DIGITARTIC : SYNTHÈSE GESTUELLE DE SYLLABES CHANTÉES

Lionel Feugère
LIMSI-CNRS et UPMC
lionel.feugere@limsi.fr

Christophe d'Alessandro
LIMSI-CNRS
cda@limsi.fr

RÉSUMÉ

Nous présentons le Digitartic, un instrument musical de synthèse vocale permettant le contrôle gestuel de l'articulation Voyelle-Consonne-Voyelle. Digitartic est situé dans la continuité de l'instrument de voyelles chantées synthétiques Cantor Digitalis, utilisant la synthèse par formants et développé dans l'environnement Max/MSP. Les analogies entre geste percussif et geste de constriction lors de la production de consonnes sont développées. Digitartic permet le contrôle temps réel de l'instant articulatoire, de l'évolution temporelle des formants, des bruits d'occlusion et de l'aspiration. Le lieu d'articulation peut varier continument par interpolation des lieux d'articulation des consonnes de référence. On discute du type de modèle de contrôle à utiliser suivant l'application recherchée, en s'appuyant sur des analogies gestuelles et des contraintes de temps réel. Un modèle de synthèse de l'articulation est présenté, utilisant les possibilités de contrôle d'une tablette graphique. Des exemples de syllabes synthétiques démontrent que le concept de contrôle gestuel de l'articulation, par analogie à la percussion, est valide.

1. INTRODUCTION

1.1. Propos de ce travail

Les voyelles de la parole correspondent à des conformations stables du conduit vocal, sur des durées qui peuvent atteindre plusieurs secondes en voix chantée (voire beaucoup plus dans certains styles de chant). Au contraire, les consonnes correspondent à des sons transitoires, relativement brefs, associés à des gestes de constriction totale ou partielle dans le conduit vocal. Les gestes consonantiques présentent ainsi des analogies avec les gestes d'attaque des sons musicaux, et en particulier les gestes de percussion digitale, ou de jeu des claviers manuels.

Dans la continuité de nos travaux sur le contrôle gestuel de la synthèse vocale, nous explorons dans cet article les analogies entre percussion manuelle et production de consonnes chantées. Par exemple, la trajectoire du geste de frappe sur une percussion est analogue à celle du geste de constriction des articulateurs sur eux-même. Le lieu et le mode d'articulation des consonnes ressemblent au lieu et au mode des frappes par exemple sur une percussion.

Ces analogies sont d'ailleurs utilisées depuis l'antiquité dans l'apprentissage de certains styles musicaux. Plus récemment, en Inde par exemple, les percussionnistes utilise un lexique de syllabes pour apprendre, mémoriser, transmettre et reproduire des séquences rythmiques, sur des instruments comme les tablas qui offrent plusieurs modes et lieux de frappe.

Une consonne n'est jamais produite seule, l'unité minimale de production de la parole étant la syllabe. La production de syllabes est un geste complexe, mettant en jeu la coordination des différents articulateurs de l'appareil vocal. Sur des durées brèves (de l'ordre d'une dizaine de millisecondes), les articulateurs (lèvres, langue, mâchoires ou luette) et la source vocale doivent se synchroniser, afin de changer dynamiquement la configuration du conduit vocal et de filtrer l'onde acoustique issue de la source glottique. La synchronisation entre les articulateurs et la vibration des plis vocaux permet de contrôler le voisement des sons produits.

Parmi les approches possibles pour synthétiser de la parole, c'est la synthèse à formant que nous avons explorée, plutôt que la synthèse par échantillonnage. Le modèle de contrôle des systèmes utilisant la synthèse par formants est constitué de règles pour la formation des unités phonologiques et des phrases ([15] pour une revue de littérature des systèmes de synthèse par formants). Nous proposons de transférer une partie de ces règles au niveau du geste de l'utilisateur. Toute la dynamique de la synthèse se retrouve alors dans le geste : la durée et l'évolution temporelle des transitions entre positions articulatoires, ainsi que le passage d'une syllabe à une autre sont contrôlés par l'utilisateur. Ce transfert du contrôle, en remplaçant des règles fixes par un geste humain, rapproche le mouvement des paramètres de synthèse du mouvement des articulateurs naturels. Les gestes manuels imitent les gestes articulatoires.

Il n'est pas proposé de produire toutes les unités phonologiques d'une langue. Il faut pour le moment restreindre la combinatoire des phonèmes. Le but n'est pas la synthèse d'un texte quelconque, mais la synthèse réaliste de syllabes, dans une perspective musicale. L'ensemble du travail est développé dans l'environnement Max/MSP [20].

1.2. Travaux antérieurs

La synthèse temps réel de l'articulation, c'est à dire le contrôle gestuel de l'articulation, a commencé avant l'arrivée de l'électricité, en 1791 avec la machine mécanique de Von Kempelen qui pouvait émettre une vingtaine de sons différents par un contrôle manuel reproduisant l'action des articulateurs [11]. Avec l'arrivée de l'élec-

tricité, 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" [18]. En 1939, le premier synthétiseur capable de produire des phrases entières fût le VODER de Homer Dudley, modification du Vocoder mais avec des commandes manuelles [10]. En 1998, Fels et al. publient le GloveTalkII qui relie des mouvements de mains aux paramètres de contrôle d'un synthétiseur à formants, mouvements reconnus à l'aide de réseaux neuronaux. Les deux premiers formants des voyelles sont contrôlés par la position de la main gauche, les consonnes sont déclenchées par la main droite, exceptées les occlusives qui le sont par une pédale [12]. En 2000, Cook et al. présentent une interface de contrôle basée sur un accordéon [4] utilisant le système SPASM basé sur des guides d'onde numérique. Le souffle est contrôlé par le soufflet de l'accordéon, la hauteur tempérée par le clavier de l'accordéon. Une série de boutons a été ajoutée pour contrôler les voyelles et les consonnes [3]. D'Alessandro et al. mentionnent dans la publication du Handsketch l'utilisation de capteurs FSR pour déclencher des syllabes mais cet aspect n'est pas implémenté [6]. En 2011, Beller et al. utilisent des mouvements percussifs captés par un accéléromètre pour déclencher des séquences de voix préenregistrées [2]. Enfin, Astrinaki et al. ont présenté en 2011 un prototype de système temps réel de synthèse à partir de texte utilisant des modèles de Markov cachés (HMMs). Les HMMs modélisent les dépendances entre les phonèmes à partir d'une base de données de voix naturelle. Ces phonèmes sont concaténés et la modification de leur spectre, de l'intonation et des durées est réalisée à partir des HMMs [1].

La synthèse par formants permet facilement l'interpolation de valeurs de référence et offre ainsi un espace continu de positions articulatoires où l'on peut s'y déplacer sans restriction. D'où l'intérêt d'un contrôle gestuel du séquencement par règles et non par échantillons. Le principal inconvénient de la synthèse par formant est sa qualité sonore à priori plus faible que celle de séquences enregistrée. Mais nous pensons que les possibilités accrues de contrôle dynamique pallient le manque de qualité statique des sons.

Dans la suite de cet article, nous commencerons par présenter le modèle de production de l'instrument, de type source-filtre avec ses règles de trajectoires formantiques, de bruit d'occlusion et d'aspiration suivant les instants articulatoires, les lieux et modes d'articulation et la force vocale. Puis nous traiterons des modèles de contrôle de ce synthétiseur, de leur intérêt face aux systèmes existants, en discutant des gestes et interfaces appropriés pour le contrôle de l'articulation, et en comparant les trajectoires formantiques de la voix de synthèse et de la voix naturelle.

2. LE MODÈLE DE PRODUCTION DE DIGITARTIC

2.1. Le modèle de source CALM

Le modèle de source glottique utilisé dans ces travaux est le RT-CALM [5], variante temps réel du modèle CALM [8]. Ce modèle travaille dans le domaine spectral et s'appuie sur une analyse des principaux modèles temporels proposés dans la littérature [9].

Les propriétés spectrales de la source sont décrites par la forme de la dérivée de son spectre : un maximum en basses fréquences, le "formant glottique" et la pente spectrale pour les fréquences médium et aigües. L'onde de débit glottique (ODG) est souvent représentée par sa dérivée en prenant en compte le filtre passe-haut associé au rayonnement aux lèvres. Notons que le "formant glottique" ne correspond pas à un formant résonantiel mais à la forme de l'ODG. La force de voix est modélisée par une augmentation de l'intensité du signal et par la diminution de la pente spectrale de l'ODG dérivée.

2.2. Règles pour les transitions articulatoires

On se limite ici à la synthèse de syllabes de type VCV où V est une voyelle (/a/ pour les exemples) et C est soit une occlusive (/p,t,k/ comme référence), soit une semi-voyelle (/w,ɥ,j/ comme référence). Dans notre système, une transition articulatoire est définie par l'évolution temporelle des paramètres suivants :

- les valeurs des 4 premiers filtres formantiques (fréquence centrale, bande passante, amplitude)
- le bruit d'aspiration, modélisé par un bruit blanc modulé par la forme de l'ODG.
- les coefficients de filtres modifiant un bruit blanc pour le bruit occlusif
- la force vocale

2.2.1. Transitions formantiques et d'aspiration

On utilise les valeurs cibles des consonnes et de la voyelle /a/ qu'on interpole pour obtenir des positions articulatoires intermédiaires à un instant temporel défini par le contrôle gestuel. On choisit une interpolation linéaire pour la correspondance entre les valeurs des formants et le paramètre de contrôle de l'instant articulatoire, et on laisse l'utilisateur modifier cette linéarité par la dynamique de son geste associé au paramètre de contrôle de l'instant articulatoire.

2.2.2. Bruits d'occlusion

On modélise les bruits d'occlusion à l'aide d'un bruit blanc filtré par une série de filtres en cascade, dont on détermine la forme à l'aide du spectre de bruit de voix réelles.

Afin de modéliser l'influence de la voyelle, on fait l'hypothèse que le bruit consonantique sera d'autant plus marqué par les résonances formantiques que son lieu de constriction est postérieur. Le bruit est filtré par les filtres for-

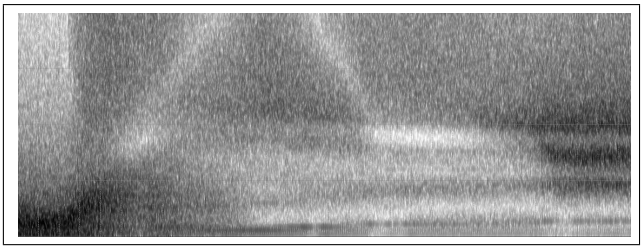

Figure 1. Sonogramme ($0 - 8000$ Hz) du bruit d'occlusion produit par le Digitartic, avec le lieu d'articulation qui évolue le long de l'axe bilabial - alvéo-dental - palatal (de gauche à droite)

mantiques correspondant à la voyelle, avec une bande passante d'autant plus étroite que le lieu d'occlusion est postérieur. Par exemple, un bruit labial est peu modifié car la source du bruit est situé à l'extrémité du conduit vocal, et un bruit provenant de la glotte est filtré comme l'est une voyelle. Cette caractéristique est illustrée par le sonogramme de la figure 1 où l'on remarque l'apparition progressive des formants quand le lieu d'origine du bruit devient postérieur.

2.3. Continuité du lieu d'articulation

Le système permet de réaliser des consonnes à des lieux intermédiaires entre deux consonnes de référence d'un même mode d'articulation, et également de contrôler l'instant précis d'articulation. Le paramètre correspondant au lieu d'articulation étant continu, on peut alors produire une infinité de pseudo-consonnes (consonnes qui n'entrent pas dans le système phonologique de la langue visée).

A partir des valeurs des fréquence / amplitude / bande-passante de leurs formants, et des valeurs des coefficients des filtres des bruits consonantiques, on construit des pseudo-consonnes intermédiaires sur l'axe bilabial - alvéo - dental - palatal. L'hypothèse utilisée revient à supposer qu'on peut interpoler ces valeurs pour obtenir des niveaux d'articulation intermédiaires. C'est peut être le cas en première approximation entre les occlusives alvéo-dentales et palatales, mais plus difficilement concevable entre les occlusives bilabiales et alvéo-dentale, vu la discontinuité entre ces deux lieux d'articulation. Du point de vue de la synthèse de voix pour la musique, cela permet d'obtenir des sons vocalement plausibles, bien que difficilement prononçable en réalité. Cependant, la perception des consonnes étant catégorielle (on associe la consonne entendue au plus proche phonème de notre langue), la perception qu'on a de cette continuité consonantique est discrète en ce qui concerne l'identification du son. La continuité des consonnes s'exprime par un changement progressif de la qualité perçue de la consonne, mais en général pas de son identification.

La figure 2 présente une série d'occlusives de l'instrument (suivies chacune de la voyelle /a/) pour lesquelles le lieu d'articulation évolue progressivement de bilabial à palatal. La configuration passe donc par les syllabes /pa/, /ta/ et /ka/. De la même manière, la figure 3 est une série de semi-voyelles pour lesquelles le lieu d'articulation varie, en passant donc par /wa/, /ɥa/ et /ja/.

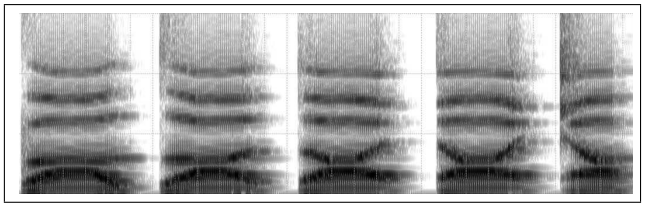

Figure 2. Sonogramme ($0 - 6000$ Hz sur 3 secondes) de "Occlusive-/a/" successifs produits par le Digitartic, avec le lieu d'articulation de l'occlusive qui évolue sur l'axe bilabial - alvéo-dental - palatal

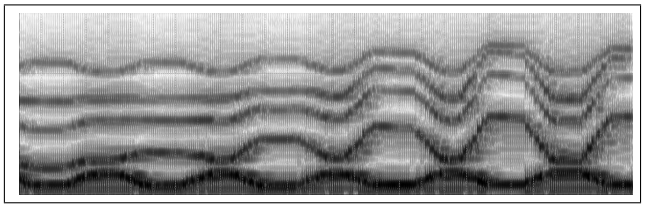

Figure 3. Sonogramme ($0 - 6000$ Hz sur 3 secondes) de "SemiVoyelle-/a/" successifs produits par le Digitartic, avec le lieu d'articulation de l'occlusive qui évolue sur l'axe bilabial - alvéo-dental - palatal

2.4. Continuité du mode d'articulation

La synthèse est dans cet article restreinte à deux modes d'articulation : les semi-voyelles et les occlusives. La différence fondamentale dans la manière de produire les consonnes de ces deux modes d'articulation est l'ajout d'un bruit blanc filtré au système source-filtre. Le reste des différences portent sur des règles distinctes, mais des processus identiques (changement des formants, du bruit d'aspiration, de la force vocale), et sur les gestes de l'utilisateur.

Dans le présent système, on considère les modes d'articulation comme discrets, c'est à dire qu'on passe des occlusives aux semi-voyelles sans intermédiaire possible. Cependant, les occlusives d'un lieu d'articulation donné présentant à peu près les mêmes valeurs formantiques que les semi-voyelles du même lieu d'articulation, on peut envisager facilement de créer une continuité entre ces deux modes d'articulation à l'aide de règles sur l'intensité des bruits (d'inexistant pour les semi-voyelles à fort pour les occlusives), de la force de voix (pleine pour les semi-voyelles tenues et nulle pour les occlusives sourdes), et de gestes de vitesse différente.

3. LE MODÈLE DE CONTRÔLE DE DIGITARTIC

On cherche à externaliser les mouvements internes de l'appareil vocal (larynx, uvule, langue, mâchoire, lèvres) dans leur ensemble par des gestes manuels. Il faut donc rechercher des interfaces qui permettent des gestes adaptés au contrôle articulatoire. Le modèle de contrôle est ce qui

permet de relier les paramètres du modèle de production à l'interface de contrôle.

3.1. Quelle interface pour quelle application ?

La multiplicité des paramètres à contrôler dans un synthétiseur vocal implique que pour un résultat optimal, on se doit de choisir une interface vraiment adaptée à l'application recherchée. Pour une approche musicale, on peut dans un premier temps ne pas chercher à reproduire tous les phonèmes de la langue étudiée. Les expériences passées ont montré que les instruments de synthèse de voix parlée sont très difficile à contrôler. Les opérateurs du VODER [10] devaient être entrainés pendant au moins un an avant de pouvoir synthétiser des phrases en public. Plus récemment, le Glove-talkII [12], 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 [19] [13].

Parmi les systèmes cités en introduction, seul le Glove-TalkII permet de contrôler l'instant d'articulation et de choisir a priori des lieux d'articulation autres que ceux de la langue choisie pour leur système. Mais dans une perspective musicale nécessitant une haute précision temporelle, 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 son récent système de synthèse temps réel destiné à la communication pour personnes muettes [16].

Dans Cantor Digitalis [14], le contrôle précis porte sur la mélodie (F_0). La tablette graphique munie de son stylet, est contrôlée par un geste proche de l'écriture. Or l'écriture requiert une haute précision spatiale : la tablette graphique est très bien adaptée au contrôle de F_0 pour la musique, qui nécessite elle aussi une haute précision de l'ordre de 4 centièmes de demi-tons.

Ici, on s'intéresse au contrôle des syllabes et leurs articulations, comme par exemple les récitations onomatopéiques des percussions indiennes. Le contrôle de F_0 est alors réduit à un contrôle ne nécessitant 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 au moins inclure les contrôles suivants de :

- F_0 et du lieu d'articulation (continus et précis)
- l'instant articulatoire (continu et rapide)
- la force vocale (continu)
- le mode d'articulation (discret)
- les voyelles (continu ou discret)

La tablette graphique permet un nombre important de contrôles (pression, position X/Y, orientations du stylet) mais le stylet ne doit pas quitter la tablette. L'analogie entre geste percussif et geste articulatoire nous encourage à contrôler l'articulation à l'aide de mouvements verticaux de type frappe. Les mouvements de la main ou des doigts sont facilement mesurables à l'aide d'accéléromètres. Des tests de mapping sont en cours.

3.2. Geste percussif et geste articulatoire

Le geste articulatoire et le geste percussif présentent des analogies assez fortes entre :

- le lieu d'articulation et le lieu de frappe
- le mode d'articulation et la manière de frapper une percussion
- le caractère voisé d'une consonne et le caractère ouvert/fermé d'une frappe (i.e main laissée ou non en contact avec la peau après la frappe pour éviter sa vibration)
- les mouvements de coarticulation et les mouvements entre deux frappes successives

En revanche, nous n'affirmons pas qu'il existe une ressemblance psycho-motrice stricte des deux types de gestes.

A l'aide d'une interface multitouch trackpad, on peut facilement modéliser une peau de percussion et faire correspondre les analogies présentées plus haut. Le problème est le contrôle de l'instant articulatoire qui se fait par analogie avec la position verticale de la main qui frappe la peau. Or, le trackpad ne permet pas de capter le geste vertical. Pour remédier à cela, un modèle de contrôle a été établi à l'aide d'un trackpad qui déclenche la transition VC lors de la pose du doigt et déclenche la transition CV lors du retrait du doigt. Le contrôle sur l'instant articulatoire est donc réduit au déclenchement des différentes transitions, mais sans correspondance continue avec le geste. Ici, le trackpad contrôle les aspects consonantiques (lieu et mode d'articulation, déclenchement des phases d'articulation) avec la main secondaires et une tablette graphique permet de contrôler les aspects vocaliques (mélodie, force vocale) avec la main préférée.

Dans l'idéal, nous aimerions avoir une captation continue de la dynamique du doigt pour le faire correspondre en temps réel avec l'instant articulatoire. On peut s'interroger alors sur les correspondances entre phase articulatoire (VC versus CV) et phase de geste percussif (remontée et descente du geste). Est-il préférable de faire correspondre la phase de frappe (descente) du doigt avec la transition CV et la phase de remontée du doigt avec la transition VC ou bien le contraire ? Cela nous amène à devoir introduire dans la discussion le temps réel et la notion d'attaque en musique.

3.3. Temps réel et contrôle articulatoire

Un instrument de synthèse vocale qui ne contrôlerait que le déclenchement des phases articulatoires et non le contrôle de l'instant articulatoire sont voués à ne pas pouvoir être joués dans des musiques nécessitant une haute précision temporelle, comparable au seuil de perception de la non simultanéité de deux évènements, 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) 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). De plus, à ce retard lié au modèle, s'ajoute le retard de la récupération des données de l'interface. En percussion, l'attaque se situe au contact de la main/doigt sur la peau. Dans l'idéal, il faudrait donc relier le mouvement descendant du geste manuel au resserrement des articulateurs pour les occlusives et au relâchement des articulateurs pour les semi-voyelles (de même pour les fricatives).

L'intérêt du trackpad vis à vis de la tablette est sa capacité à capter plusieurs doigts simultanément. Mais divers problèmes ont été rencontrés avec le trackpad qui nous ont amené à ne pas l'utiliser pour le profit de la tablette graphique seule :

- latence d'autant plus grande que le nombre de doigts en contact est important, allant jusqu'à des latences perceptibles de l'ordre de 100 ms ;
- bugs dans l'external Max/MSP qui récupère les données du trackpad ("fingerpinger" [1])

Ainsi, pour le moment, le geste percussif se fait dans le plan horizontal d'une tablette graphique. Capter le mouvement vertical de la main ou du doigt en temps réel n'est pas sans poser de problèmes techniques. Les mouvements sont d'amplitude assez faibles (moins d'une dizaine de 10 cm) et la résolution doit être suffisamment importante pour que le pas d'échantillonnage ne soit pas audible. Nous avons pensé à des systèmes de captation vidéo ou de gants haptiques, mais le temps de calcul ou la précision est mauvaise (jusqu'à plusieurs dizaines de millisecondes). La meilleure solution trouvée, à la fois sur la facilité d'utilisation et de la rapidité de la réponse, seraient des accéléromètres placées sur les doigts. La vitesse est récupérable par intégration. La position par double intégration fournit trop d'erreurs de calcul. Il faudrait donc faire correspondre la vitesse à l'instant articulatoire.

4. RÉSULTATS

Un schéma général de l'instrument utilisant une tablette graphique avec la deuxième configuration (explicitée ci-après) est donné à la figure 4. La première configuration est définit par les correspondances suivantes :

- pression du stylet sur la tablette $\longleftrightarrow$ force vocale
- position du stylet sur l'axe Y $\longleftrightarrow$ lieu d'articulation
- angle entre stylet et axe X $\longleftrightarrow$ instant articulatoire
- position du stylet sur l'axe X $\longleftrightarrow$ F_0

Dans la deuxième configuration, les contrôles du lieu d'articulation et de l'instant articulatoire sont inversés afin de pouvoir profiter de la bonne résolution spatiale et temporelle de la captation de la position du stylet :

- angle entre stylet et axe X $\longleftrightarrow$ F_0
- position du stylet sur l'axe X $\longleftrightarrow$ instant articulatoire

Alors que la première configuration permet de chanter à une hauteur mélodique précise, la production de syllabe est facilité et de meilleure qualité avec la deuxième configuration car la résolution de la position du stylet est plus

1 . www.anyma.ch/2009/research/multitouch-external-for-maxmsp/, consultée le 13 avril 2012

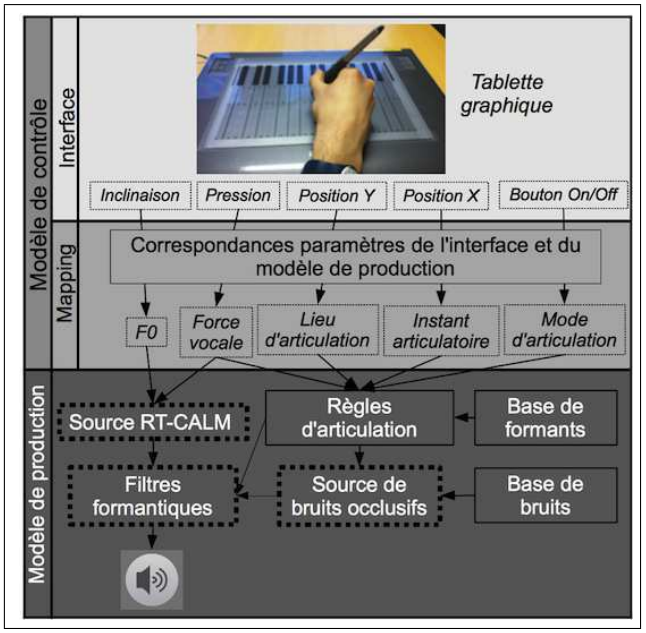

Figure 4. Shéma de fonctionnement général du Digitartic

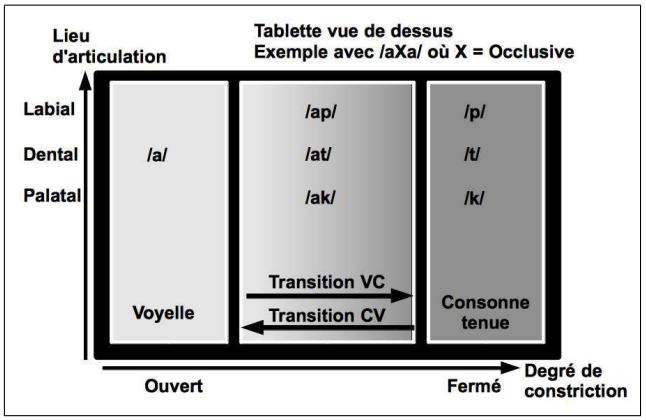

Figure 5. Tablette vue de dessus avec son mapping spatial pour le contrôle de l'instant articulatoire suivant l'axe X et du lieu d'articulation suivant l'axe Y

grande que celle de l'angle entre le stylet et l'axe X, et que le geste est plus simple (déplacement rectiligne de la main versus torsion du poignée tout en maintenant le stylet en contact fixe). Le contrôle de l'instant articulatoire et du lieu d'articulation dans l'espace 2D de la tablette, comme il est défini dans le deuxième configuration, est illustré par la figure 5.

Nous comparons ci-dessous quelques séquences VCV produites via la deuxième configuration avec les mêmes séquences VCV produite par une voix naturelle (figure 6 pour les semi-voyelles et figure 7 pour les occlusives).

Ces images appellent les remarques suivantes :

- la dynamique des trajectoires est bien reproduite
- les amplitudes des formants F_3 à F_5 des consonnes sont de façon générale trop élevées par rapport à ce qu'on peut observer en voix naturelle (figure 6).
- l'aspiration sur la transition VC dure trop longtemps pour les occlusives synthétiques comparées à la voix

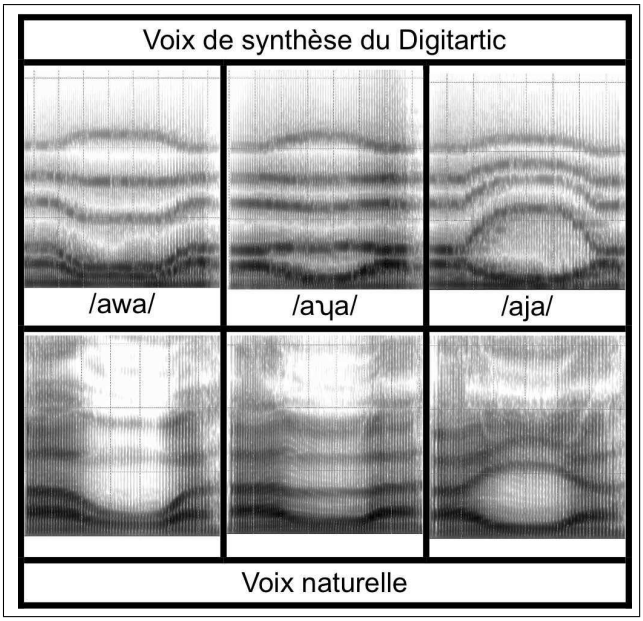

Figure 6. Sonogrammes (0-6000 Hz) des semi-voyelles de Digitartic et de voix naturelle

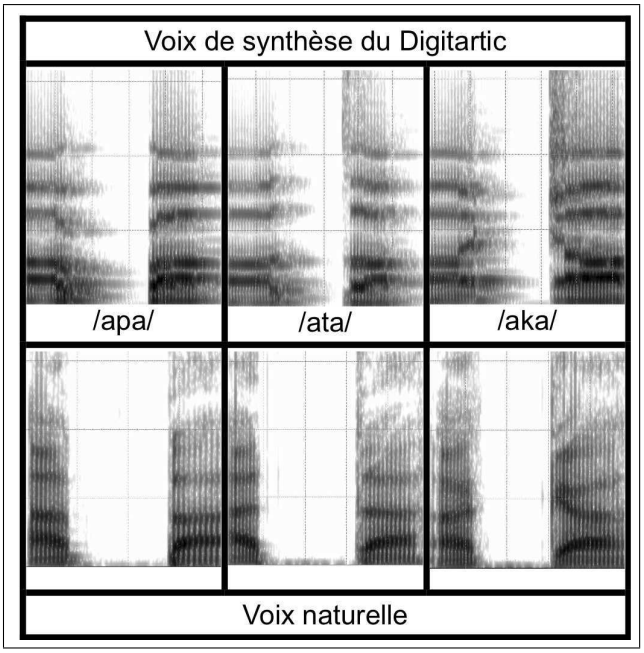

Figure 7. Sonogrammes (0-6000 Hz) des occlusives de Digitartic et de voix naturelle

naturelle (figure 7).

– les transitions avec le Digitartic présentent de petites discontinuités contrairement à la voix naturelle visible en zoomant sur les figures 6 et 7. Les raisons peuvent être : le geste rapide qui est exigé et la résolution spatiale et/ou temporelle limitée ne permettent peut être pas d'avoir suffisamment de données reçues de la tablette pour accomplir une trajectoires consonantiques bien continue puisqu'on a un contrôle directe sur la trajectoire de l'ensemble des formants ; Digitartic possède un module qui atténue en temps réel l'amplitudes des formants quand les harmoniques de F_0 et les filtres formantiques coïncident [14], pouvant créer des variations d'amplitudes très brèves quand les formants évoluent rapidement.

5. PERSPECTIVES

Le système Digitartic permet de démontrer qu'il est possible de contrôler finement et précisément par le geste manuel des transitions consonantiques convaincantes. L'analogie entre geste de constriction dans l'appareil vocal et geste de percussion manuelle semble donc prometteuse.

Nous avons pour ambition d'élargir les possibilités articulatoires du Digitartic aux autres modes d'articulation, comme les fricatives et les nasales. Le point de mire est toujours la performance temps réel et nous sommes en train de tester l'utilisation d'accéléromètres placés sur les doigts comme interface pour le contrôle de l'instant articulatoire et de l'intensité consonantique.

La synthèse d'articulation est introduite progressivement dans notre chorale, Chorus Digitalis [14] [17], un groupe musical qui se réunit de façon hebdomadaire pour faire de la musique à l'aide de synthèse vocale à contrôle gestuel. Quelques consonnes enrichissent déjà la palette sonore des chanteuses et chanteurs virtuels, ainsi que plusieurs modules pour améliorer la qualité de la synthèse par formants, dont des interactions source-filtre, et la modélisation de modulations naturelles de la source.

Enfin, notre instrument devra être validé par des expériences perceptives pour évaluer la qualité des consonnes produites et d'autres part par des expériences plus quantitatives pour mesurer les capacités de notre système à reproduire les transitions articulatoires.

6. REMERCIEMENTS

Les auteurs tiennent à remercier les relecteurs pour leurs bons conseils.

Ce travail est mené dans le cadre du projet Européen FEDER OrJo [2] dans lequel le LIMSI fournit des instruments vocaux pour le logicel Méta-Malette [7] qui permet de jouer des instruments virtuels audio-visuels à plusieurs.

7. REFERENCES

[1] Astrinaki, M., Babacan, O., d'Alessandro, N., Dutoit, T., *sHTS : a streaming architecture for statistical parametric speech synthesis*, International Workshop on Performative Speech and Singing Synthesis, Vancouver, BC, CA, March 14-15, 2011.

[2] Beller, G., *Gestural Control of Real-Time Concatenative Synthesis in Luna Park*, P3S 2011, International Workshop on Performative Speech and Singing Synthesis, Vancouver, BC, CA, March 14-15, 2011.

2 . Voir le site du projet OrJo http ://pucemuse.com/orjo

[3] Cook., P. R., *Identification of control parameters in an articulatory vocal tract model, with applications to the synthesis of singing*, PhD thesis, Stanford University, 1991.

[4] Cook, P., Leider, C., *SqueezeVox : A New Controller for Vocal Synthesis Models*, Proceedings of the ICMC (International Computer Music Conference), Berlin, 2000.

[5] D'Alessandro, N., d'Alessandro, C., Le Beux, S., Doval, B. *Real-time CALM Synthesizer New Approaches in Hands-Controlled Voice Synthesis*, Proceedings of the 2006 International Conference on New Interfaces for Musical Expression (NIME06), Paris, France, pp. 266-271, 2006.

[6] D'Alessandro, N., Dutoit, T. *HandSketch Bi-Manual Controller, Investigation on Expressive Control Issues of an Augmented Tablet* , Proceedings of the 2007 Conference on New Interfaces for Musical Expression (NIME07), New York, NY, USA, pp. 78-81, 2007.

[7] De Laubier, S., Goudard., V., *Puce Muse - La Méta-Mallette*, Journée d'Informatique Musicale, 2008.

[8] Doval, B., d'Alessandro, C., Henrich, N., *The voice source as a causal / anticausal linear filter*. In ISCA, editor, Proceedings of Voqual'03, *Voice Quality : Functions, analysis and synthesis*, Geneva, Switzerland, 2003.

[9] Doval, B., d'Alessandro, C., Henrich, N. , *The spectrum of glottal flow models*. Acta Acustica, 92 :1026–1046, 2006.

[10] Dudley, H. *Remaking speech*, The Journal of the Acoustical Society of America, 11(2) :169-177, 1939.

[11] Dudley, H., Tarnoczy, T. H., *The speaking machine of Wolfgang von Kempelen*, J. Acoust. Soc. Am., vol. 22, no. 2, pp. 151-166, 1950.

[12] Fels, S., Hinton, G., *Glove-TalkII : A neural network interface which maps gestures to parallel formant speech synthesizer controls*, IEEE Transactions on Neural Networks, pp. 205–212, Vol 9, No. 1, 1998.

[13] Fels, S., Pritchard, R., Lenters, A., *ForTouch : A Wearable Digital Ventriloquized Actor*, Proceedings of the International Conference on New Interfaces for Musical Expression, pp. 274–275, 2009.

[14] Feugère, L., Le Beux, S., d'Alessandro, C., *Chorus Digitalis : Polyphonic gestural singing*. P3S 2011, International Workshop on Performative Speech and Singing Synthesis, Vancouver, BC, CA, March 14-15, 2011

[15] Klatt, D. H., *Review of text-to-speech conversion for English*, J. Acoust. Soc. Am., Volume 82, Issue 3, pp. 737-793, 1987

[16] Kunikoshi, A., Qiao, Y., Saito, D., Minematsu, N., Hirose, K., *Gesture Design of Hand-to-Speech Converter Derived from Speech-to-Hand Converter Based on Probabilistic Integration Model*, Proceedings of Interspeech, pp. 3025-3028, 2011.

[17] Le Beux, S., Feugère, L., d'Alessandro, C., *Chorus Digitalis : experiments in chironomic choir singing*. Proceedings of Interspeech 2011, 28-31, Florence, Italie, August 2011.

[18] Lienard, J.-S., *Les processus de la communication parlée*, Masson, Paris, 1977.

[19] Pritchard, B., Fels, S., *GRASSP : Gesturally-Realized Audio, Speech and Song Performance*, Proceedings of the International Conference on New Interfaces for Musical Expression, pp. 272–276, 2006.

[20] Puckette, M., Zicarelli, D., *Max/MSP*, Cycling 74/IRCAM, version 5.1, 1990-2010.