée (voix soprano 1), assez différentes les unes des autres. La voix soprano 1 est lente et expressive sur une voyelle tournant autour de /ɔ/ ; la soprano 2 marque la pulsation avec une voix soufflée alternant entre /u/ et /a/ ; la voix de ténor et d'alto utilisent des battements sur le cage thoracique pour couper le souffle et obtenir une succession rapide de silence et de voyelle /e/ ou /o/ en suivant la tonalité du morceau ; enfin la voix basse chante un motif rythmique cyclique sur la syllabe /ko/.

L'interprétation est bien réussie. Une voix soufflée (voir section 2.5) est utilisée pour la soprano 2. L'effet saccadé des voix de ténor et d'alto est reproduit facilement en relevant rapidement le stylet après chaque note (geste percussif en visant la voyelle et la note). Enfin, on utilise la voyelle /o/ à la place de la syllabe /ko/ pour la voix de basse (à l'époque, l'instrument de syllabes n'était pas achevé).

Les trois morceaux restant sont des compositions, il est donc difficile de les comparer à un existant. *Canticum Novum* est une pièce de style « rock progressif » pour *Cantors Digitalis*, une basse, des tablas et une guitare. Elle met sur le même plan les voix de synthèse et les autres instruments, en alternant accompagnement et solo. L'accompagnement à la voix se fait sous la forme d'un tapis sonore par imitation de chant diphonique en jouant des voix de basse et en modifiant doucement la première ou deuxième fréquence centrale des filtres formantiques (modélisant la 1ère et la 2ème résonance du conduit vocal).

Picato est le morceau d'introduction du concert donné à Vancouver en mars 2011. C'est un morceau improvisé, assez démonstratif, qui permettait de donner un bref aperçu des différentes voix utilisées alors (basse, ténor, alto, soprano). Il est constitué de chants diphoniques, de longs portamentos et de notes piquées.

Enfin, la suite jouée à la conférence NIME était composée de deux morceaux classiques alternés par des pièces contemporaines, d'une durée totale de 15 minutes. La première partie contemporaine reprenait le principe des chants diphoniques du concert du Printemps de la Culture, combinées à des improvisations et quelques grognements d'animaux. La seconde partie contemporaine débute sur des syllabes parlées dénuées de sens phonologique, d'abord

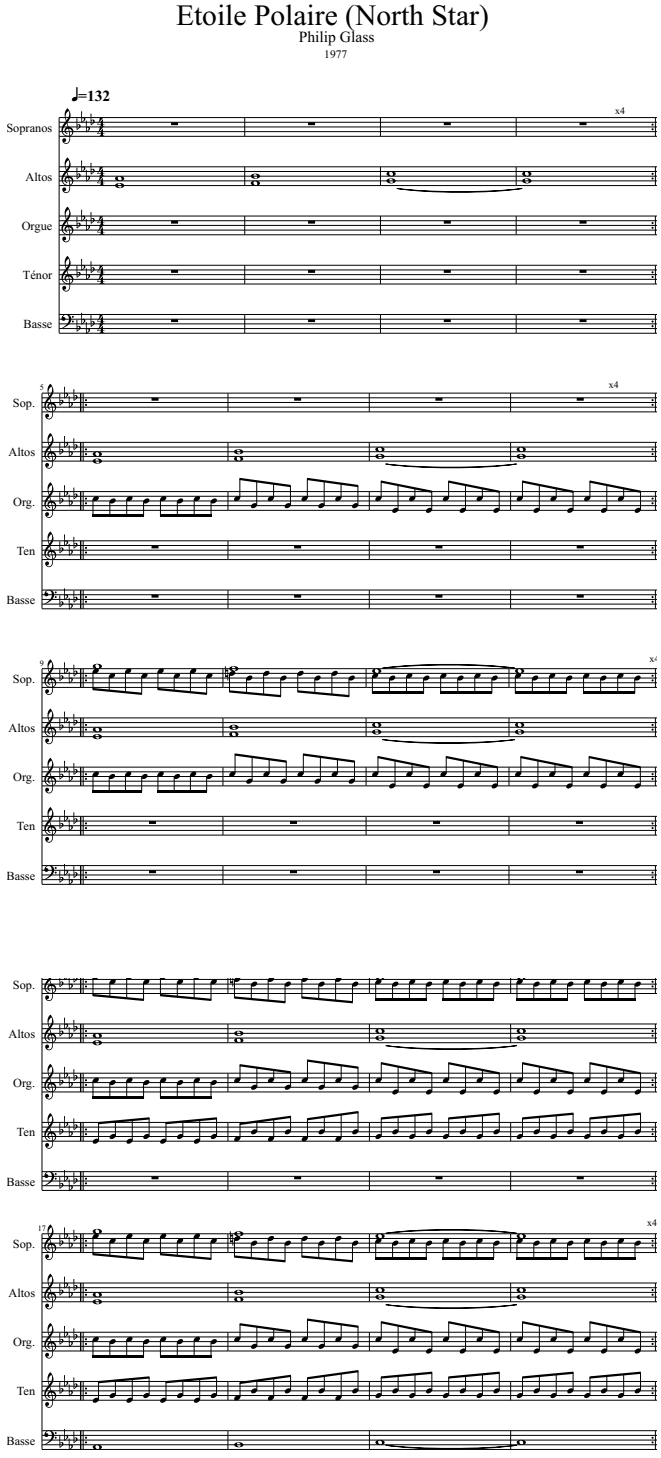


Figure D.11 – *Partition du morceau North Star (Philip Glass), relevé par M. Delorme*

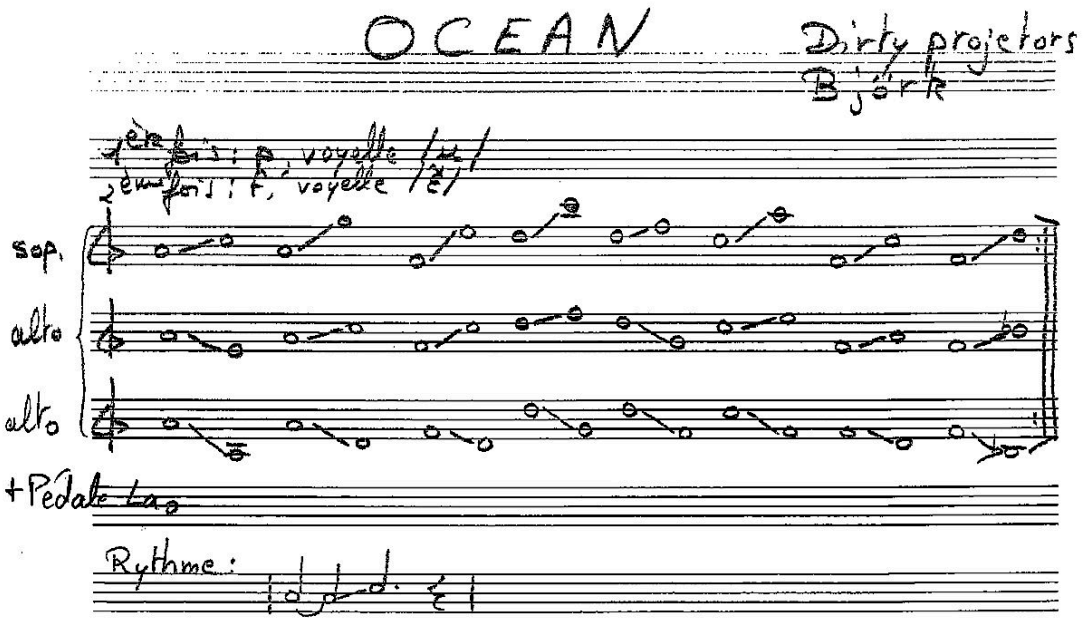


FIGURE D.12 – *Partition du morceau Ocean (Bjork & Dirty Projector), relevé par B. Doval*

en monologue puis en dialogue, où un des musiciens faisait intervenir le Digitartic. Puis le dialogue se transforme en une succession de courtes comptines enfantines et airs connus sur un ton comique, jusqu'à que les synthétiseurs miment une « explosion de rire » contagieuse (voir [OW13] pour l'analyse et la synthèse de rires). Cette seconde partie se termine par un glissando donnant l'effet de monter indéfiniment, en reprenant le principe de Shepard [She64].

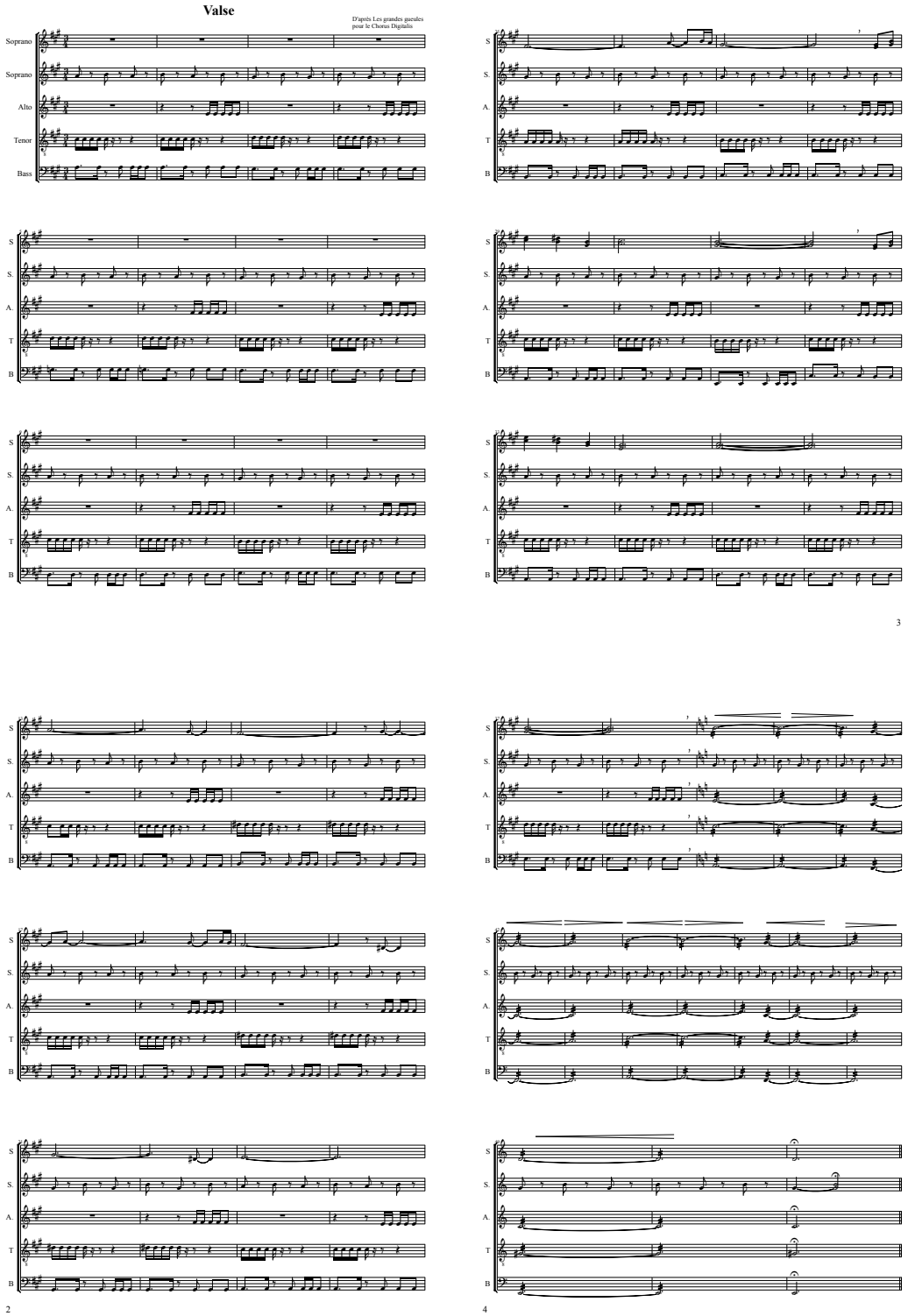


Figure D.13 – *Partition du morceau Valse (Bruno Lecossois), relevé par L. Feugère et B. Doval*

Annexe E

Techniques pour le temps-réel

Le travail de cette thèse a été financé par le projet FEDER *OrJo*[1] (Orchestre de Joysticks). La société *Puce Muse* y développe un logiciel, la *Méta-Mallette* [DLG08], destinée à jouer des instruments audio-visuels à plusieurs et à partir d'un même ordinateur. L'idée est de transformer le *Personnal Computer* en *Collective Computer*. Le LIMSI a pour tâche de développer des instruments de voix numériques.

Au début, nous avons voulu utiliser la Méta-Mallette pour jouer de nos voix de synthèse. Ainsi, avec seulement un ordinateur, il aurait été possible de calculer toutes les voix de la chorale, de centraliser les réglages, et surtout d'utiliser la facilité d'utilisation de la méta-mallette qui est conçue pour gérer les ensembles d'instruments numériques, notamment pour l'affectation des interfaces et le réglage du son.

Mais nos instruments nécessitent trop de puissance de calcul pour être utilisés à plusieurs sur le type d'ordinateur à notre disposition, à savoir typiquement des MacBookPro5,1 datant de 2008 (Processeur Intel Core 2 Duo de 2,4 GHz). Alors qu'au repos le patch Max/MSP utilise 20-25% du processeur (surtout à cause du modèle de source glottique), le taux d'utilisation approche les 90% si les mouvements sur la tablettes sont rapides, utilisation partagée entre le driver d'acquisition des données de la tablette et les patchs Max/MSP.

Ci-dessous sont énumérés quelques conseils pour les personnes utilisant des instruments temps réels sous Max/MSP et ne disposant pas d'ordinateur très puissant.

- Dans le système d'exploitation de l'ordinateur :
  1. fermer tous les autres programmes ;
  2. retirer tout périphérique USB qui pourrait enclencher une indexation de fichiers et donc consommer du CPU (programme Spotify sous Mac par exemple) ;
  3. libérer de l'espace sur le disque dur de travail s'il est trop faible (au moins 10 Go semblent nécessaire sous Mac OSX).
- Dans Max/MSP :
  1. fermer les fenêtres Max/MSP inutiles, surtout si elles contiennent des objets graphiques nécessitant du temps de calcul, tel un spectrogamme ;
  2. séparer et exporter en application le patch contenant l'objet qui réceptionne les données de la tablette Wacom, et les renvoyer via le protocole UDP (objets *udpsend* / *udpreceive*) vers le patch Max/MSP de synthèse audio. De cette manière, il

1. http://pucemuse.com/orjo/

semblerait que Mac traite en priorité la réception des données issues de la tablette Wacom (conseil donné par Jean-Michel Couturier) ;

3. utiliser un limiteur de débit de données à leur réception de la tablette par l'objet Max/MSP *s2m.wacom*. Dans Max/MSP, l'objet *speedlim* permet d'échantillonner le flux de données à une période à partir de 1 ms.

Annexe F

Liste des productions scientifiques

F.1 Communications

Revue international

C. d'Alessandro, L. Feugère, S. Le Beux, O. Perrotin, A. Rilliard (ordre alphabétique).
Drawing melodies : Evaluation of Chironomic Singing Synthesis.
J. Acoust. Soc. Am. (soumis).

Conférences internationales avec actes

L. Feugère, C. d'Alessandro, B. Doval [FdD13].
Performative voice synthesis for edutainment in acoustic phonetics and singing : a case study using the "Cantor Digitalis".
5th International Conference on Intelligent Technologies for Interactive Entertainment, July 3-5, 2013 Mons, Belgium, à paraître dans « Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering series. Volume 0124 ».

L. Feugère, C. d'Alessandro [Fd13].
Digitartic : bi-manual gestural control of articulation in performative singing synthesis.
13th International Conference on New Interfaces for Musical Expression (NIME 2013), Daejon, Korea, 27/05 au 30/05, 2013, 331-336.

S. Le Beux, L. Feugère, C. d'Alessandro [LBFd11].
Chorus digitalis : experiment in chironomic choir singing.
12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011), Firenze, Italy, 27/08 au 31/08, 2011, 2005-2008. Paru dans Proceedings of the conference ISSN : 1990-9772.

L. Feugère, S. Le Beux, C. d'Alessandro [FLBd11].
Chorus digitalis : polyphonic gestural singing.
1st International Workshop on Performative Speech and Singing Synthesis (P3S 2011), Vancouver (Canada), 14/03 au 15/03, 2011, 4p.

Conférences nationales avec actes

L. Feugère, C. d'Alessandro [Fd12].
Digitartic : synthèse gestuelle de syllabes chantées.
Journées d'Informatique Musicale (JIM 2012), Mons, Belgique, 09/05 au 11/05, 2012, 219-225.

S. De Laubier, G. Bertrand, H. Genevois, V. Goudard, B. Doval, L. Feugère, S. Le Beux, C. d'Alessandro [DLBG+12].
OrJo et la Méta-Mallette 4.0.
Journées d'Informatique Musicale (JIM 2012), Mons, Belgique, 09/05 au 11/05, 2012, 227-232.

Conférences nationales sans acte

L. Feugère.
Modèles de contrôle instrumental de la synthèse vocale.
Journées Jeunes Chercheurs en Acoustique Audition et Signal Audio (JJCAAS), 7-9 décembre 2011, Orange Labs, Rennes.

L. Feugère.
Contrôle gestuel d'occlusives pour les instruments de synthèse vocale.
Journées Jeunes Chercheurs en Acoustique Audition et Signal Audio (JJCAAS), 17-19 novembre 2010, Paris.

Conférence invitée

L. Feugère et B. Doval.
Contrôle par le geste d'une voix chantée de synthèse. Comment faire vivre une voix de synthèse ?
Journée Science et Musique 2012, Diapason, campus Universitaire de Beaulieu, Rennes, 13 octobre 2012.

Rapport interne

L. Feugère, S. Le beux, C. d'Alessandro.
OrJo : développement d'instruments virtuels sur la synthèse vocale (LIMSI-CNRS).
Rapport de fin du projet FEDER OrJo, 2012.

F.2 Concerts Arts-Sciences

O. Perrotin, L. Feugère, C. d'Alessandro, B. Doval.
Concert Chorus Digitalis - Saison 3.1.
CuriositAS (CuriositAS 2013), Orsay, France, 7/10, 2013.

L. Feugère, O. Perrotin, C. d'Alessandro, B. Doval.
Concert Chorus Digitalis - Saison 3.0.
Journées nationales du Développement Logiciel (JDEV 2013), Palaiseau, France, 04/09 au 06/09, 2013.

N. d'Alessandro, C. d'alessandro, L. Feugère, M. Astrinaki, J. Wang, O. Perrotin, A. Pon, B. Doval.
Vox Tactum Meets Chorus Digitalis : Seven Years of Singing Surfaces.
13th International Conference on New Interfaces for Musical Expression, Daejeon + Seoul, Korea Republic, May 27-30, 2013.

L. Feugère et B. Doval.
Interprétation d'un raga indien, en duo tablas-tablette.
Journée Science et Musique 2012, Diapason, campus Universitaire de Beaulieu, Rennes, 13 octobre 2012.

L. Feugère, C. d'Alessandro, B. Doval.
Concert Chorus Digitalis - Saison 1.0.
Journées Arts Sciences, Printemps de la Culture 2012, 24 mai 2012.

L Feugère, S. Le Beux, C. d'Alessandro, B. Doval.
Concert Chorus Digitalis - Saison 0.1.
1st International Workshop on Performative Speech and Singing Synthesis (P3S 2011), Vancouver (Canada), 14/03 au 15/03, 2011.

F.3 Prix

L. Feugère.
Prix Jeune Chercheur « Science et Musique 2012 ».
Organisé par l'IRISA et parrainé par l'Association Française d'Informatique Musicale (AFIM) et la Fondation Rennes/Action Métivier .

Liste des tableaux

Table des figures

Bibliographie

[AdD12] Maria Astrinaki, Nicolas d'Alessandro et thierry Dutoit : Mage - a platform for tangible speech synthesis. *In Proceedings of the 12th Conference on New Interfaces for Musical Expression (NIME'12)*, pages 353–356, Ann Arbor, Michigan, USA, May 21-23 2012.

[AdP+12] Maria Astrinaki, Nicolas d'Alessandro, Benjamin Picart, Thomas Drugman et thierry Dutoit : Reactive and continuous control of hmm-based speech synthesis. *In IEEE Workshop on Spoken Language Technology (SLT 2012)*, Miami, Florida, USA, December, 2-5 2012.

[Bel11] Greg Beller : Gestural control of real-time concatenative synthesis in luna park. *In P3S (Performative Speech and Singing Synthesis)*, 2011.

[Bir] Peter Birkholz : http ://www.vocaltractlab.de/, visité le 01/04/2013.

[Bir07] Peter Birkholz : Articulatory synthesis of singing. *In Interspeech*, pages 4001–4004, Antwerp, Belgium, August, 27-31 2007. INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, ISCA.

[Bre83] David Brewster : *Letters on Natural Magic.* Numéro 267-271. Chatto and Windus, Piccadilly, London, 1883.

[BS03] Jose Bretos et Johan Sundberg : Measurements of vibrato parameters in long sustained crescendo notes as sung by ten sopranos. *Journal of Voice*, 17:343–352, 2003.

[Cad88] Claude Cadoz : Instrumental gesture and musical composition. *In Proceedings of the 1988 International Computer Music Conference (ICMC1998)*, pages 1–12, San Francisco, 1988.

[Car81] René Carré : Couplage conduit vocal-source vocale. *In 12ème Journée d'Étude sur la Parole*, pages 233–245, Montreal, 25, 26, 27 mai 1981 1981.

[CK02] Alain Cheveigné et Hideki Kawahara : Yin, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.*, 111(4):1917–1930, April 2002.

[CL00] Perry R. Cook et Colby N. Leider : squeezevox : A new controller for vocal synthesis models. *In Proceedings of the 2000 International Computer Music Conference (ICMC2000)*, Berlin, August 2000.

[Cok68] Cecil H. Coker : Speech synthesis with a parametric articulatory model. *In Proc. Speech. Symp.*, Kyoto, Japan, 1968.

[Coo91] Perry R. Cook : *Identification of control parameters in an articulatory vocal tract model, with applications to the synthesis of singing.* Thèse de doctorat, Stanford University, 1991.

[Coo11] Perry R. Cook : Thoughts, issues, and questions on designing and using computer controllers with vocal models for performance. *In P3S (Performative Speech and Singing Synthesis)*, 2011.

[CT89] Calliope et J. P. Tubach : *La parole et son traitement automatique.* 1989.

[d'A06] Christophe d'Alessandro : *Voice Source Parameters and Prosodic Analysis (in Methods in Empirical Prosody Research)*, pages 63–88. Walter de Gruyter, 2006.

[d'A09] Nicolas d'Alessandro : *Realtime and Accurate Musical Control of Expression in Voice Synthesis.* Thèse de doctorat, 2009.

[DBGP07] Simone Dalla Bella, Jean-François Giguère et Isabelle Peretz : Singing proficiency in the general population. *J. Acoust. Soc. Am.*, 121(2):1182–1189, February 2007.

[dD07] Nicolas d'Alessandro et Thierry Dutoit : Handsketch bi-manual controller, investigation on expressive control issues of an augmented tablet. *In Proceedings of the 7th Conference on New Interfaces for Musical Expression (NIME'07)*, New York, USA, 2007.

[dD09] Nicolas d'Alessandro et thierry Dutoit : Advanced techniques for vertical tablet playing a overview of two years of practicing the handsketch 1.x. *In Proceedings of the 9th Conference on New Interfaces for Musical Expression (NIME'09)*, 2009.

[DdH03] Boris DOVAL, Christophe D'ALESSANDRO et Nathalie HENRICH : The voice source as a causal/anticausal linear filter. *In* ISCA, éditeur : *Proceedings of Voqual'03 : Voice Quality : Functions, analysis and synthesis*, Geneva, Switzerland, 2003.

[DdH06] Boris DOVAL, Christophe D'ALESSANDRO et Nathalie HENRICH : The spectrum of glottal flow models. *Acta Acoustica*, 92:1026–1046, 2006.

[ddLB+05] Christophe D'ALESSANDRO, Nicolas D'ALESSANDRO, Sylvain LE BEUX, Juraj SIMKO, Feride ÇETIN et Hannes PIRKER : The speech conductor : Gestural control of speech synthesis. Rapport technique Final Project Report #6, eNTERFACE'05, Mons, Belgium, July 18th – August 12th 2005.

[ddLBD06] Nicolas D'ALESSANDRO, Christophe D'ALESSANDRO, Sylvain LE BEUX et Boris DOVAL : Real-time calm synthesizer : new approaches in hands-controlled voice synthesis. *In Proceedings of the 2006 International Conference on New Interfaces for Musical Expression (NIME06)*, pages 266–271, Paris, France, 2006.

[DH97] Jianwu DANG et Kiyoshi HONDA : Acoustic characteristics of the piriform fossa in models and humans. *J. Acoust. Soc. Am.*, 101(1):456–465, January 1997.

[DLBG+12] Serge DE LAUBIER, Guillaume BERTRAND, Hugues GENEVOIS, Vincent GOUDARD, Boris DOVAL, Lionel FEUGÈRE, Sylvain LE BEUX et Christophe D'ALESSANDRO : Orjo et la méta-mallette 4.0. *In Actes des Journées d'Informatique Musicale (JIM 2012)*, pages 227–232, Mons, Belgique, 9-11 mai 2012.

[DLG08] Serge DE LAUBIER et Vincent GOUDARD : Puce muse - la méta-mallette. *In Journée d'informatique musicale*, 2008.

[dPW+12] Nicolas D'ALESSANDRO, Aura PON, Johnty WANG, David EAGLE, Ehud SHARLIN et Sidney S. FELS : A digital mobile choir : Joining two interfaces towards composing and performing collaborative mobile music. *In Proceedings of the 12th Conference on New Interfaces for Musical Expression (NIME'12)*, University of Michigan, Ann Arbor, May 21-23 2012.

[DR73] Raymond G. DANILOFF et Hammarberg ROBERT : On defining coarticulation. *Journal of Phonetics*, 1:239–248, 1973.

[dRLB11] Christophe D'ALESSANDRO, Albert RILLIARD et Sylvain LE BEUX : Chironomic stylization of intonation. *J. Acoust. Soc. Am.*, 129(3):1594–1604, March 2011.

[DRW39] Homer DUDLEY, R. R. RIESZ et S. S. A. WATKINS : A synthetic speaker. *Journal of the Franklin Institute*, 227(6):739–764, 1939.

[DT50] Homer DUDLEY et Thomas H. TARNOCZY : The speaking machine of wolfgang von kempelen. *J. Acoust. Soc. Am.*, 22(2):151–166, 1950.

[Fan60] Gunnar FANT : *Acoustic theory of speech production*. Mouton, 1960.

[Fd12] Lionel FEUGÈRE et Christophe D'ALESSANDRO : Digitartic : synthèse gestuelle de syllabes chantées. *In Actes des Journées d'Informatique Musicale (JIM 2012)*, pages 219–225, Mons, Belgique, 9-11 mai 2012.

[Fd13] Lionel FEUGÈRE et Christophe D'ALESSANDRO : Digitartic : bi-manual gestural control of articulation in performative singing synthesis. *In Proceedings of the 13th Conference on New Interfaces for Musical Expression (NIME'13)*, pages 331–336, Daejeon + Seoul, Korea Republic, May 27-30 2013.

[FdD13] Lionel FEUGÈRE, Christophe D'ALESSANDRO et Boris DOVAL : Performative voice synthesis for edutainment in acoustic phonetics and singing : a case study using the "cantor digitalis". *In Proceeding of the 5th International Conference on Intelligent Technologies for Interactive Entertainment*, Mons, Belgium., July 3-5 2013.

[FH92] Sidney S. FELS et Geoffrey E. HINTON : Glove-talk : A neural network interface between a data-glove and a speech synthesizer. *IEEE Transactions on neural networks*, 3(6):1–7, November 1992.

[FH98] Sidney S. FELS et Geoffrey E. HINTON : Glove-talk ii : a neural network interface which maps gesture to parallel formants. *IEEE*, 9(1):205, 1998.

[FLBd11] Lionel FEUGÈRE, Sylvain LE BEUX et Christophe D'ALESSANDRO : Chorus digitalis : polyphonic gestural singing. *In 1st International Workshop on Performative Speech and Singing Synthesis (P3S 2011)*, Vancouver (Canada), 14/03 au 15/03 2011.

[FLBdD11] Lionel FEUGÈRE, Sylvain LE BEUX, Christophe D'ALESSANDRO et Boris DOVAL : Chorus digitalis performance. *In 1st International Workshop on Performative Speech and Singing Synthesis (P3S 2011)*, Vancouver, Canada, University of British Columbia, 14/03 au 15/03 2011.

[FPL09] Sidney S. FELS, Robert PRITCHARD et Allison LENTERS : Fortouch : A wearable digital ventriloquized actor. *In Proceedings of the 9th Conference on New Interfaces for Musical Expression (NIME'09)*, 2009.

[FS58] James L. FLANAGAN et Michael G. SASLOW : Pitch discrimination for synthetic vowels. *J. Acoust. Soc. Am.*, 30(5):435–442, May 1958.

[FSG12] Samia FRAJ, Jean SCHOENTGEN et Francis GRENEZ : Development and perceptual assessment of a synthesizer of disordered voices. *J. Acoust. Soc. Am.*, 132(4):2603–2615, October 2012.

[Fuc] Susanne FUCHS : http ://benoit.susannefuchs.org/tutorial3.html, visité le 01/04/2013.

[Gen99] Hugues GENEVOIS : *Geste et pensée musicale : de l'outil à l'instrument (dans "Les nouveaux gestes de la musique")*, pages 35–45. 1999.

[GHNO03] Masataka GOTO, Hiroki HASHIGUCHI, Takuichi NISHIMURA et Ryuichi Ryuichi Oka OKA : Rwc music database : Music genre database and musical instrument sound database. *In Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR 2003)*, 2003.

[GHWS10] Maeva GARNIER, Nathalie HENRICH, Joe WOLFE et John SMITH : Vocal tract adjustments in the high soprano range. *jasa*, 2010.

[GIKGK13] Roni Y. GRANOT, Rona ISRAEL-KOLATT, Avi GILBOA et Tsafrir KOLATT : Accuracy of pitch matching significantly improved by live voice model. *Journal of Voice*, 27(3):390.e13–390.e20, March 2013.

[Gor87] John W. GORDON : The perceptual attack time of musical tones. *J. Acoust. Soc. Am.*, 82(2):88–105, 1987.

[GR94] Martine GARNIER-RIZET : *Elaboration d'un module de règles phonético-acoustiques pour un système de synthèse à partir du texte pour le français.* Thèse de doctorat, Université de la Sorbonne nouvelle, 1994.

[Hau99] Jean HAURY : *Petite histoire illustrée de l'interface clavier (dans "Les nouveaux gestes de la musique")*, pages 93–110. 1999.

[HdD01] Nathalie HENRICH, Christophe D'ALESSANDRO et Boris DOVAL : Spectral correlates of voice open quotient and glottal flow asymmetry : theory, limits and experimental data. *In Proc. Eurospeech*, Aalborg, September 2001.

[HdDC05] Nathalie HENRICH, Christophe D'ALESSANDRO, Boris DOVAL et Michèle CASTELLENGO : Glottal open quotient in singing : Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency. *J. Acoust. Soc. Am.*, 117(5):1417–1430, 2005.

[Hen01] Nathalie HENRICH : *Etude de la source glottique en voix parlée et chantée : modélisation et estimation, mesures acoustiques et électroglottographiques, perception.* Thèse de doctorat, Université Paris 6, Novembre 2001.

[Hes59] Donald A. HESS : Pitch, intensity, and cleft palate voice quality. *In Journal of Speech and Hearing Research*, volume 2, pages 113–125, 1959.

[HKT+04] Kiyoshi HONDA, Tatsuya KITAMURA, Hironori TAKEMOTO, Satoru FUJITA et Mokhtari PARHAM : Resonance characteristics of hypopharyngeal cavities. *In Proceedings of the International Conference on Voice Physiology and Biomechanics (ICVPB'04)*, pages 81–82, Marseille, France, 2004.

[HMC89] Christian HAMON, Eric MOULINES et Francis CHARPENTIER : A diphone synthesis system based on time-domain prosodic modifications of speech. *In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 238–241, 23-26 May 1989.

[Hol83] J.N. HOLMES : Formant synthesizers : cascade or parallel ? *Speech Communication*, 2:251–273, 1983.

[HSW11] Nathalie HENRICH, John SMITH et Joe WOLFE : Vocal tract resonances in singing : Strategies used by sopranos, altos, tenors, and baritones. *J. Acoust. Soc. Am.*, 129(2):1024–1035, February 2011.

[IPA05] International Phonetic Association Chart, available under a Creative Commons Attribution-Sharealike 3.0 Unported License. Copyright © 2005. 2005.

[JSW04] Elodie JOLIVEAU, John SMITH et Joe WOLFE : Vocal tract resonances in singing : the soprano voice. *J. Acoust. Soc. Am.*, 116(4):2434–2439, 2004.

[KB08] Bernd J. KRÖGER et Peter BIRKHOLZ : Articulatory synthesis of speech and singing - state of the art and suggestions for future research. *In COST 2102 School (Vietri)*, pages 306–319, 2008.

[Kes02] Loic KESSOUS : Bi-manual mapping experimentation, with angular fundamental frequency control and sound color navigation. *In Proceedings of the International Conference on New Interfaces for Musical Expression (NIME'02)*, pages 113–114, 2002.

[Kes04] Loic KESSOUS : *Contrôles gestuels bi-manuels de processus sonores*. Thèse de doctorat, Université de Paris VIII, 9 novembre 2004.

[KHT05] Tatsuya KITAMURA, Kiyoshi HONDA et Hironori TAKEMOTO : Individual variation of the hypopharyngeal cavities and its acoustic effects. *Acoust. Sci. & Tech.*, 26(1):16–26, 2005.

[Kip88] James KIPPEN : *The tabla of Lucknow – a cultural analysis of a musical tradition*. Cambridge University Press, 1988.

[Kla80] Dennis H. KLATT : Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Am.*, 67(3):971–995, March 1980.

[Kla87] Dennis H. KLATT : Review of text-to-speech conversion for english. *J. Acoust. Soc. Am.*, 82(3): 737–793, September 1987.

[KMKdC99] Hideki KAWAHARA, MASUDA-KATSUSE et Alain de CHEVEIGNÉ : 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*, 27:187–207, 1999.

[KO07] Hideki KENMOCHI et Hayato OSHITA : Vocaloid – commercial singing synthesizer based on sample concatenation. *In Interspeech*, 2007.

[KQS+11] Aki KUNIKOSHI, Yu QIAO, Daisuku SAITO, Nobuaki MINEMATSU et Keikichi HIROSE : Gesture design of hand-to-speech converter derived from speech-to-hand converter based on probabilistic integration model. *In Interspeech*, page 3025, August 2011.

[Lar77] Bjorn LARSSON : Music and singing synthesis equipment (musse). *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR)*, 18(1):38–40, 1977.

[Lav] http ://www.phonetique.ulaval.ca/illust.html, Laboratoire de Phonétique et Phonologie, Université Laval, Québec, website lastcheck on 22/01/2013.

[Lav94] John LAVER : *Principles of phonetics*. Cambridge, 1994.

[LB09] Sylvain LE BEUX : *Contrôle gestuel de la prosodie et de la qualité vocale*. Thèse de doctorat, 2009.

[LBFd11] Sylvain LE BEUX, Lionel FEUGÈRE et Christophe D'ALESSANDRO : Chorus digitalis : experiment in chironomic choir singing. *In* Proceedings of the conference ISSN : 1990-9772, éditeur : *12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*, pages 2005–2008, Firenze, Italy, 27/08 au 31/08 2011.

[LC04] Jean-Louis LEFEBVRE et Dominique CHEVALIER : Cancer de l'hypopharynx. *EMC-Oto-rhino-laryngologie*, 1:274–289, 2004.

[LDB99] Jean-Sylvain LIÉNARD et Maria-Gabriella DI BENEDETTO : Effect of vocal effort on spectral properties of vowels. *J. Acoust. Soc. Am.*, 106(1):411–422, July 1999.

[Léo04] Gilles LÉOTHAUD : *Ethnomusicologie générale - Les techniques vocales (chapitre 7)*, 2004.

[Lié77] Jean-Sylvain LIÉNARD : *Les processus de la communication parle*. Masson, Paris, 1977.

[LM60] W. LOTTERMOSER et Fr.-J. MEYER : Fequenzmessungen an gesungenen akkorden. *Akustica*, 10:181–184, 1960.

[Mae79] Shinji MAEDA : An articulatory model based on statistical analysis. *J. Acoust. Soc. Am.*, 65, 1979.

[Max] http ://www.cycling74.com/products/maxmsp.html, website lastcheck on 22/01/2013.

[MC90] Eric MOULINES et Francis CHARPENTIER : Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453–467, 1990.

[MHWL09] Mark T. MARSHALL, Max HARTSHORN, Marcello M. WANDERLEY et Daniel J. LEVITIN : Sensor choice for parameter modulations in digital musical instruments : Empirical evidence from pitch modulation. *Journal of New Music Research*, 38(3):241–253, 2009.

[MMF76] John MORTON, Steve MARCUS et Clive FRANKISH : Perceptual centers (p-centers). *Psychological Review*, 83(5):405–408, September 1976.

[MW06] Mark T. MARSHALL et Marcello M. WANDERLEY : Evaluation of sensors as input devices for computer music interfaces. *In* T. Voinier R. KROLAND-MARTINET et S. YSTAD, éditeurs : *Third International Symposium, CMMR 2005*, pages 130–139, Isa, Italy, September 26-28, 2005 2006. Springer-Verlag Berlin Heidelberg.

[NT96] Jan NORDMARK et Sten TERNSTRÖM : Intonation preferences for major thirds with non-beating ensemble sounds. *TMH-QPSR, KTH*, 37(1):57–62, 1996.

[Oha11] John J. OHALA : Christian gottlieb kratzenstein : pionneer in speech synthesis. *In The 17th International Congress of Phonetic Sciences (ICPhS XVII)*, Hong Kong, 17-21 August 2011.

[Orl90] Robert F. ORLIKOFF : Vowel amplitude variation associated with the heart cycle. *J. Acoust. Soc. Am.*, (88):2091, 1990.

[OW13] Jieun OH et Ge WANG : Lolol : Laugh and out loud on laptop. *In Proceedings of the 13th Conference on New Interfaces for Musical Expression (NIME'13)*, pages 190–195, Daejeon + Seoul, Korea Republic, May 27-30 2013.

[PB52] Gordon E. PETERSON et Harold L. BARNEY : Control methods used in a study of vowels. *J. Acoust. Soc. Am.*, 24(2):175–184, March 1952.

[PB07] Peter Q. PFORDRESHER et Steven BROWN : Poor-pitch singing in the absence of "tone deafness". *Music Percept.*, 25:95–115., 2007.

[PBM+10] Peter Q. PFORDRESHER, Steven BROWN, Kimberly M. MEIER, Michel BELYK et Mario LIOTTI : Imprecise singing is widespread. *J. Acoust. Soc. Am.*, 128(4):2182–2190, October 2010.

[Per82] Pascal PERRIER : *Etude d'un modèle continu des cordes vocales sous forme de deux poutres bi-articules. Premières simulations, Thèse Doct. Ing. INP, Grenoble, 1982.* Thèse de doctorat, Institut National Polytechnique de Grenoble, Grenoble, 1982.

[PF06] Bob PRITCHARD et S. Sidney FELS : Grassp : Gesturally-realized audio, speech and song performance. *In Proceedings of the 6th Conference on New Interfaces for Musical Expression (NIME'06)*, pages 272–276, 2006.

[PFd+11] Robert PRITCHARD, Sidney S. FELS, Nicolas D'ALESSANDRO, Marguerite WITVOET, Johnty WANG, Cameron HASSALL, Helene DAY-FRASER et Meryn CADELL : Performance : What does a body know ? *In Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems (CHI EA'11)*, 2011.

[PI03] Aniruddh D. PATEL et John R. IVERSEN : Acoustic and perceptual comparison of speech and drum sounds in the north indian tabla tradition : An empirical study of sound symbolism. *In Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)*, pages 925–928., Barcelona, 3-9 August 2003.

[Pon79] J. PONTEUS : *Mimmi, en utrustning för konsonantsyntes avsedd att komplettera MUSSE. (Mimmi, an equipment for consonant synthesis intended as a complement for MUSSE).* In swedish., Department of Speech Music Hearing, KTH, 1979.

[Pra94] Eric PRAME : Measurements of the vibrato rate of ten singers. *J. Acoust. Soc. Am.*, 96(4):1979–84, 1994.

[Ram91] Christophe RAMSTEIN : *Analyse, representation et traitement du geste instrumental.* Thèse de doctorat, Institut National Polytechnique de Grenoble, Décembre 1991.

[RHC09] Bernard ROUBEAU, Nathalie HENRICH et Michèle CASTELLENGO : Laryngeal vibratory mechanisms : The notion of vocal register revisited. *Journal of Voice*, 23(4):425–438, July 2009.

[RPB84] Xavier RODET, Yves POTARD et Jean-Baptiste BARRIÈRE : The chant project : From the synthesis of the singing voice to synthesis in general. *Computer Music Journal*, 8(3):15–31, Autumn 1984.

[Sal86] Elliot SALTZMAN : Task dynamic coordination of the speech articulators : A preliminary model. *H. Heuer and C. Fromm (Eds.), Experimental Brain Research*, 15:129–144, 1986.

[SBVB06] Diemo SCHWARZ, Greg BELLER, Bruno VERBRUGGHE et Sam BRITTON : Real-time corpus-based concatenative synthesis with catart. *In Proc. of the 9 th Int. Conference on Digital Audio Effects (DAFx-06)*, Montreal, Canada, September 18-20 2006.

[She64] Roger N. SHEPARD : Circularity in judgements of relative pitch. *J. Acoust. Soc. Am.*, 36(12):2346–2353, December 1964.

[Ste98] Kenneth N. STEVENS : *Acoustic Phonetics.* The MIT Press, 1998.

[Sun69] Johan SUNDBERG : Articulatory differences between spoken and sung vovels in singers. *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR)*, 10(1):033–046, 1969.

[Sun01] Johan SUNDBERG : Level and center frequency of the singer's formant. *Journal of Voice*, 15(2): 176–186, 2001.

[Sun06] Johan SUNDBERG : The kth synthesis of singing. *Advances in cognitive Psychology*, 2(2-3):131–143, 2006.

[TCSW06] Daniel TRUEMAN, Perry R. COOK, Scott SMALLWOOD et Ge WANG : Plork : The princeton laptop orchestra, year 1. *In Proc. of the 2006 International Computer Music Conference (ICMC2006)*, New Orleans, USA, 2006.

[Ter93] Sten TERNSTRÖM : Perceptual evaluations of voice scatter in unison choir sounds. *Journal of Voice*, 7(2):129–135, June 1993.

[Ter03] Sten TERNSTRÖM : Choir acoustics : An overview of scientific research published to date. *International Journal of Research in Choral Singing*, 1(1):3–12, 2003.

[Tit73] Ingo TITZE : The human vocal cords : a mathematical model. *Phonetica*, 28:129–170, 1973.

[Tro39] Nikolaï Sergueïevitch TROUBETZKOY : *Principes de phonologie*. 1939.

[TS88] Sten TERNSTRÖM et Johan SUNDBERG : Intonation precision of choir singers. *J. Acoust. Soc. Am.*, 84(1):59–69, July 1988.

[VUK96] Roel VERTEGAAL, Tamas UNGVARY et Michael KIESLINGER : Towards a musician's cockpit : Transducers, feedback and musical function. *In Proc. of the 1996 International Computer Music Conference (ICMC1996)*, pages 308–311, 1996.

[WD04] Marcello M. WANDERLEY et Philippe DEPALLE : Gestural control of sound synthesis. *Proceedings of the IEEE*, (92):632–644, 2004.

[WVIR00] Marcello M. WANDERLEY, Jean-Philippe VIOLLET, Fabrice ISART et Xavier RODET : On the choice of transducer technologies for specific musical functions. *In Proc. of the 2000 International Computer Music Conference (ICMC2000)*, pages 244–247, 2000.

[YTM+99] Takayoshi YOSHIMURA, Keiichi TOKUDA, Takao MASUKO, Takashi KOBAYASHI et Tadashi KITAMURA : Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis. *IE- ICE Transactions On Information And Systems*, 83(11):2347–2350, 1999.

[ZGS84] Jan ZERA, Jan GAUFFIN et Johan SUNDBERG : Synthesis of selected vcv-syllables in singing. *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR)*, 25(2-3): 119–125, 1984.

[ZTB09] Heiga ZEN, Keiichi TOKUDA et Alan W. BLACK : Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064, 2009.

[ZWMC07] Michael ZBYSZYNSKI, Matthew WRIGHT, Ali MOMENI et Daniel CULLEN : Ten years of tablet musical interfaces at cnmat. *In Proceedings of the 7th Conference on New Interfaces for Musical Expression (NIME'07)*, pages 100–105, New York, USA, 2007.

Lionel FEUGÈRE

SYNTHÈSE PAR RÈGLES DE LA VOIX CHANTÉE CONTRÔLÉE PAR LE GESTE ET APPLICATIONS MUSICALES

Résumé

Le travail de cette thèse porte sur la modélisation de la production et du contrôle de voix chantée synthétique dans la perspective de la lutherie numérique. Nous présentons deux instruments : le Cantor Digitalis, se focalisant sur le contrôle de voyelles chantées et sur l'individualisation des voix ; et le Digitartic, destiné au contrôle de l'articulation de syllabes de type Voyelle-Consonne-Voyelle. Ils permettent, à l'aide de tablettes graphiques augmentées, des applications musicales interactives nécessitant un contrôle temporel fin des paramètres de la production vocale. La pertinence musicale de ces instruments a été établie avec notre ensemble Chorus Digitalis en participant à plusieurs concerts. Nous avons étudié en situation musicale la justesse inter-musiciens et les gestes utilisés pour réaliser les tâches musicales nécessaires à la reproduction d'un large répertoire, constitué de musiques actuelles et traditionnelles (chorale baroque, chant khayal d'Inde du Nord). Notamment, une expérience visant à analyser la faculté à contrôler la fréquence fondamentale du Cantor Digitalis a été entreprise. Les sujets devaient imiter des intervalles et quelques mélodies suivant trois modalités (avec leur propre voix, à la tablette sans et avec retour audio). Les résultats montrent une aptitude plus grande des sujets à jouer de manière précise avec la tablette plutôt qu'avec leur propre voix, tandis que l'apport de l'audio sur le jeu à la tablette est nulle dans ces conditions expérimentales. Les deux instruments sont regroupés dans une application écrite en Max/MSP fournissant également un outil pédagogique audio-visuel et interactif sur le fonctionnement de la voix.

Mots-clefs : synthèse vocale, contrôle gestuel, voix chantée, gestes musicaux, instruments numériques, orchestre numérique.

Abstract

This thesis deals with the production and control modeling of a synthetic singing voice in the context of making a digital musical instrument. Two instruments are presented: the Cantor Digitalis, focusing on singing vowel control and voice individualization, and the Digitartic, which aims at controlling the articulation of Vowel-Consonant-Vowel syllables. Using an augmented graphic tablet, these instruments allow interactive musical applications with fine temporal control of voice production parameters. The relevance of these musical instruments was established through several public performances of the Chorus Digitalis ensemble. The gestures of the musicians were studied along with the musical tasks required for playing the selected repertoire which was composed of traditional world music (baroque choral, North Indian khayal singing) as well as more contemporary pieces. In particular, an experiment was conducted to analyze the ability to control the fundamental frequency of the Cantor Digitalis. Participants were asked to imitate intervals and melodies according to three tempos using three different modalities (one's own voice, tablet, and tablet with audio feedback). Results showed that precision was better with the tablet modalities than with one's own voice, while no significant difference was found between the tablet with and without audio feedback. Both instruments have been unified into one Max/MSP application, which provides an audio-visual and interactive educational tool for understanding voice production.

Keywords: voice synthesis, gestural control, singing voice, musical gestures, digital musical instrument, digital orchestra.