Contrôle gestuel de la synthèse vocale

Les instruments Cantor Digitalis et Digitartic

Lionel Feugère [1,2], **Christophe d'Alessandro** [1]

1. *LIMSI, CNRS, Université Paris-Saclay,*
   Bât 508, rue John von Neumann, Campus Universitaire, F-91405 Orsay
   lionel.feugere@limsi.fr,cda@limsi.fr

2. *Sorbonne Universités, UPMC Univ Paris 06*
   UFR d'Ingéniérie, 4 place Jussieu, 75252 Paris cedex 05
   lionel.feugere@limsi.fr

RÉSUMÉ. *Deux instruments de synthèse vocale sont présentés : le Cantor Digitalis et le Digitartic. Ces deux instruments utilisent des gestes bimanuels dérivés de l'écriture ou du dessin sur une tablette graphique. Le signal est calculé par un synthétiseur vocal paramétrique, comprenant des modèles de source voisée et de bruits consonantiques et une structure à formants série/parallèle, pour le conduit vocal. Le Cantor Digitalis permet de chanter des voyelles et des semi-voyelles. Le Digitartic permet de chanter des syllabes, avec des consonnes plosives, fricatives, liquides et nasales. La question de la synchronisation des gestes consonantiques et des appuis rythmiques demandés par la musique est discutée. Ces instruments permettent un jeu musical expressif et sont régulièrement utilisés en concert.*

ABSTRACT. *Two singing synthesis instruments are presented: Digitalis Cantor and Digitartic. Both instruments use bimanual writing or drawing gestures on graphic tablets. The voice signal is computed with the help of a parametric synthesizer, including a voice source model, consonantal noise models and series/parallel formant filters. Cantor Digitalis is a vowel and semi-vowel singing instrument. Digitartic allows for singing syllables, including plosives, fricative, liquid and nasal consonants. The issue of consonant gestures and musical beat synchronization is discussed. These instruments allow for expressive musical performances. They are regularly used for concerts.*

MOTS-CLÉS : *synthèse vocale, synthèse syllabique, instrument de musique numérique, contrôle gestuel, contrôle de l'articulation, contrôle rythmique.*

KEYWORDS: *voice synthesis, syllable synthesis, digital musical instrument, gestural control, articulation control, rythmic control.*

DOI:10.3166/TS.32.417-442

Extended abstract

Gestural control of speech or singing synthesis is difficult, because of the very fast articulators motions encountered in speech and singing. For singing, another difficult question is accurate rhythmic coordination and precision, because syllables must coincide with musical beats, at a given tempo. Then, the precise location of beats for different syllables must be controlled.

Two singing synthesis instruments are presented: Digitalis Cantor and Digitartic. Both instruments use bimanual writing or drawing gestures on graphic tablets. The voice signal is computed with the help of a parametric synthesizer, including a voice source model, consonantal noise models and series/parallel formant filters. Cantor Digitalis is a vowel and semi-vowel singing instrument. Digitartic is an extension of Cantor Digitalis and allows for singing syllables, including plosives, fricative, liquid and nasal consonants. Any in-between canonical place of articulation is possible by linear interpolation of the consonant parameters, for each mode of articulation.

In this paper, the focus is given on Digitartic through the issue of consonant gestures and musical beat synchronization. Three modes of Vowel-Consonant-Vowel (VCV) articulation are discussed according to three levels of rhythmic precision and musical context. A VCV articulation is composed of the onset phase (articulators approaching the position of maximum constriction), the medial phase (maximum of constriction) and the offset phase (constriction release). Offset phase of plosives is very short compared to other consonants.

The first control mode consists of triggering the syllable at the beginning of the onset phase. However, when the syllable starts on the musical beat, it is perceived with a delay depending on the duration of articulation phases. Anticipating precisely this delay is very difficult. The second control mode of control allows for triggering the VCV dissyllable in two steps. In the first step, the onset phase is triggered, and in the second step, the offset phase is triggered. In this way, plosives can be synchronized with musical beats without any delay. The third control mode is a continuous control of the phases of articulation, without any triggering. This requires a fast synthesis engine, a high interface sampling rate, as well as an expert control gesture, fast and precise enough to reproduce speech articulation phases.

The continuous control mode of articulation is performed by a back-and-forth gesture with the pen of the non-preferred hand, along the vertical dimension of the graphic tablet. Place of articulation is continuously controlled along the horizontal dimension, and the mode of articulation is assigned to different areas on the tablet. This back-and-forth gesture is analog to the somewhat symmetric articulation of the VCV dissyllable. The gesture amplitude allows for different degrees of articulation (hypoarticulation to hyperarticulation). Controlling durations of each phase of articulation is another mean to increase expressiveness. The preferred hand is controlling pitch, vocal effort and vowel quality on another graphic tablet. Then it is possible to modify pitch and vocal effort during each phase of articulation. Cantor Digitalis and Digitartic allow for expressive musical performances. They are regularly used for concerts.

1. Introduction

1.1. Présentation

La voix reste le parent pauvre des instruments de synthèse musicale. Des « chœurs » existent depuis longtemps dans les synthétiseurs, mais l'effet est très éloigné d'une voix soliste ou même chorale de qualité. Cette absence s'explique surtout, à notre avis, pour des raisons de contrôle, à cause de la variabilité et de l'expressivité requises pour la synthèse d'une voix chantée.

Pour traiter cette question, et grâce au développement des nouvelles interfaces humain-machine (IHM) et à celui des environnements pour la synthèse audio temps-réel, plusieurs groupes de recherche ont proposé depuis une douzaine d'années des instruments de synthèse vocale (Wanderley *et al*., 2000 ; Kessous, 2004 ; Zbyszynski *et al*., 2007 ; D'Alessandro *et al*., 2007 ; Le Beux *et al*., 2011 ; Astrinaki *et al*., 2012).

La synthèse vocale « instrumentale » est particulièrement difficile car les mains et l'appareil vocal sont des systèmes musculaires radicalement différents. La rapidité et la précision des articulateurs lors de la phonation permettent une coordination impressionnante des gestes (de l'ordre du millimètre et de la dizaine de millisecondes). Ce sont des gestes experts développés depuis la naissance, et correspondant à la fonction vitale du langage.

Cet article [1] discute du contrôle gestuel de la synthèse vocale en présentant le Cantor Digitalis et le Digitartic, deux instruments de musique régulièrement joués en chœur, dans le cadre de l'ensemble Chorus Digitalis.

Un synthétiseur vocal comprend deux éléments : un ordinateur, avec le moteur de synthèse, et une interface de contrôle (Miranda, Wanderley, 2006 ; Cook, 2005). Dans le cas de l'articulation vocale, un troisième élément particulièrement important est le répertoire de gestes de contrôle. En effet, aucun des gestes instrumentaux connus (utilisant claviers, clefs, touches, cordes, archet, etc.) ne semble bien adapté, car a priori trop éloignés de l'ensemble des gestes articulatoires naturels. La mise en œuvre du synthétiseur demande donc une réflexion approfondie sur le contrôle gestuel de l'articulation en vue du jeu musical, en particulier pour le contrôle rythmique. Nous défendons l'idée que les gestes dérivés de l'écriture ou du dessin, capturés à l'aide de tablettes graphiques sont particulièrement appropriés pour ces tâches. Nous nous concentrons dans cet article sur les réflexions et mises en œuvres des gestes de contrôle. Le lecteur intéressé par les détails d'implémentation est invité à consulter le site http://cantordigitalis.limsi.fr, où il trouvera, entre autres, les programmes sources du Cantor Digitalis accompagnés d'une documentation détaillée.

1. Certains éléments du présent article ont été présentés dans les conférences JIM2012 et NIME2013 (Feugère, d'Alessandro, 2012 ; 2013).

1.2. L'évolution de la synthèse de voix chantée

Historiquement, la voix de synthèse est apparue dans des pièces de musique contemporaine grâce au programme « Chant » (Rodet *et al.*, 1984). « Chant » est un système de synthèse par règles, utilisant un modèle paramétrique source/filtre de production vocale [2]. L'avantage de la synthèse paramétrique est sa souplesse et son faible coût en calcul et en mémoire. Dans la ligne de « Chant », d'autres groupes ont développé des synthétiseurs de voix chantée à formants par règles (Berndtsson, 1995 ; Cook, 1993).

Dans la génération suivante des synthétiseurs vocaux, le son est calculé par la concaténation et la modification d'échantillons de voix enregistrés au préalable. Un exemple marquant de ce type de technique en voix chantée, utilisée en post-production, est la bande son du film « Farinelli » (Depalle *et al.*, 1995) ; une voix de Castrat virtuelle est synthétisée par le mélange et la mise en forme d'une voix d'homme et d'une voix de femme, chantant la même partition. La synthèse automatique de chant à partir du texte est quant à elle commercialisée par Yamaha, avec le système Vocaloid (Kenmochi, Oshita, 2007). Ce système très populaire permet de synthétiser de la voix de haute qualité. Mais sa souplesse étant limitée, il est difficile d'aller au delà des échantillons fournis, et surtout, il n'y a aucun contrôle expressif en temps-réel. Ce n'est donc pas un instrument, mais un synthétiseur de studio personnel. La synthèse paramétrique statistique est la dernière génération de synthétiseurs de parole, et par extension de voix chantée. C'est un domaine très intéressant, puisque l'apprentissage statistique permet d'ajuster automatiquement les paramètres de nouvelles voix. Cependant la synchronisation précise des notes de musique et des syllabes est difficile (Astrinaki *et al.*, 2012) et les résultats actuels en voix chantée sont d'une qualité sonore insuffisante pour des instruments musicaux.

1.3. Interface de contrôle : chironomie

Dans le développement d'un instrument vocal, l'expressivité musicale prime, et donc la question du contrôle gestuel est centrale. Contrôler la synthèse vocale avec les mains reste une tâche difficile, malgré des efforts de recherche soutenus et anciens. Avec le premier système, pour la parole, les opérateurs du VODER (Dudley *et al.*, 1939) étaient soumis à plusieurs mois d'entraînement avant de pouvoir synthétiser quelques phrases en public. Plus récemment, les GloveTalk, des systèmes de synthèse de parole utilisant deux gants de données et une pédale d'expression pour contrôler un synthétiseur à formant par règles (Fels, Hinton, 1992 ; 1998) restent peu intelligibles et très artificiels, même au prix d'un entraînement intense (Pritchard, Fels, 2006 ; Fels *et al.*, 2009). Les gants permettent de piloter plusieurs paramètres continus simultanément. Par contre, il n'y a pas de référence absolue sur la position des capteurs : chanter

2. Une version temps-réel de « Chant » a été réalisée dès 1984 sur processeur de traitement de signal TMS320, et installée dans la première exposition permanente du Musée des Sciences de la Villette à Paris (Déchelle *et al.*, 1984). Mais le contrôle des paramètres était limité à des potentiomètres.

juste est quasiment impossible avec un tel système, à cause de la difficulté d'atteindre la position exacte des notes. Pour les mêmes raisons, un joystick, la WII ou tout autre périphérique du même type, sans repérage spatial, se prêtent mal à un contrôle mélodique précis, continu et rapide. De même le Theremin, conçu comme un instrument mélodique joué par les mains libres dans l'espace, reste très difficile à maîtriser.

Les premiers essais d'utilisation de la tablette graphique en conjonction avec un synthétiseur vocal sont apparus il y a une quinzaine d'années (Wanderley *et al.*, 2000 ; Kessous, 2002). Ce type d'interface semble capable de contrôler finement les variations de la voix, et c'est le choix adopté pour le Cantor Digitalis et le Digitartic.

Après avoir essayé toutes sortes d'IHM disponibles, il nous est apparu que les gestes manuels réglés (« chironomie »), c'est-à-dire les gestes de dessin, permettaient un contrôle du son suffisamment fin pour jouer de la musique. Jouer de la musique signifie ici être capable de jouer une partition ou d'improviser dans différents styles, en contrôlant parfaitement la mélodie, le rythme, les nuances, le timbre, comme avec un autre instrument musical. Les gestes manuels sont « réglés » (et non des gestes libres) par des repères, comme par exemple un dessin de clavier ou les limites d'une tablette graphique. Les gestes peuvent être directs, avec les doigts, ou bien indirects, par la médiation d'un stylet. Nous utilisons des tablettes graphiques, éventuellement tactiles, équipées de patrons qui représentent une échelle musicale et/ou un espace vocalique, ou bien des lieux articulatoires ou des phases articulatoires, comme expliqué plus bas (notamment les figures 2 et 4).

D'autres types d'interface peuvent bien sûr être envisagées. Une interface de type « accordéon » (Cook, Leider, 2000) permet un contrôle direct du souffle, comme dans la voix. Dans cet instrument, l'intonation est contrôlée par un clavier muni d'un ruban continu. L'idée du contrôle de l'intonation par un ruban continu existe depuis longtemps dans l'Onde Martenot par exemple. Cette idée a été intégrée récemment au Haken Continuum [3] et au Soundplane [4], des instruments à clavier 2D tactile continu. Le module de synthèse du Cantor Digitalis a récemment été intégré à ces deux instruments [5] [6]. La voix étant multidimensionnelle, un simple clavier ne peut de toute façon pas suffire à reproduire toutes ses variations (force de voix, tension, articulation, etc.).

La méthode de synthèse est décrite dans la partie suivante. Le Cantor Digitalis, un instrument permettant de chanter des voyelles et des semi-voyelles à l'aide d'une tablette graphique est présenté dans la partie 3. Le Digitartic est une extension du Cantor Digitalis aux consonnes. Il est décrit dans la partie 4 en discutant des aspects spécifiques au contrôle de l'articulation des consonnes. La dernière partie présente l'usage, des évaluations, et les limites de ces deux instruments.

3. http://www.hakenaudio.com/Continuum/
4. http://madronalabs.com/soundplane
5. https://youtu.be/R2XRfhu95Dc
6. https://youtu.be/oVQMHX4bQuo

2. Méthode de synthèse

2.1. Les composants d'un instrument vocal

Un instrument vocal est au service de l'intention musicale de celui qui en joue. Le musicien vise une certaine tâche musicale, une mélodie par exemple avec une voyelle donnée, une certaine qualité vocale, etc. Cette intention musicale s'exprime dans le contexte de nos instruments par des gestes manuels, mouvements du stylet et des doigts sur la tablette. Après le choix du type de voix et d'autres présélections, l'interface permet de contrôler des paramètres musicaux immédiatement perçus par le musicien, comme la mélodie, l'effort vocal ou les voyelles.

Ces paramètres musicaux, de haut-niveau, doivent être convertis en paramètres du synthétiseur, ou paramètres de bas-niveau. Un ensemble de règles de synthèse permettent de calculer les paramètres de bas-niveau à partir des paramètres de haut-niveau. Les paramètres de bas-niveau pilotent le synthétiseur paramétrique, qui produit les échantillons de synthèse. Ces échantillons sont transformés en signal acoustique par la chaine de conversion numérique-analogique puis électroacoustique. Ils sont perçus par le musicien, qui réagit en conséquence. La boucle de perception-action du jeu musical est bouclée. En résumé :

1. l'intention musicale conduit au contrôle des paramètres de haut-niveau, effectué par des gestes, qui reproduisent par analogies certains mouvements de l'appareil vocal, comme des mouvements intonatifs ou des mouvements articulatoires ;

2. ces gestes, mouvements des stylets ou des doigts, sont capturés par une ou deux tablettes graphiques ;

3. un ensemble de règles transforme les paramètres graphiques en paramètres de bas-niveau ;

4. le signal numérique est calculé par un synthétiseur paramétrique, à partir des paramètres de bas-niveau ;

5. les échantillons sont convertis en son et diffusés.

Les étapes 2, 3 et 4 forment la méthode de synthèse, qui est explicitée maintenant.

2.2. Paramètres de haut-niveau

Les paramètres de haut-niveau sont les paramètres musicaux vus du côté du musicien. Ils correspondent aux boîtes extérieures situées en haut et en bas de la figure 1 (dont les termes sont expliqués dans les paragraphes suivants). Ces paramètres sont commandés de façon intuitive par le musicien. Ils peuvent se grouper en 3 catégories :

les paramètres morphologiques représentent les propriétés du chanteur virtuel, soit la taille du conduit vocal, les valeurs moyennes de tension vocale et de bruit dans la voix, la tessiture, le système vocalique. Ces paramètres sont en principe fixés sous forme de presets pour une voix donnée, mais rien n'interdit de jouer avec dans le synthétiseur ;

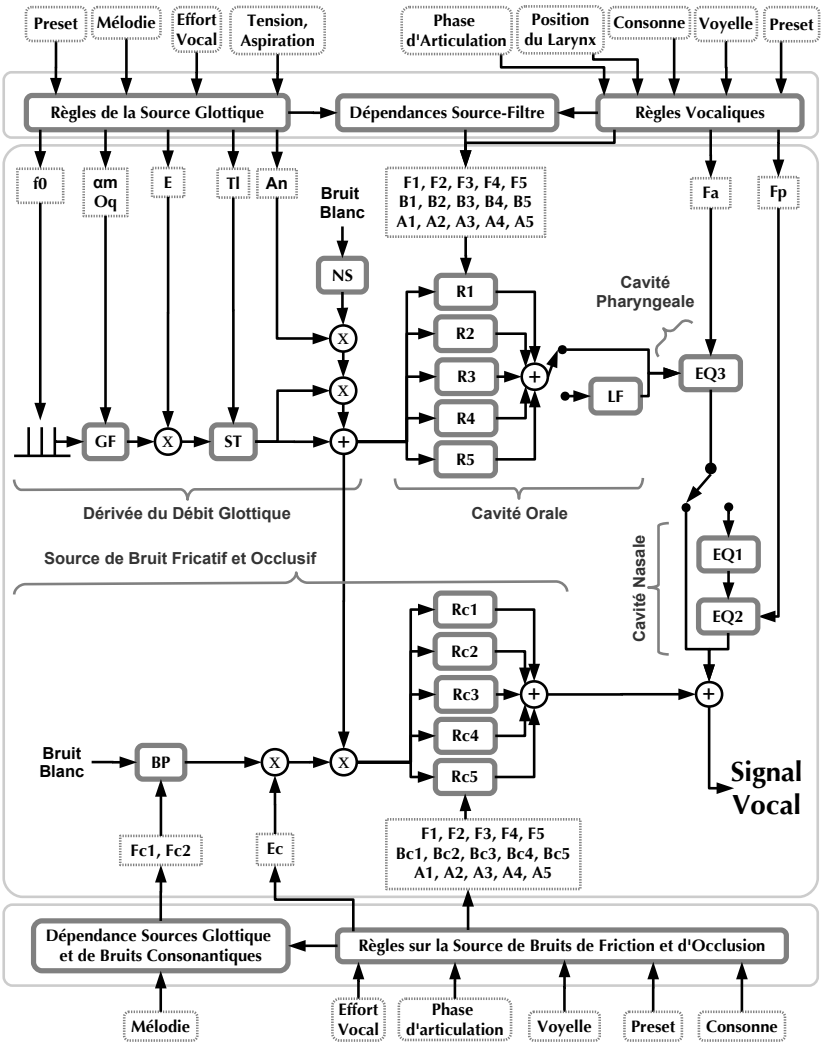

Figure 1. Structure du système de synthèse

les paramètres intonatifs comprennent la hauteur mélodique, l'effort vocal, la tension vocale ou le bruit dans la voix. De ces paramètres va dépendre une grande part de l'expressivité ;

les paramètres articulatoires comprennent les voyelles, les lieux, modes et phases d'articulation des consonnes, la position du larynx.

2.3. Synthétiseur paramétrique

Pour garantir le maximum de souplesse, l'instrument vocal est un synthétiseur paramétrique à formants. Ainsi, tous les paramètres vocaux peuvent être contrôlés à volonté par les gestes du musicien, le signal étant entièrement calculé, sans utilisation d'échantillons enregistrés. Le synthétiseur est de type source/filtre, avec une structure hybride à formants série/parallèle (Klatt, 1980 ; Holmes, 1983). Un ensemble de règles d'interactions source-filtre permet de raffiner le modèle. Un schéma fonctionnel du synthétiseur est porté sur la figure 1. Deux branches traitent d'une part de la composante voisée, d'autre part de la composante de bruit consonantique.

Pour la partie voisée, de façon classique, le signal vocal est calculé en filtrant par les résonances du conduit vocal un signal correspondant à la source d'excitation, le tout filtré par un filtre de rayonnement aux lèvres. Cette partie est représentée dans la moitié supérieure de la figure 1. On peut ramener à la source le terme de rayonnement aux lèvres, qui correspond à une dérivation, en considérant la dérivée du débit glottique. Le modèle de dérivée du débit glottique utilisé est le CALM (Causal Anticausal Linear Model) (Doval *et al.*, 2003). L'onde de débit glottique est une impulsion asymétrique, formée par un résonateur anticausal GF, filtré par un filtre passe-bas ST qui règle la pente spectrale. Ce modèle a 5 paramètres: $f0$, la fréquence fondamentale, qui correspond à la mélodie, Oq le quotient ouvert, αm l'asymétrie, Tl la pente spectrale, E l'amplitude de voisement. Ces paramètres se combinent pour rendre compte de la tension et de l'effort vocal (Doval *et al.*, 2006). Un bruit d'aspiration dont le spectre de puissance est donné par NS et l'amplitude par An permet de compléter la source vocale. Le filtre associé au conduit vocal comprend 5 résonateurs du second ordre en parallèle ($R1$ à $R5$, avec les fréquences centrales $F1$ à $F5$, les largeurs de bande $B1$ à $B5$ et les amplitudes spectrales $A1$ à $A5$) pour modéliser les formants vocaux. Cette branche parallèle est tout en série avec un filtre biquadratique $EQ3$, de fréquence centrale Fa qui permet d'ajouter l'antirésonance de l'hypopharynx. Deux filtres biquadratiques $EQ1$ et $EQ2$ en série, de fréquence centrale Fp, représente l'effet des cavités nasales, pour les occlusives nasales. Un filtre passe-bas supplémentaire, LF, modélise le spectre lors de la phase de tenue des plosives voisées.

Un circuit en parallèle (représenté dans la moitié inférieure de la figure 1) modélise les bruits occlusifs et fricatifs des consonnes. Un bruit blanc d'amplitude Ec est filtré par un passe-bande BP de fréquences de coupure $Fc1$ et $Fc2$, éventuellement modulé par la source voisée pour les consonnes voisées. Cette source de bruit est filtrée par des résonateurs du second ordre en parallèle ($Rc1$ à $Rc5$, avec les fréquences centrales $F1$ à $F5$, les largeurs de bande $Bc1$ à $Bc5$ et les amplitudes spectrales $A1$ à $A5$) pour colorer le bruit en fonction du contexte vocalique.

2.4. Règles de synthèse

Dans un système de synthèse de la parole par règles, le synthétiseur paramétrique est piloté par un ensemble de règles de synthèse. La suite de phonèmes et les consignes prosodiques sont transformées par des règles en paramètres de synthèse. En plus des

règles, dans nos instruments de synthèse, une partie de l'évolution des paramètres de synthèse est contrôlée dynamiquement par le geste (comme la mélodie, l'évolution des voyelles ou les phases des consonnes). Il faut cependant toujours un ensemble de règles pour gérer certains aspects de la transformation entre paramètres de haut-niveau et paramètres de bas-niveau :

les règles de la source vocale transforment les paramètres de haut-niveau, comme la hauteur mélodique, l'effort vocal et la tension vocale, en paramètres du modèle d'onde de débit glottique ;

les règles du bruit consonantique règlent l'amplitude et le spectre du bruit en fonction de la consonne, du contexte vocalique, de la phase d'articulation, de l'effort vocal, et de la morphologie du chanteur ;

les règles vocaliques calculent les paramètres des résonateurs et l'effet des cavités pharyngée et nasales en fonction des voyelles et des consonnes, de leur phase d'articulation, et de la morphologie du chanteur ;

les règles de dépendance source-filtre permettent de reproduire des phénomènes récemment étudiées pour la voix chantée. Pour les voix aigües, les formants s'accordent avec les harmoniques dans le suraigu du registre. L'effort vocal modifie la fréquence du premier formant. Les amplitudes des formants sont atténuées en cas de coïncidence trop grande avec des harmoniques ;

la règle de dépendance des sources glottiques et de bruits consonantiques modifie les fréquences de coupure de la source de bruit afin de modéliser la déformation du conduit vocal avec la mélodie.

3. Cantor Digitalis : gestes mélodiques et vocaliques

Le Cantor Digitalis est un instrument vocal bi-manuel qui permet de chanter des voyelles ou des semi-voyelles. L'interface de contrôle est une tablette graphique équipée d'un patron imprimé, représenté sur la figure 2. Cette figure montre que pour contrôler la hauteur musicale et la force vocale, on utilise un stylet tenu par la main dominante, sur un clavier figuré. La main secondaire permet de contrôler l'espace vocalique en utilisant avec les doigts une surface tactile dans un espace figuré au-dessus du clavier. Cette interface de contrôle présente une latence (du geste de contrôle au son émis en passant par le modèle de production) inférieure à 10-20 ms. Cela induit causalité directe entre geste et son, comme avec les instruments acoustiques (Genevois, 1999). La résolution spatiale est également élevée, limitée par la largeur de la mine du stylet (0,25 mm) ou de celle du doigt.

3.1. Gestes mélodiques

Le stylet de la main dominante contrôle la hauteur mélodique et l'effort vocal (un paramètre de haut-niveau qui contrôle la pente spectrale et l'asymétrie de la dérivée de l'onde de débit glottique, l'intensité sonore et le souffle). L'effort vocal nécessitant

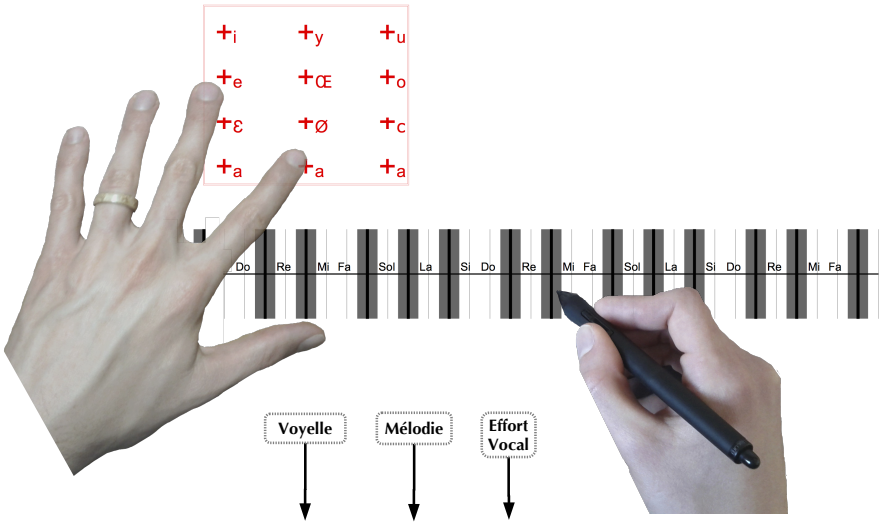

Figure 2. Configuration bi-manuelle du Cantor Digitalis, avec espace vocalique 2-D.

une précision moindre, il est relié 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 nécessite une grande précision. En faisant correspondre $f0$ à la position du stylet suivant l'axe X de la tablette, une précision au $1/2$ millimètre près est nécessaire, ne disposant que de $0,65$ cm par demi-ton. Cela est rendu possible à la fois par rapport aux caractéristiques techniques de la tablette et par rapport au geste.

Avec Cantor Digitalis, la façon la plus simple de chanter consiste à fixer le timbre, la voyelle et à ne contrôler que les aspects mélodico-rythmiques, l'intonation. L'utilisation de la chironomie en contrôle intonatif est particulièrement efficace. Les gestes mélodiques sont très proches des gestes d'écriture ou de dessin au trait. Ils sont donc efficaces lorsqu'on utilise un stylet et la main dominante. La vitesse des gestes intonatifs est celle des notes de musique. On peut considérer, pour la musique vocale, que des doubles croches, avec un tempo de 80 à la noire représentent un mouvement rapide. La durée d'une note est de l'ordre de 187 ms. Donc les notes de musique ont à peu près une durée comparable à celle d'une syllabe en parole. La question de l'ornementation, plus rapide, est traitée par de petits mouvements du stylet autour de la note visée. Elle peut s'opérer par des gestes variés, par un mouvement de va-et-vient le long de l'axe X ou par des petits cercles autour de la note comme illustré à la figure 3 par une succession de quatre notes vibrées.

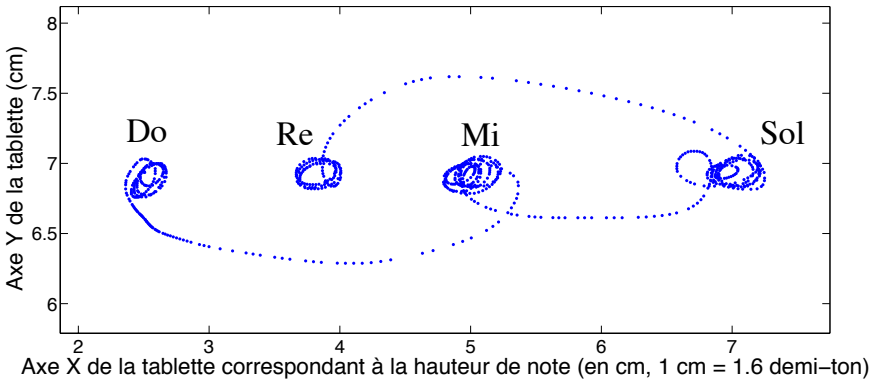

Figure 3. Trace du stylet produisant quatre notes consécutives (Do-Mi-Sol-Ré) avec vibrato sur chacune d'entre elle

3.2. Contrôle vocalique

La main secondaire permet de contrôler continûment l'espace vocalique. Sur le patron de la tablette, les voyelles sont représentées en deux dimensions, suivant les fréquences centrales de leurs deux premiers formants, ou triangle vocalique (Peterson, Barney, 1952). La position du doigt dans cet espace va déterminer la valeur des formants. Le triangle vocalique est ici 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. La surface est un carré de 5x6 cm de façon à pouvoir le parcourir avec l'index sans lever le poignet. L'instrumentiste peut ainsi mieux se concentrer sur la main dominante qui nécessite plus de précision.

Le contrôle continu des voyelles permet de produire également des semi-voyelles (/ɥ,w,j/). Les semi-voyelles sont produites en se déplaçant rapidement d'une voyelle à une autre par la trajectoire la plus courte dans le triangle vocalique. Dans ce cas, c'est la dynamique du geste qui contrôle continûment les phases d'articulation.

La configuration de contrôle tactile des voyelles et semi-voyelles dans un espace à deux dimensions a été souvent utilisée en concerts [7].

7. Des semi-voyelles sont produites dans « The lion sleeps tonight » de Solomon Linda : https://youtu.be/UtZkRsjy0ow

3.3. Typologies de voix et autres variations

De nombreuses possibilités de réglage des presets ou des paramètres moyens du Cantor Digitalis permettent d'obtenir une grande variété de sons vocaux, ou dérivés des sons vocaux. Des types vocaux prédéfinies sont proposés (le quatuor vocal, Soprano, Alto, Ténor, Basse, des voix pour le chant traditionnel bulgare, des voix soufflées, d'enfants, monstrueuses...). Il est possible de régler à volonté la taille apparente du conduit vocal, la tension, le souffle, la raucité dans la voix, les valeurs prédéfinies des formants des voyelles. Enfin, un algorithme sophistiqué de correction automatique de la mélodie permet d'assister le joueur dans les mouvements mélodiques rapides (Perrotin, d'Alessandro, 2013).

4. Digitartic : gestes mélodiques, vocaliques et consonantiques

4.1. Interface de contrôle

Digitartic s'est développé en parallèle du Cantor Digitalis, avec comme objectif d'étendre les possibilités de synthèse à un ensemble de consonnes. Musicalement, cela permet d'introduire des syllabes, comme dans le scat, ou de prononcer des mots. Cependant, la phonétique du français est riche (structures syllabiques complexes, clusters consonantiques à l'attaque ou à la coda, voyelles nasales, nombreux types de consonnes). Il nous faut donc réduire notre ambition à un sous-ensemble d'allophones et à des structures syllabiques simples, sans clusters consonantiques. La figure 4 montre le principe du fonctionnement du Digitartic, avec deux tablettes graphiques et deux stylets.

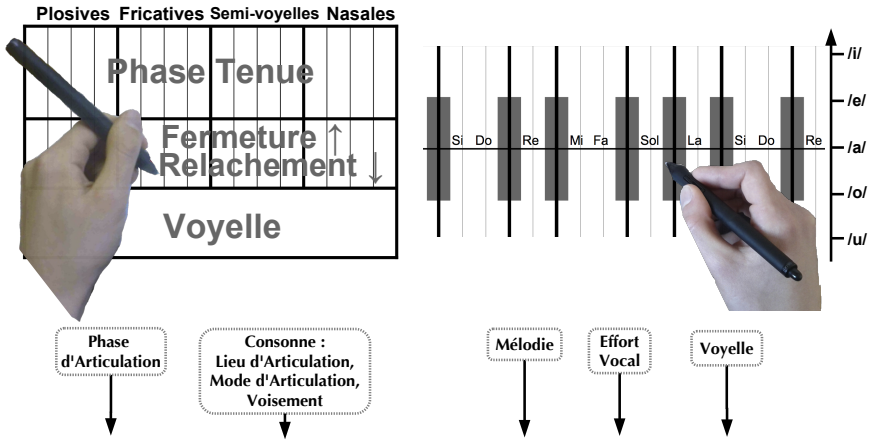

Figure 4. Représentation schématique du fonctionnement du synthétiseur Digitartic, pour un droitier

Une première tablette graphique pour la main préférée permet de contrôler les voyelles et la source glottique de la même manière qu'avec le Cantor Digitalis (partie droite de la figure 4). A la différence du Cantor Digitalis, les voyelles sont ici contrôlées par la main dominante suivant l'axe vertical de la tablette (et non plus sous la forme d'un triangle vocalique, par manque de degré de liberté). L'ordre des cinq voyelles a été établi en maximisant la similarité des voyelles adjacentes sur l'axe Y de la tablette pour avoir les transitions les plus naturelles, et l'espace vocalique du français. Les voyelles sont en principe distribuées dans un espace à deux dimensions (correspondant aux deux premiers formants). Il faut les ramener sur un espace à une seule dimension. Ainsi, l'ordre des voyelles sur l'axe 1D correspond à la projection des deux arêtes du triangle vocalique opposées à l'axe du deuxième formant. Les deux 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.

Une seconde tablette graphique est utilisée comme interface de contrôle de l'articulation des consonnes, par la main secondaire (partie gauche de la figure 4). Cette seconde 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. Un patron imprimé sur la tablette donne des repères (voir figure 5 pour les détails).

4.2. Gestes articulatoires : lieux et modes d'articulation

Les gestes articulatoires sont généralement plus rapides que les gestes intonatifs. La syllabe, « note de musique » minimale correspond généralement à un ou plusieurs phonèmes, avec une attaque, un noyau vocalique et une coda. Le noyau vocalique, en français une voyelle, est caractérisé par la vibration des plis vocaux et une configuration des organes articulateurs relativement stable pendant sa production. Le geste articulatoire est alors limité à la sélection de la voyelle.

Les consonnes correspondent à des mouvements rapides de constriction des organes articulateurs. Le lieu d'articulation est celui de la constriction la plus forte dans le conduit vocal. Le mode d'articulation dépend de la nature de la constriction et de la présence ou de l'absence de voisement (Laver, 1994). Une obstruction totale donne des consonnes occlusives /p,b,t,d,k,g,m,n/. Une obstruction partielle, mais suffisante pour produire un bruit de friction important, donne des consonnes fricatives /f,v,s,z,ʃ,ʒ,ʁ/. Une obstruction plus faible donne les approximantes, liquides /l/ ou /r/, ou semi-voyelles /ɥ,w,j/. Dans le mode d'articulation intervient également la nasalité pour distinguer consonnes orales et nasales, et le voisement, pour distinguer consonnes sourdes et voisées. Les zones de contrôle de la main secondaire sur la tablette sont schématisées sur la figure 5 :

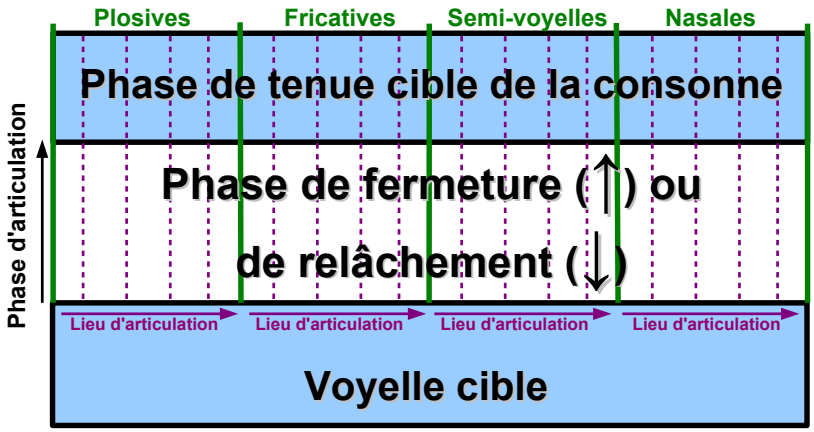

Figure 5. Représentation schématique des différentes zones de contrôle de la tablette contrôlant les consonnes, vue de dessus

le lieu d'articulation de la consonne est contrôlé continûment avec la position du stylet suivant l'axe X de la tablette, avec des positions de référence correspondant aux consonnes du français (lignes verticales pointillées) ;

la phase d'articulation est contrôlée continûment avec la position du stylet suivant l'axe Y (mouvement vertical dans la zone « Phase de fermeture ou de relâchement ») ;

le voisement de la consonne est modifié avec le bouton du stylet ;

le mode d'articulation correspond aux différentes régions discrètes de la tablette séparées suivant l'axe X (lignes verticales continues).

4.3. Phases articulatoires et ictus rythmique

La production de consonnes demande de synchroniser les organes articulateurs (lèvres, langue, mâchoire et luette) pour changer dynamiquement la forme du conduit vocal, sur des durées brèves (quelques dizaines de millisecondes). L'articulation d'une consonne comprend trois phases (Laver, 1994, p. 133-134) : la phase de fermeture des articulateurs vers la position de constriction maximale ; la phase de tenue, correspondant au maximum de constriction ; la phase de relâchement des articulateurs, qui ouvrent la constriction.

La tenue produit un arrêt du flux d'air pour les plosives, un arrêt du flux d'air dans le conduit oral seul pour les occlusives nasales, 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 dépend de la dynamique des articulateurs (durée minimale nécessaire pour passer d'une configuration à une autre), et de contraintes de sens pour faire des distinctions phonologiques ou expressives.

Dans le cas du chant, une contrainte supplémentaire provient de l'organisation rythmique. L'appui rythmique généré par le mouvement d'articulation de la consonne doit se synchroniser avec l'ictus rythmique, c'est à dire l'appui rythmique correspondant au début de la note de musique. L'appui rythmique dû à l'articulation d'une syllabe consonne-voyelle CV correspond à son « centre perceptif », ou P-center (Gordon, 1987 ; Morton *et al.*, 1976)). Il se situe au début de la phase de relâchement de la consonne pour les plosives, c'est à dire au début de la transition entre la consonne et la voyelle, au moment de l'explosion due au relâchement brusque des articulateurs. Dans ce cas il faut anticiper la phase médiane de façon à produire l'explosion en même temps que l'ictus rythmique. Pour les autres consonnes, le centre perceptif se situe plutôt au début de la voyelle, c'est à dire en fin de phase de relâchement.

Trois modes de contrôle de l'articulation de dissyllabe VCV, correspondant à 3 degrés de précision rythmique sont proposés (voir figure 6). Plus le contrôle est détaillé, plus fine est la synchronisation, mais plus le nombre de paramètres à contrôler simultanément est important, donc la tâche difficile. Le choix du mode sera fonction du contexte musical et du nombre de paramètres à contrôler simultanément, afin de trouver le meilleur compromis.

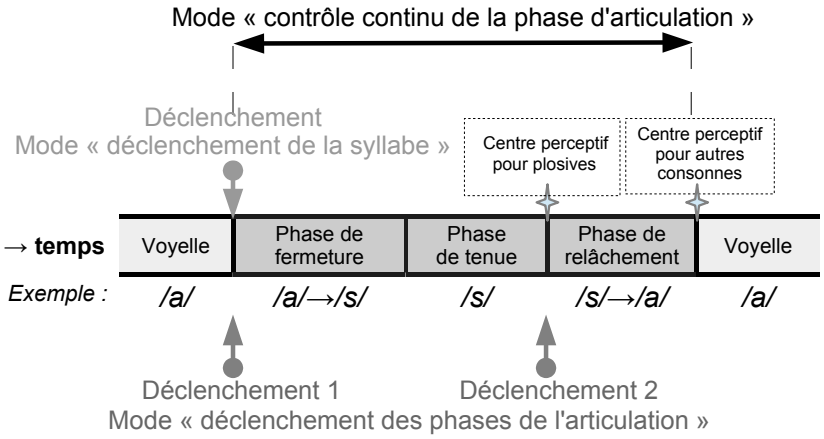

Figure 6. Les différents modes de contrôle suivant l'axe temporel des phases d'articulation et le centre perceptif d'attaque de la syllabe

Déclenchement à la fermeture. Le mode le plus simple pour jouer une dissyllabe est de la déclencher à partir du début de la phase de fermeture, en une seule fois. Si le déclenchement se fait en même temps que l'ictus rythmique, la syllabe sera perçue avec un retard correspondant à la phase de fermeture ajoutée à la phase de tenue pour les plosives, ou à toute la durée de l'attaque syllabique pour les autres consonnes. Anticiper ce retard en déclenchant avant l'ictus rythmique est très difficile à réaliser de manière précise. Ce mode est

le plus simple, mais le moins précis rythmiquement. Il pourrait convenir pour de la parole non musicale, ou si l'on n'utilise que des syllabes CV avec une plosive pour consonne (ou des séquences de syllabes CV avec la voyelle qui s'arrête avant le déclenchement de la plosive), comme dans le cas de la pièce Luna Park (Beller, 2011).

Double déclenchement : fermeture/relâchement. Pour plus de précision temporelle, le déclenchement de la consonne d'attaque d'une syllabe peut se faire en deux temps : le premier pour la phase de fermeture, et le second pour la phase de relâchement. On peut ainsi anticiper l'ictus rythmique et le synchroniser avec le centre perceptif dans le cas des plosives, mais pas pour les autres consonnes où un retard de la durée de la phase de relâchement sera présent par rapport à l'ictus rythmique (voir figure 6).

Contrôle continu de l'articulation. On peut également contrôler l'articulation de manière continue, sans déclenchement. Cela exige une temps de calcul de la synthèse et un envoi des données de l'interface vers le synthétiseur en moins de 20 ms, ainsi que le choix d'un geste de contrôle suffisamment rapide et précis. Si ces conditions sont respectées, il est possible d'obtenir un résultat similaire au contrôle de la voix naturelle, en ce qui concerne la synchronisation entre articulation et ictus rythmique. D'autre part, ce mode de contrôle permet également de synchroniser ses propres gestes pour contrôler différents paramètres tels que l'effort vocal et la hauteur mélodique, afin d'accentuer des syllabes sur certaines phases articulatoires et d'avoir ainsi accès à une plus grande variété d'articulation. Le geste de contrôle peut être unidirectionnel ou bien sous la forme d'un aller-retour entre la voyelle et la phase de tenue au milieu. Cela permet d'atteindre plus ou moins la cible de la phase de tenue et ainsi d'hypo-articuler. C'est ce mode de contrôle qui est utilisé généralement dans Digitartic.

4.4. Cibles acoustiques : formants et bruits consonantiques

Différentes bases de formants on été utilisées (Garnier-Rizet, 1994 ; Klatt, 1980 ; Stevens, 1998) et adaptées à notre synthétiseur. Toutes ces valeurs ont été ajustées avec la voyelle /a/, en situation /a/C/a/, et ont été conservées pour les autres contextes vocaliques. C'est une approximation assez grossière étant donnée que la forme du conduit vocal sur la phase de tenue dépend de la voyelle précédent ou suivant cette consonne. Le Digitartic fournit donc une qualité optimale pour des dissyllabes /a/C/a/.

Une autre approximation est que le spectre du bruit consonantique, indépendamment de son évolution dans le temps, dépend seulement du lieu d'articulation, du contexte vocalique et du voisement. Ainsi, les bruits de friction et d'explosion ont un spectre similaire, la différence venant essentiellement de la durée et de l'évolution de l'intensité. Pour la synthèse du bruit consonantique, on procède de la façon suivante : chaque lieu d'articulation est associé à une source de bruit spécifique, bruit blanc filtré par un filtre Butterworth du second ordre (BP sur la figure 1) dont les fréquences de coupure $Fc1$ et $Fc2$ dépendent principalement du lieu d'articulation ; l'amplitude Ec

du bruit est modulée par l'onde de débit glottique dans le cas des consonnes voisées, et en fonction de plusieurs paramètres (phase d'articulation, lieu d'articulation, mode d'articulation et voisement de la consonne) ; pour reproduire les effets de coarticulation, le bruit est filtré partiellement par le spectre de la voyelle adjacente (en utilisant les mêmes filtres formantiques que les voyelles mais avec une bande passante Bc plus importante).

Pour améliorer la cohérence des bruits consonantiques avec le modèle de débit glottique, on ajoute certaines dépendances. Les fréquences de coupures $Fc1$ et $Fc2$ du filtre BP sont rendues dépendantes de $f0$, et les bandes-passantes Bc des résonateurs Rc reliées à l'effort vocal. Ces dépendances modélisent en fait le changement de forme du conduit vocal avec $f0$, ou avec l'effort vocal qui modifie à son tour le bruit consonantique.

4.5. Contrôle de l'articulation des syllabes

Digitartic utilise le mode de contrôle continu de la phase d'articulation décrite plus haut, afin d'assurer le contrôle de l'ictus rythmique, et pour augmenter la palette de possibilités articulatoires et par conséquent les capacités expressives de l'instrument. Digitartic permet d'articuler des dissyllabes de type $V1CV2$, où $V1$ et $V2$ sont des voyelles quelconques, interpolées suivant l'axe /i,e,a,o,u/ et où C désigne une consonne du français parmi /p,b,t,d,k,g,f,v,s,z,ʃ,ʒ,w,ɥ,j,m,n,ŋ/ et leurs interpolations suivant le lieu d'articulation pour un même mode d'articulation. Les syllabes s'enchaînent à volonté pour créer des énoncés plus long.

Certaines séquences VCV montrent une évolution symétrique des paramètres par rapport à la phase de tenue. Les fricatives, semi-voyelles et occlusives nasales présentent une structure temporelle plutôt symétrique autour de la phase médiane de la séquence VCV. Les plosives présentent quant à elles une dissymétrie par rapport à la phase médiane (un silence, ou une barre de voisement). En effet, l'explosion des plosives ne se manifeste que lors de la phase de relâchement et non lors de la phase de fermeture. Ainsi on traitera de manière symétrique la synthèse des phases de fermeture et de relâchement si la consonne est une fricative, une semi-voyelle ou une occlusive nasale. Concernant les plosives, 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 phases de fermeture et de relâchement. Un exemple d'évolution des principaux paramètres que sont voisement, aspiration et formants est donné à la figure 7.

Dans le modèle de contrôle, pour un mode d'articulation donné (plosives, fricatives, semi-voyelles, nasales, voisées ou non), on dispose sur la tablette de deux paramètres articulatoires (se reporter à la figure 4 et 7) :

– la dimension horizontale, correspond à la position du lieu d'articulation de la phase de tenue cible, calculée par interpolation des valeurs de paramètres des deux phases 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

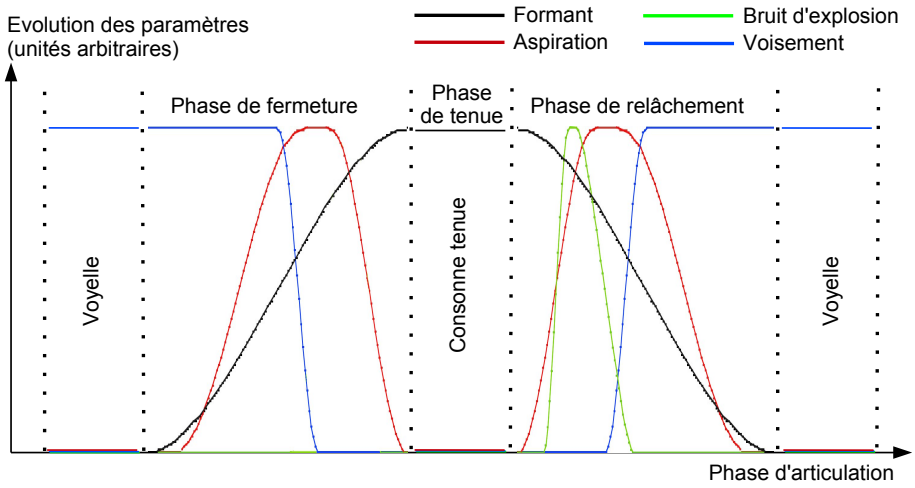

Figure 7. Exemple d'évolution des paramètres de la transition articulatoire pour une séquence V/p/V

valeur des formants consonantiques cibles, 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.

– la dimension verticale, correspond à la phase d'articulation, c'est-à-dire à la position articulatoire entre les deux états « voyelle cible » et « phase de tenue cible », contrôlant l'évolution temporelle des formants, du bruit fricatif et occlusif, du taux d'aspiration, du voisement, et de la nasalité.

4.6. Dynamique du geste des différentes consonnes et rendu sonore

Sur la tablette, la distance séparant la zone de phase de tenue de celle de la voyelle (cf figure 5) 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 dynamique du geste modifie les trajectoires formantiques (voir figure 7). Ce geste doit être adapté à la dynamique des modes d'articulation, qu'on regroupe sous deux catégories :

– Les plosives 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 de ces occlusives de synthèse n'est pas bien corrélé au ralentissement du geste articulatoire naturel des plosives qui voudrait que l'explosion ait lieu quelle 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 la vitesse d'articulation désirée, suivant son adaptation à un contexte rythmique, on modifie la dynamique en conséquence. Pour produire des syllabes se rapprochant de la voix parlée, on doit effectuer le geste de transition sur environ 60 à 80 ms.

Pour les consonnes ne présentant pas de bruit d'occlusion ou de friction, il est possible d'hypo-articuler en n'atteignant pas avec le stylet la cible de phase de tenue (voir page suivante pour un lien vers un exemple sonore).

Il est également possible (par interpolation) d'articuler des consonnes avec des lieux d'articulation intermédiaires des consonnes canoniques du français, par exemple entre /w/ et /ɥ/, en ciblant un lieu intermédiaire sur la tablette. La figure 8 donne le spectrogramme d'une série de semi-voyelles et de plosives dont le lieu d'articulation varie entre labial et palatal.

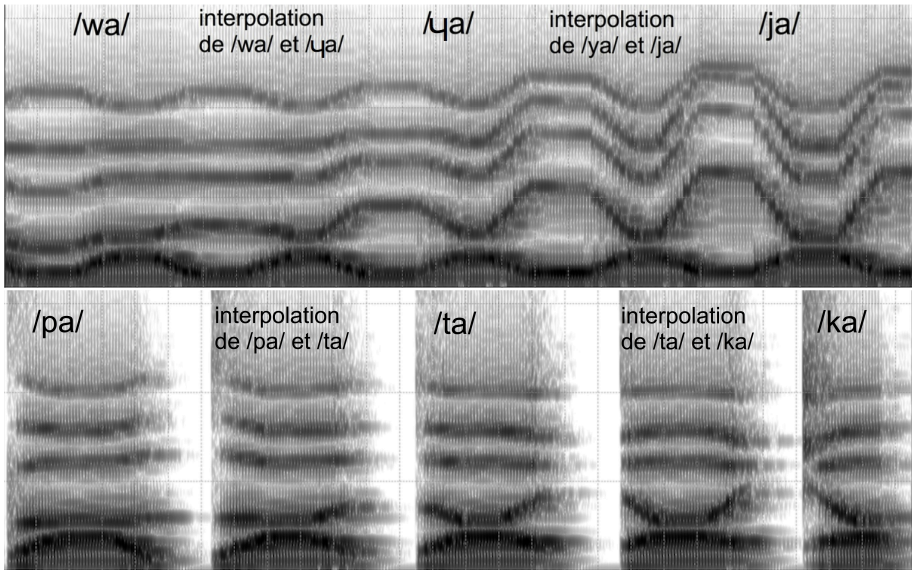

Figure 8. Spectrogrammes ($0 - 6000$ *Hz) de successifs C-/a/ (C1-/a/-C2-/a/-C3-...) produits par le Digitartic, avec le lieu d'articulation évoluant sur l'axe bilabial - alveolaire - palatal. En haut : des semi-voyelles. En bas : des plosives sourdes.*

5. Discussion

5.1. Apprentissage, usage et distinction musicales

Le Cantor Digitalis est simple à apprendre, avec l'utilisation à la fois d'un stylet pour la mélodie et des doigts pour les voyelles. C'est de plus un instrument peu encombrant, une seule tablette, et riche de nombreuses possibilités de timbre. L'apprentissage du Cantor Digitalis est rapide, de l'ordre de quelques heures pour un usage de base (comme pour un piano où la production du son ne nécessite pas ou peu d'apprentissage, contrairement à un violon). L'interaction avec des enfants lors de fêtes de la science ont révélé un accès très facile à l'instrument. Cependant, un jeu expert peut se développer, au prix de centaines d'heures de travail (les joueurs les plus expérimentés du Cantor Digitalis ont actuellement 4 ans de pratique), et la virtuosité est comparable à ce que l'on peut faire avec des instruments traditionnels.

Le Digitartic est quant à lui pratiqué dans un cercle beaucoup plus restreint de musiciens à cause du contrôle plus difficile des consonnes, ainsi que par le stade moins mature de son développement. De plus, dans l'état actuel, deux tablettes graphiques sont utilisées de façon à capter des gestes bi-manuels avec la plus grande résolution à disposition, à savoir l'usage d'un stylet associé à chaque tablette (la fonction tactile peut être utilisée en mono-tablette en parallèle de l'usage du stylet avec l'autre main, mais celle-ci présente une résolution insuffisante en regard de la rapidité des mouvements articulatoires). Enfin, le lieu et la phase d'articulation pouvant 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.

Les instruments Cantor Digitalis et Digitartic sont régulièrement réunis au sein du Chorus Digitalis, un ensemble musical de voix de synthèse [8]. La qualité musicale du Cantor Digitalis a récemment été reconnue lors d'une compétition annuelle des nouveaux instruments de musique, la *Margaret Guthman Musical Instrument Competition*, dont le premier prix a été remportée par le Cantor Digitalis en 2015 [9] [10]. Le répertoire parcourt différentes traditions musicales comme la musique baroque, le chant Khayal d'Inde du Nord, ou encore les chants bulgares. D'autres styles plus contemporains ont été travaillés tels des chorales aux racines jazz [11], ou des polyphonies populaires comme *Le lion est mort ce soir* [12].

Le Cantor Digitalis est distribué sous licence libre CeCILL. Le Digitartic devrait également faire l'objet d'une distribution libre ultérieurement.

---

8. https://cantordigitalis.limsi.fr/chorusdigitalis_fr.php
9. http://guthman.gatech.edu/
10. http://nymag.com/next/2015/03/listen-to-a-chorus-made-of-singing-synthesizers.html
11. Voir un ensemble de six musiciens interprétant « Valse » de Bruno Lecossois : https://youtu.be/9-9_YCJRe-4
12. dont le lien web est cité plus haut.

5.2. Évaluations

5.2.1. Évaluation du contrôle mélodique

Les instruments développés dans ce travail sont basés sur des analogies entre gestes manuels, comme des gestes de dessin et d'écriture, et de production vocale. Nous avons montré ailleurs que les gestes manuels étaient comparables (voire meilleurs) que la voix pour la justesse et la précision mélodique (d'Alessandro *et al*., 2014) (ainsi que pour l'imitation de l'intonation prosodique (d'Alessandro *et al*., 2010)). Cette qualité du contrôle mélodique est clairement une des clés du naturel et de l'expressivité de nos instruments vocaux.

Pour un musicien maîtrisant déjà un autre instrument, au prix d'un travail important, un haut degré de virtuosité peut être atteint avec un stylet sur une tablette graphique [13].

5.2.2. Caractérisation qualitative de la dynamique gestuelle du contrôle des syllabes

La comparaison de spectrogrammes de séquences VCV pour des séquences de synthèse contrôlées par le geste et des séquences de voix naturelles montre qu'il est possible de reproduire la dynamique des trajectoires des formants de consonnes naturelles. Sur les figures 9 et 10, nous comparons 3 plosives synthétiques et 3 semi-voyelles synthétiques à leurs doubles naturels.

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 9), mais leurs valeurs sont plutôt reliées à la personnalisation de la voix et ont peu de sens phonologique ; l'intensité ne décroît pas assez rapidement sur la phase de fermeture dans cet exemple.

Le lecteur pourra écouter quelques enregistrements de consonnes [14] ainsi qu'à un extrait vidéo issu d'un concert-performance montrant comment sont joués quelques /papa/ [15]. Les exemples issus de (Feugère, 2013) et réalisés avec le Digitartic, donne un aperçu de la richesse possible du contrôle temporel de l'articulation des syllabes : accentuation de différentes phases d'articulation, durées différentes des phases de fermeture et de relâchement, précision rythmique et rapidité du geste d'articulation [16], ainsi que différents degrés d'hypo-articulation [17].

---

13. Voir par exemple un raga en style Khayal (Inde du Nord) joué plusieurs fois en concert : https://youtu.be/hRxr3ZWVjM4
14. Enregistrements de consonnes synthétiques contrôlées gestuellement (Feugère, 2013) https://tel.archives-ouvertes.fr/tel-00926980/file/7_apatakavazajawauayamana.wav
15. https://youtu.be/d4TV-IcK8c8?t=6m40s (à partir de 6'40)
16. https://tel.archives-ouvertes.fr/tel-00926980/file/8_vitessesArticulatoires_awa.wav
17. https://tel.archives-ouvertes.fr/tel-00926980/file/8_differentDegresArticulatoires_aja.wav

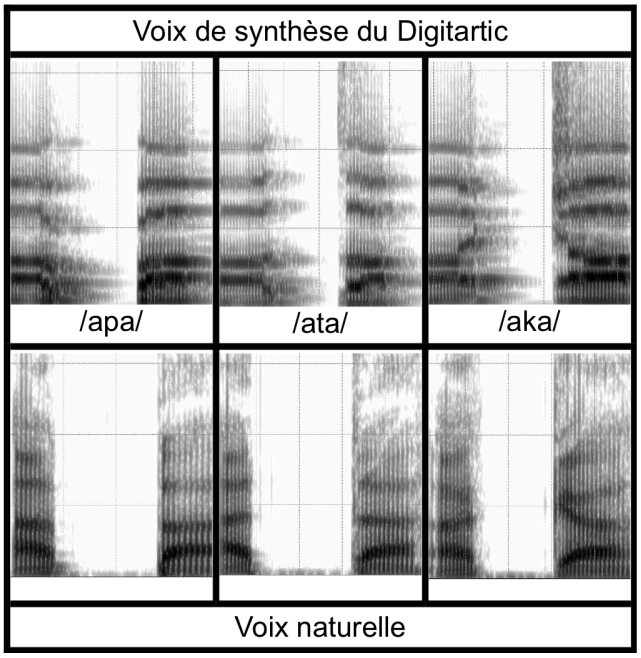

Figure 9. Spectrogrammes (sur environ 400 ms, $0 - 6000$ *Hz) de séquences* VCV *du Digitartic et de voix naturelles, pour trois occlusives*

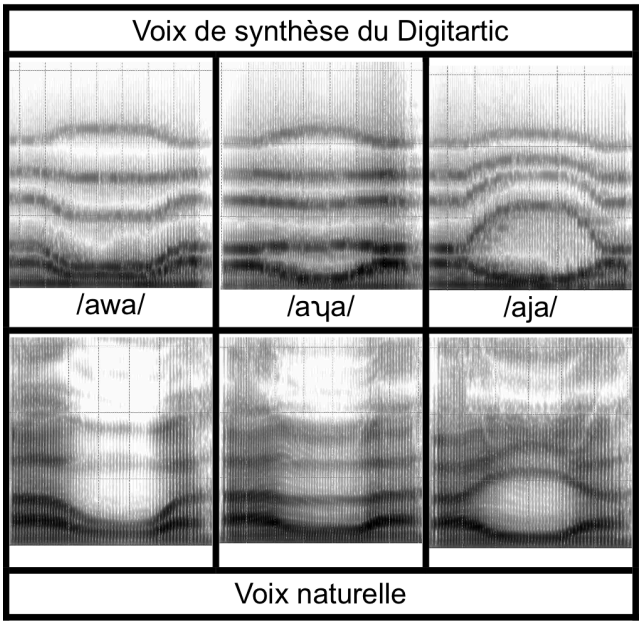

Figure 10. Spectrogrammes (sur environ 700 ms, $0 - 6000$ *Hz) de séquences* VCV *du Digitartic et de voix naturelles, pour trois semi-voyelles*

5.3. Limitations

Pour une utilisation où l'articulation se limite à des voyelles et semi-voyelles, il est préférable d'utiliser le Cantor Digitalis, plus facile d'accès que le Digitartic. Pour chanter des onomatopées comprenant des plosives, nasales et fricatives, le Digitartic est nécessaire.

Le modèle de contrôle principal du Digitartic permet de reproduire n'importe quelle séquence VCV, parmi toutes les consonnes 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 sonnent moins naturelles lorsqu'elles sont dans un contexte vocalique autre que /a/. Ainsi, il reste à étendre l'articulation à d'autres contextes vocaliques, ce qui impliquerait une interpolation à 3 dimensions entre la dimension de lieu d'articulation, celle de l'axe vocalique et celle de la phase d'articulation correspondante, et pour chaque mode d'articulation donné.

Vu la complexité du contrôle de l'articulation vocale, les choix ont été faits en fonction du but musical de l'instrument. La contrepartie de la possibilité d'avoir un contrôle fin sur l'articulation de la plupart des consonnes du français est qu'il est difficile de produire de la parole. Seulement certaines syllabes sont possibles, à penser sous forme musicale d'onomatopées et enchaînées à une vitesse plus faible que celle de l'articulation naturelle.

La synthèse paramétrique permet une certaine liberté dans le contrôle du modèle de par sa nature, en termes d'individualisation des voix, contrairement à des systèmes à base d'échantillons de voix naturelle transformés, comme le système de modification de voix temps-réel Calliphony (Le Beux *et al.*, 2007) qui offre moins de flexibilité sur le signal. Par contre, la qualité sonore de la synthèse paramétrique, en particulier en ce qui concerne les consonnes, est en général de moins bonne qualité que celle utilisant des segments pré-enregistrés.

6. Conclusion et perpectives

Ces instruments, Cantor Digitalis et Digitartic, démontrent 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 semble possible, malgré la rapidité des gestes d'articulation. Nos études ont montré que l'analogie entre gestes manuels et gestes mélodiques de la source glottique fonctionne de manière satisfaisante.

Les interfaces tactiles se développant actuellement, notamment avec stylet, il est probable que les instruments numériques en tirent avantage, pour une diffusion à plus grande échelle et dans le grand public.

La prochaine étape de développement du Cantor Digitalis se concentre sur la reproduction de voix naturelles variées, et sur des interfaces pour permettre à l'utilisateur de paramétrer aisément les voix de synthèse.

Concernant le Digitartic, une extension des diphones réalisables ainsi que des évaluations formelles de sa qualité sonore et de contrôle sont envisagées. Les mises en œuvres menées pour le contrôle du Digitartic, pour la synchronisation rythmique très précise qu'exige le jeu musical, pourront être appliquées à d'autres méthodes de synthèse comme le Calliphony. Enfin, la position distincte des centres perceptifs des syllabes CV comprenant une plosive et de celles comprenant les autres consonnes du français peuvent laisser suggérer qu'il serait plus intuitif d'avoir un geste de contrôle également distinct : double déclenchement « fermeture/relâchement » pour les occlusives et « contrôle continu de la phase d'articulation » pour les autres consonnes.

Bibliographie

Astrinaki M., D'Alessandro N., Picart B., Drugman T., Dutoit T. (2012, December, 2-5). Reactive and continuous control of HMM-based speech synthesis. In *IEEE workshop on spoken language technology (SLT 2012).* Miami, Florida, USA.

Beller G. (2011). Gestural control of real-time concatenative synthesis in Luna Park. In *1st international workshop on performative speech and singing synthesis (P3S 2011).*

Berndtsson G. (1995). The KTH rule system for singing synthesis. *STL-QPSR*, vol. 36, n° 1, p. 1-22.

Cook P. R. (1993). SPASM, a real-time vocal tract physical model controller; and singer, the companion software synthesis system. *Computer Music Journal*, vol. 17, n° 1, p. 30-44.

Cook P. R. (2005, May 26-28). Real-time performance controllers for synthesized singing. In *Proceedings of the 5th conference on new interfaces for musical expression (NIME'05).* Vancouver, BC, Canada.

Cook P. R., Leider C. N. (2000, August). squeezeVox: A new controller for vocal synthesis models. In *Proceedings of the 2000 international computer music conference (ICMC2000).* Berlin.

d'Alessandro C., Feugère L., Le Beux S., Perrotin O., Rilliard A. (2014, June). Drawing melodies: Evaluation of chironomic singing synthesis. *J. Acoust. Soc. Am.*, vol. 135, n° 6, p. 3601-3612.

d'Alessandro C., Le Beux S., Rilliard A. (2010, 12-16 Avril). Contrôle gestuel du modèle source/filtre de production de la voix. In *10ème congrès français d'acoustique.* Lyon.

D'Alessandro N., Woodruff P., Fabre Y., Dutoit T., Le Beux S., Doval B. *et al.* (2007, March). Real time and accurate musical control of expression in singing synthesis. *Journal on Multimodal User Interfaces*, vol. 1, n° 1, p. 31-39.

Déchelle F., d'Alessandro C., Rodet X. (1984). Synthèse temps-réel sur microprocesseur TMS 320. In *Proc. of the 1984 international computer music conference (ICMC1984)*, p. 15.

Depalle P., Garcia G., Rodet X. (1995). A virtual castrato (!?). In *Proc. of the 1994 international computer music conference (ICMC1994)*, p. 357-360.

Doval B., d'Alessandro C., Henrich N. (2003). The voice source as a causal/anticausal linear filter. In ISCA (Ed.), *Proceedings of voqual'03 : Voice quality : Functions, analysis and synthesis.* Geneva, Switzerland.

Doval B., d'Alessandro C., Henrich N. (2006). The spectrum of glottal flow models. *Acta Acoustica*, vol. 92, p. 1026-1046.

Dudley H., Riesz R. R., Watkins S. S. A. (1939). A synthetic speaker. *Journal of the Franklin Institute*, vol. 227, n° 6, p. 739-764.

Fels S. S., Hinton G. E. (1992, November). Glove-talk: A neural network interface between a data-glove and a speech synthesizer. *IEEE Transactions on neural networks*, vol. 3, n° 6, p. 1-7.

Fels S. S., Hinton G. E. (1998). Glove-talk II : a neural network interface which maps gesture to parallel formants. *IEEE Transactions on neural networks*, vol. 9, n° 1, p. 205.

Fels S. S., Pritchard R., Lenters A. (2009). Fortouch: A wearable digital ventriloquized actor. In *Proceedings of the 9th conference on new interfaces for musical expression (NIME'09).*

Feugère L. (2013). *Synthèse par règles de la voix chantée contrôlé par le geste et applications musicales*. Thèse de doctorat non publiée, Université Pierre et Marie Curie, Ecole doctorale Sciences Mécaniques, Acoustique, Electronique et Robotique (SMAER), Paris, France.

Feugère L., d'Alessandro C. (2012, 9-11 mai). Digitartic : synthèse gestuelle de syllabes chantées. In *Actes des journées d'informatique musicale (JIM 2012)*, p. 219-225. Mons, Belgique.

Feugère L., d'Alessandro C. (2013, May). 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)*, p. 331-336. Daejeon, Korea Republic.

Garnier-Rizet M. (1994). *Elaboration d'un module de règles phonético-acoustiques pour un système de synthèse à partir du texte pour le français*. Thèse de doctorat non publiée, Université de la Sorbonne nouvelle.

Genevois H. (1999). Geste et pensée musicale : de l'outil à l'instrument (dans "les nouveaux gestes de la musique"). In E. Parenthèse (Ed.),, p. 35-45.

Gordon J. W. (1987). The perceptual attack time of musical tones. *J. Acoust. Soc. Am.*, vol. 82, n° 2, p. 88-105.

Holmes J. (1983). Formant synthesizers: cascade or parallel? *Speech Communication*, vol. 2, p. 251-273.

Kenmochi H., Oshita H. (2007). Vocaloid – commercial singing synthesizer based on sample concatenation. In *Interspeech.*

Kessous L. (2002). 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)*, p. 113–114.

Kessous L. (2004). *Contrôles gestuels bi-manuels de processus sonores*. Thèse de doctorat non publiée, Université de Paris VIII.

Klatt D. H. (1980, March). Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Am.*, vol. 67, n° 3, p. 971-995.

Laver J. (1994). *Principles of phonetics* (Cambridge, Ed.). Cambridge.

Le Beux S., Feugère L., d'Alessandro C. (2011, 27/08 au 31/08). Chorus digitalis : experiment in chironomic choir singing. In P. of the conference ISSN: 1990-9772 (Ed.), *12th annual conference of the international speech communication association (INTERSPEECH 2011)*, p. 2005-2008. Firenze, Italy.

Le Beux S., Rilliard A., d'Alessandro C. (2007, August 22-24). Calliphony: A real-time intonation controller for expressive speech synthesis. In *6th ISCA workshop on speech synthesis*, p. 345-350. Bonn, Germany.

Miranda E. R., Wanderley M. M. (2006). New digital musical instruments: Control and interraction beyond the keyboard. *A-R Editions*, n° Middleton, WI, USA, p. 1-18.

Morton J., Marcus S., Frankish C. (1976, September). Perceptual centers (P-centers). *Psychological Review*, vol. 83, n° 5, p. 405-408.

Perrotin O., d'Alessandro C. (2013, May 27-30). Adaptive mapping for improved pitch accuracy on touch user interfaces. In K. R. Daejeon + Seoul (Ed.), *Proceedings of the 13th conference on new interfaces for musical expression (NIME'13)*, p. 186-189.

Peterson G. E., Barney H. L. (1952, March). Control methods used in a study of vowels. *J. Acoust. Soc. Am.*, vol. 24, n° 2, p. 175-184.

Pritchard B., Fels S. S. (2006). GRASSP: Gesturally-realized audio, speech and song performance. In *Proceedings of the 6th conference on new interfaces for musical expression (NIME'06)*, p. 272-276.

Rodet X., Potard Y., Barrière J.-B. (1984, Autumn). The CHANT project: From the synthesis of the singing voice to synthesis in general. *Computer Music Journal*, vol. 8, n° 3, p. 15-31.

Stevens K. N. (1998). *Acoustic phonetics*. The MIT Press.

Wanderley M. M., Viollet J.-P., Isart F., Rodet X. (2000). On the choice of transducer technologies for specific musical functions. In *Proc. of the 2000 international computer music conference (ICMC2000)*, p. 244–247.

Zbyszynski M., Wright M., Momeni A., Cullen D. (2007). Ten years of tablet musical interfaces at cnmat. In *Proceedings of the 7th conference on new interfaces for musical expression (NIME'07)*, p. 100-105. New York, USA.

Article soumis le 26 mars 2015
Accepté le 13 novembre 2015.