Une nouvelle organologie de la voix : chironomie et prosodie de la parole et du chant

Christophe d'Alessandro

Une nouvelle organologie de la voix : chironomie et prosodie de la parole et du chant

Christophe d'Alessandro

Institut Jean Le Rond D'Alembert, équipe LAM
Sorbonne Université – CNRS – Paris
christophe.dalessandro@sorbonne-universite.fr

RESUME

La synthèse vocale performative offre une nouvelle approche pour l'analyse et la synthèse de la parole expressive et du chant. Deux instruments vocaux, à la confluence entre nouvelles interfaces pour l'expression musicale, traitement du signal vocal et phonétique acoustique, sont présentés. Cantor Digitalis est un synthétiseur à formants muni d'un modèle spectral de la source glottique, piloté par des gestes bimanuels sur une surface tactile et un stylet. Voks est un vocodeur piloté par des gestes bimanuels, sur une surface tactile et un stylet, ou par un thérémine, augmenté d'un contrôle biphasique du séquencement syllabique. Les unités prosodiques, leur organisation et leur contrôle pour la construction des instruments vocaux sont discutés suivant trois lignes : la stylisation chironomique de l'intonation et de l'effort vocal, le contrôle biphasique du rythme syllabique. La conclusion évoque les applications musicales, pédagogiques, voire cliniques de cette nouvelle organologie de la voix.

ABSTRACT

The building of vocal instruments: chironomy and prosodic organization.

Performative speech synthesis is a new avenue for expressive speech and singing analysis and synthesis. At the confluence between new interfaces for musical expression, voice signal processing and acoustic phonetics, two vocal instruments are presented. Cantor Digitalis, a formant synthesizer using a spectral glottal model, is controlled using bimanual gestures on a graphic tablet. Voks is based on sampling and vocoding. It is controlled by bimanual gestures (graphic tablet or theremin) and biphasic syllabic sequencing. Prosodic units and their organization for vocal instruments are discussed along three lines: chironomic stylization of intonation and vocal effort, biphasic syllabic rhythm control. Musical, pedagogical and clinical applications of this new organology of voice are discussed.

MOTS-CLES : synthèse vocale, prosodie, instrument vocal, interface musicale expressive.

KEYWORDS: voice synthesis, prodody, voice instrument, expressive musical interface.

1 Introduction

1.1 Synthèse vocale expressive

La motivation de cette recherche est déjà ancienne (d'Alessandro et al., 2005) : après 33 ans de recherche, la synthèse à partir du texte en français restait bien incapable d'expression (d'Alessandro,

2001) . La synthèse vocale est de qualité acceptable pour lire des textes ou pour les systèmes d'information depuis plusieurs dizaines d'années. En dépit des progrès considérables en apprentissage automatique, statistique puis neuronal, ces machines à lire manquent toujours d'expressivité. Ce n'est pas seulement une affaire de dimension de corpus, de mémoire ou de vitesse de calcul : la qualité du rendu est aujourd'hui excellente lorsque l'on sait spécifier le contexte d'énonciation et l'expression voulue. Mais spécifier l'expression adaptée à la situation, en contexte d'énonciation, en interaction et avec une voix qui incarne un interlocuteur est toujours un défi pour la synthèse vocale. Rendre expressive la synthèse est un problème à double face : côté pile, la spécification de l'expression dans une situation donnée, côté face, la synthèse effective de l'expression. Nous laissons à d'autres le côté pile. Le côté face est abordé ici sous l'angle de la synthèse vocale performative : l'organologie de la voix ou construction d'instruments vocaux. C'est un humain qui joue l'instrument, le problème est de le faire sonner comme une voix et de permettre la variation expressive subtile. La synthèse vocale par contrôle gestuel nous semblait possible, malgré la qualité décevante de tentatives précédentes (Voder, SqueezeVox, GloveTalk, Vocalise, Voicer …). En faisant l'hypothèse d'un traitement central (perception, planification, effectuation) de l'expression vocale s'ouvre la possibilité d'utiliser les membres pour le contrôle expressif d'un modèle de l'appareil vocal. La boîte à outils, à développer le cas échéant, emprunte à la phonétique acoustique, au traitement du signal et à l'informatique musicale. Les interfaces et le mode d'emploi des instruments sont à inventer, en mêlant performance, perception et création. L'angle de la chironomie apporte des méthodes et des résultats pour revisiter ou réviser plusieurs questions dans le domaine prosodique : la stylisation intonative, le rôle de la qualité vocale, la production et la perception du rythme. La synthèse vocale performative invente de nouveaux instruments musicaux, permet d'envisager de nouvelles méthodes pour l'enseignement de la prosodie des langues étrangères, pour la suppléance vocale avec un larynx artificiel, pour l'analyse par la synthèse des phonostyles tant pour les sciences de la parole que pour la musicologie.

1.2 La construction des instruments vocaux

La construction des instruments vocaux repose sur trois hypothèses principales

1. **Codage de l'intonation**. Les motifs mélodiques, rythmiques et intensif sont considérés comme des gestes intonatifs, qui véhiculent les informations linguistiques et expressives. En suivant la théorie du codage d'évènement (theory of Event Coding) la représentation des stimuli sous-jacentes à la perception et les représentations d'action sous-jacentes à la planification motrice partagent un support de représentation commun, et indépendant de la modalité effectivement utilisée pour percevoir ou produire la prosodie. La théorie de « résonance perceptuelle » suggère que l'action initie et guide le processus de perception, en orientant la perception dans le domaine des effets de l'action.
2. **Substitution entre modalités.** La précision et la qualité du contrôle de la synthèse vocale performative sont suffisantes pour reproduire des schémas d'intonation chironomiques indiscernables des schémas d'intonation de la parole. Ainsi l'intonation peut être transférée de l'appareil vocal à d'autres modalités, dans notre cas les gestes de la main.
3. **Renforcement multimodal.** La synthèse vocale performative est envisagée dans le cadre du principe idéo-moteur : l'action est planifiée pour obtenir les résultats dont la perception est anticipée ; penser une action initie et guide sa réalisation. Après une période d'apprentissage, réciproquement, les effets d'une action peuvent être prédits et ainsi l'action planifiée en conséquence. Comme la synthèse performative met en œuvre les modalités auditive, visuelle et kinesthésique, l'apprentissage suscite et tire parti du renforcement multimodal.

2 Cantor Digitalis

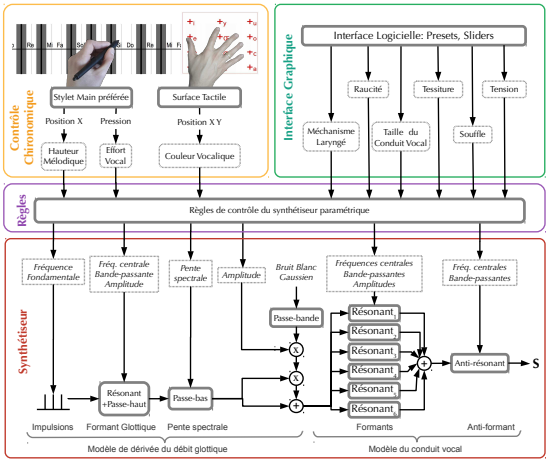

Figure 1 : Schéma de principe de Cantor Digitalis. En haut à gauche : contrôle gestuel ; en haut à droite : interface utilisateur ; au centre règles de synthèse ; en bas : synthétiseur paramétrique à formants avec source glottique spectrale (d'Alessandro et al. 2019).

Une première famille d'instruments vocaux a abouti au Cantor Digitalis, un synthétiseur par règles à formants piloté par le geste manuel (Feugère et al. 2017) dont le schéma de principe est donné Figure 1. La main préférée dessine l'intonation (dimension X) et l'effort vocal (pression), à l'aide d'un stylet sensible sur une tablette graphique. L'intonation varie du grave à l'aigu (de gauche à droite), sur un ambitus donné. La pression du stylet contrôle l'effort vocal, en passant par un modèle spectral élaboré de l'onde de débit glottique. La main non-préférée définit la voyelle (orale uniquement) suivant les deux dimensions d'aperture et d'antériorité, à l'aide d'un doigt. De nombreuses possibilités de préréglages ou de jeu avec ce modèle vocal permettent la production de sons très variés : taille apparente du conduit vocal, typologie vocale, variation de la source glottique, jeux de voyelles et de formants. Les modes d'affectation des entrées gestuelles aux paramètres de synthèse sont nombreux, et permettent différents modes de jeux.

Le synthétiseur adopte une structure série/parallèle pour les formants et anti-formants, avec le modèle spectral linéaire de source glottique (Perrotin et al., 2021) conçu pour permettre des variations expressives perceptivement pertinentes avec peu de paramètres. Un ensemble de règles régit les rapports entre données gestuelles et paramètres de synthèse : règles sur l'effort vocal, sur la correspondance entre voyelles et paramètres formantiques, sur la qualité vocale. L'instrumentiste dispose d'un instrument mélodique continu, comme le Violon, l'onde Martenod, ou de Théremine, sonnant comme une voix. Le jeu porte sur la hauteur mélodique, le rythme, l'effort vocal, la voyelle, mais sans consones. Cantor Digitalis est avant tout destiné à la musique dans la mesure où il ne produit que des voyelles : il ne parle pas. Une extension du Cantor Digitalis, le Digitartic permet de jouer des consonnes grâce à la main non-préférée. Mais la très courte durée des consonnes et la précision demandée semblent au-delà des possibilités motrices de la main et des doigts, rendant le contrôle très difficile, au détriment de la qualité segmentale (Feugère et d'Alessandro, 2017).

3 Voks

Pour parler, synthétiser n'importe quel énoncé, une approche radicalement différente de la synthèse est nécessaire. Une seconde famille d'instruments vocaux a abouti à Voks (Delalez et d'Alessandro, 2017, Locqueville et al., 2020), synthèse vocale performative par échantillonnage, dont le schéma de principe est porté Figure 2. Avant de pouvoir jouer un texte, il faut acquérir les échantillons sonores correspondant au contenu segmental (par enregistrement, synthèse à partir du texte, exploration de base de données) et les enrichir de points de contrôle rythmique. Une fois ce matériau disponible, le rythme, l'intonation, l'effort vocal sont contrôlés par les gestes manuels. L'intonation et l'effort vocal sont contrôlés de la même manière que pour le Cantor Digitalis : la position du stylet contrôle l'intonation, et la pression l'effort vocal. Comme alternative spectaculaire, le thérémine peuvent remplacer la tablette pour l'intonation et l'effort vocal : la main droite contrôle l'intonation en fonction de la distance à l'antenne verticale et la main gauche l'effort vocal en fonction de la distance à l'antenne horizontale. Le rythme peut être pilotés par plusieurs modes de contrôle du temps de synthèse. Dans le mode « défilement » ou « frottement » (scrub) la pointe du stylet dans la dimension Y définit la position de lecture de l'échantillon. On peut ainsi jouer le son à l'endroit ou à l'envers, sauter d'une position à l'autre. Dans le mode « vitesse » (speed), l'appui du stylet sur la tablette déclenche la lecture des échantillons, avec une vitesse qui dépend de la position Y du stylet. Une vitesse nulle, au centre de la tablette lit les échantillons à leur vitesse nominale d'enregistrement ; une vitesse positive accélère la lecture alors qu'une vitesse négative ralentit la lecture. Ces deux modes sont surtout utiles pour la pratique musicale.

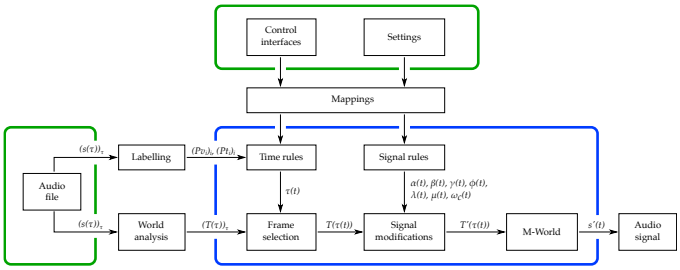

Figure 2 : Schéma de principe de Voks. En haut le contrôle gestuel et l'interface utilisateur ; au centre le synthétiseur par échantillonnage ; à gauche la base de données d'échantillons étiquetés (Locqueville et al. 2020)

Pour le contrôle prosodique du rythme de parole il faut repérer des unités rythmiques, ici les syllabes, à l'aide de points de contrôle. Des marques sont apposées sur tous les noyaux syllabiques (ainsi que sur les silences suprasegmentaux) et des contre-marques sont apposée entre les noyaux vocaliques, dans les groupes consonantiques ou silences formés par les attaques et codas des syllabes). Un bouton à deux états, ou bien un ou deux potentiomètres (comme des pédales) permettent de séquencer à volonté le flux syllabique, et ainsi de produire le rythme désiré. Les ancres rythmiques des marques et contre-marques peuvent être apposées sur d'autres unités que les syllabes : les pieds par exemple pour les langues à rythme accentuel plutôt que syllabique. La qualité de la source vocale, la longueur apparente du conduit vocal sont définis par des paramètres supplémentaires. Pour jouer Voks, il faut donc commencer par sélectionner le contenu segmental désiré (phrase, mot, texte …). Ensuite le jeu ressemble à celui du Cantor Digitalis, sauf que le déroulement temporel est piloté par un des modes de jeu rythmique. Le jeu par points de contrôle biphasique est le plus riche expressivement. Cet

instrument ne permet pas de jouer un texte sans le préparer. La question de la planification et de la réalisation gestuelle d'un énoncé quelconque au moment même de la synthèse reste encore ouverte.

4 Chironomie de l'intonation

Le contrôle de l'intonation par une interface gestuelle est une tâche exigeante. Une interface de type clavier, avec des degrés, ne convient pas du tout : elle impose en effet un système d'échelle musicale fixe à l'intonation vocale, oblitérant toute possibilité d'expression. Une interface de contrôle continu est nécessaire. La tablette graphique s'est révélée particulièrement efficace pour imiter l'intonation vocale, en effectuant des gestes proches de ceux du dessin ou de l'écriture. Cependant, avec l'imitation gestuelle, tous les détails des courbes intonatives ne sont pas reproduits, en particulier la micro-prosodie due aux effets acoustiques des consonnes sur la source vocale. La chironomie de l'intonation est donc un procédé de stylisation mélodique.

4.1 Stylisation chironomique de l'intonation

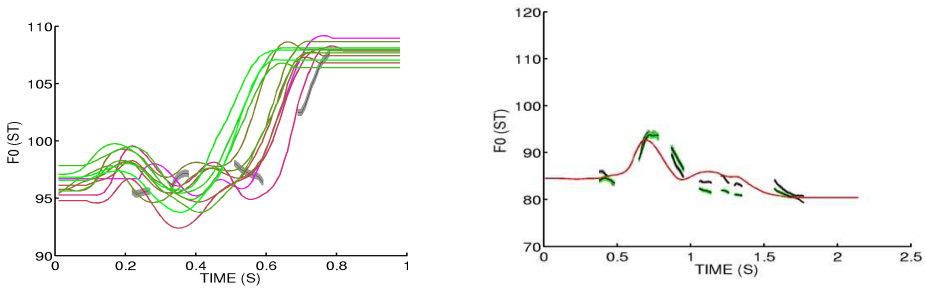

Figure 3. Imitation chironomique de l'intonation :à gauche, exemples de tracés gestuels, phrase de 4 syllabes. À droite, meilleures imitations : gestuelle, en trait continu, vocale, en trait vert, de la phrase de référence de 6 syllabes, en trait gris (d'Alessandro et al. 2011).

La précision et la validité perceptive de la stylisation chironomique de l'intonation a été étudiée à l'aide d'un paradigme d'imitation (d'Alessandro et al. 2011). La tâche des 15 sujets est de reproduire l'intonation d'un ensemble de phrases (phonétiquement équilibrées, de 1 à 9 syllabes, un locuteur, une locutrice) à l'aide d'un stylet sur une tablette graphique en utilisant le système Calliphony (Le Beux et al., 2007). Le déroulement temporel des stimuli n'est pas modifié et il n'y a aucun repère visuel sur la tablette graphique. Les sujets doivent aussi imiter au mieux avec leur propre voix les phrases de référence. La Figure 3 montre des exemples de tracé chironomique, ainsi que la courbe intonative de référence et une courbe intonative obtenue par imitation vocale. Des mesures (distance des moindres carrés, dissemblance de programmation dynamique) entre les copies de synthèse, de voix naturelles et les phrases de référence permettent de quantifier la qualité de l'imitation intonative. Les distances obtenues avec la stylisation chironomique sont comparables à celle obtenues pour l'imitation vocale, avec un léger avantage pour l'imitation vocale. Un test de discrimination permet d'évaluer l'équivalence perceptive des copies gestuelles et vocale. L'intonation chironomique, dans un grand nombre de cas, est perceptuellement indiscernables ou quasiment indiscernables de l'intonation naturelle, en particulier pour la voix féminine. Ainsi les mouvements de la main sont perceptuellement équivalents aux mouvements intonatifs dans cette tâche de synthèse prosodique. La chironomie est un procédé efficace de stylisation de l'intonation. Les courbes intonatives stylisées montrent des échelles de temps de l'ordre de la syllabe : les détails micro-prosodiques gommés par la stylisation semblent perceptivement négligeables. Les contours chironomiques sont continus sur la

durée d'un groupe de souffle, sans levée de stylet ou rupture dans les silences articulatoires. Les silences prosodiques par contre entraînent des interruptions de contour par levée du stylet.

4.2 Justesse et précision du chant chironomique

La justesse et la précision mélodique du chant chironomique a été étudiée à l'aide du Cantor Digitalis (d'Alessandro et al. 2014). Un masque reprenant une disposition de clavier est posé sur la tablette graphique. La position juste des notes ne correspond pas au centre du dessin de la touche mais à la ligne gauche de la touche. L'intonation est continue sur la tablette, donc le centre de la touche correspond au quart de ton entre les deux notes matérialisées par les deux lignes aux bords de la touche. La justesse (moyenne) et la précision (écart type) des mélodies produites, par rapport aux notes musicales, sont mesurées dans trois expériences pour des groupes de 20 et 28 sujets. La tâche des sujets est de produite des intervalles musicaux et de courtes mélodies, à différents tempi, en utilisant la chironomie avec retour auditif, la chironomie muette (sans retour auditif), la chironomie aveugle (sans référence visuelle) et leur propre voix. La condition de chironomie muette permet de faire la part des modalités audio-visuo-motrice et visuo-motrice dans la tâche. La Figure 4 montre un exemple des courbes mélodiques obtenues pour la chironomie et la voix ainsi que. Les résultats montrent (Figure 4, à droite) la justesse et la précision obtenues par l'ensemble des sujets avec le contrôle chironomique. Plusieurs sujets ont produit de meilleurs résultats avec la chironomie qu'en chantant, d'autres sujets ont obtenu des résultats comparables. Pour la chironomie, la précision moyenne des notes est inférieure à 12 cents et la précision moyenne des intervalles est inférieure à 25 cents pour tous les sujets. La condition de chironomie muette indique que les compétences acquises pour l'écriture et le dessin sont pleinement mises en œuvre pour le chant chironomique, mais que le retour auditif améliore la précision des intervalles. Un intervalle est en effet visuellement plus difficile à apprécier qu'une position absolue. La condition de chironomie aveugle indique que la référence visuelle renforce considérablement la justesse et la précision. Cette étude valide l'utilisation musicale des instruments vocaux.

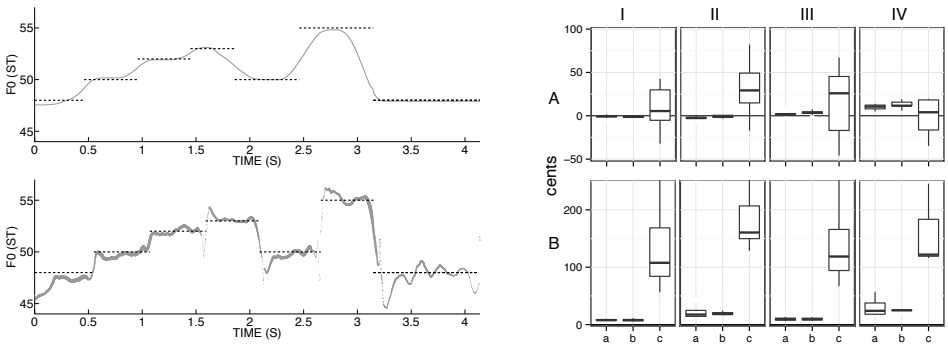

Figure 4. Mélodies chironomique (en haut à gauche) et vocale (en bas à gauche), avec les notes de référence (traits pointillés). À droite, distributions de la justesse (A) et de la précision (B) pour les 3 modalités (a : chironomie, b : chironomie muette, c : voix) selon 2 types de tâches (intervalles ou mélodies) et de mesures (sur les notes ou intervalles) : I intervalles/notes, II mélodie/notes, intervalles/intervalles III, mélodies/intervalles IV (d'Alessandro et al. 2014).

4.3 Les modalités sensori-motrices dans la chironomie de l'intonation

Le contrôle chironomique de l'intonation implique plusieurs modalité sensori-motrices (audition, vision, proprioception ou kinesthésie) ainsi qu'un processus de planification, de comparaison et de mémorisation. Le contrôle vocal est intéroceptif, mais il est kinesthésique et extéroceptif pour la synthèse performative. L'imitation mise en œuvre dans les évaluations précédentes est une tâche complexe asynchrone, selon la séquence suivante : (1) Écoute du stimulus acoustiques, en se concentrant sur le contour d'intonatif ; (2) Mémorisation du contour, stocké sous forme de trace acoustique, phonétique ou motrice ; (3) Planification et réalisation du geste chironomique ou vocal équivalent, en utilisant cette trace mnésique ; (4) Comparaison de l'énoncé produit avec l'original, ou sa trace mnésique immédiate. (5) Le processus est répété jusqu'à la satisfaction du sujet. La première étape est contrainte par les seuils perceptifs de hauteur (absolus, différentiel, avec intégration temporelle). Ces seuils entraînent un lissage du contour, et la perte probablement à ce stade de la micro-prosodie. La trace mnésique résultante de ces contours intonatifs lissés est sans doute représentés par des valeurs cibles ancrées à moments précis ou par une représentation globale des trajectoires. La planification et la réalisation du geste chrironomique est un acte moteur dont la cinématique et la dynamique suivent des lois semblables à celles qui régissent les gestes d'écriture. Notons que les gestes spécifiques utilisés par les différents sujets pour accomplir la même tâche n'ont pas été analysés en détail pour le moment. Certains sujets utilisent des mouvements plutôt circulaires, d'autres plutôt linéaires. La forme spécifique du dessin ne semble pas très importante, tant que les hauteurs sont atteintes avec un timing correct. Enfin, le processus de décision, pour accepter un contour ou pas, est également basé sur la mémoire du contour d'origine ou sur une représentation mentale de ce contour. Puisque audition, vision et proprioception interviennent dans la chironomie, le rôle joué par chacune de ces modalités mérite d'être étudié. La prééminence de la modalité visuelle sur la modalité kinesthésique est connue dans la littérature. Nos travaux sur la chironomie dans le chant ont suggéré une possible prééminence de la vue sur l'écoute dans le jeu du Cantor Digitalis.

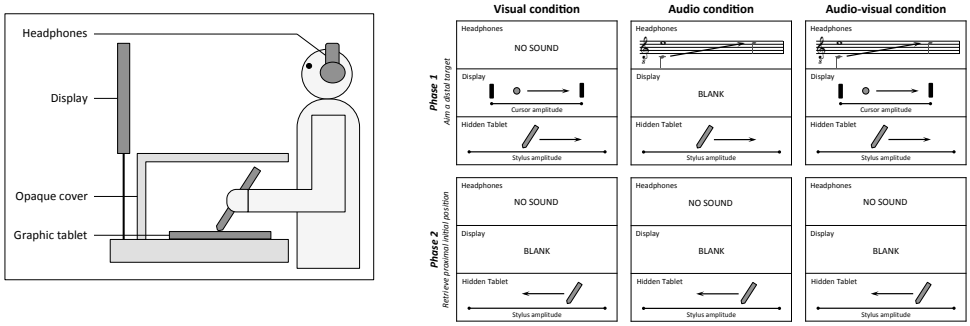

Figure 5. À gauche, dispositif expérimental pour l'étude des interférences visuo-audio-proprioceptives dans la tâche chironomique. À droite, les deux phases de la tâche chironomique selon les trois conditions : visuelle, audio et audiovisuelle ((Perrotin et d'Alessandro, 2016).).

Nous avons donc conduit une étude spécifique sur ce point, afin d'examiner les possibles interférences entre modalités dans les tâches qui ressemblent au jeu musical, avec des notes, ce qui dans le domaine des interfaces humains-machine revient à des tâches d'acquisition de cibles visuelles, audios et audiovisuelles (Perrotin et d'Alessandro, 2016). Les expériences utilisent un paradigme de réplication de mouvements, en déplaçant un curseur visuel sur un écran et/ou une référence de hauteur mélodique par le contrôle de la position du stylet sur une tablette graphique cachée de la vue du sujet (Figure 5,

à gauche). La tâche est divisée en deux phase. Phase Aller : atteindre une cible visuelle sur l'écran, une cible auditive (hauteur mélodique, intervalle mélodique), ou une cible audio-visuelle (combinant les deux). Phase Retour : réplication en sens inverse du geste en revenant à la position initiale du stylet. Dans la phase de retour, les références visuelles et audios sont supprimées, la tâche s'appuie donc sur la proprioception uniquement (comme montré en Figure 5, à droite). Différents gains (amplitudes) perturbent la relation entre les mouvements du stylet, les mouvements du curseur visuel et les mouvements de hauteur mélodique. L'analyse des résultats porte sur les écarts d'amplitude des mouvements du stylet entre phases aller et retour selon les conditions de perturbation. Ces amplitudes varient en fonction des gains visuels, audios et audiovisuels perturbés. Perturber la modalité visuelle entraine les plus grandes variations d'écart. Cela indique que la modalité interfère plus avec la modalité motrice. Dans la condition audiovisuelle, la vision domine sur l'audition. Ce résultat confirme la contribution de la vision pour la chironomie de l'intonation. L'aide visuelle facilite le jeu, au moins dans la phase d'apprentissage de l'instrument. Il est notoire que des instruments sans référence visuelle comme le thérémine sont plus difficiles à maîtriser que par exemple les instruments à clavier, au moins au commencement. Cette étude confirme une organisation spatiale interne sous-jacente de la perception de hauteur (gauche/droite et bas/haut correspondant à grave/aigue).

5 Chironomie du rythme

5.1 Position et débit : contrôle continu du rythme

La chironomie de l'intonation utilise des gestes continus sur une surface ou dans l'espace. Cette approche peut s'appliquer au contrôle du temps de synthèse. Une première méthode contrôle directement la position temporelle de lecture de l'échantillon par la position d'un stylet sur une ligne temporelle qui représente la forme d'onde. Le temps peut être lu à l'endroit ou à l'envers, tout comme dans le scratch des platinistes. Produire de la parole rythmiquement juste et expressive est cependant difficile, à cause de la rapidité des mouvements nécessaires et du manque de points de repère. La seconde méthode utilise la vitesse, ou débit, plutôt que la position. Le débit peut être diminué ou augmenté en ralentissant ou accélérant le flux temporel. Un tel contrôle abouti à des effets spectaculaires mais n'est pas assez souple et rapide pour, par exemple, accentuer une seule syllabe.

5.2 Arsis et Thesis : contrôle biphasique du rythme

Le contrôle chironomique du rythme demande donc une approche différente de celle de l'intonation. Le déroulement du temps en musique comme en parole, le rythme prosodique ou musical, s'appuie sur la récurrence d'évènements temporels. Alors que dans l'histoire de la musique occidentale les théories rythmiques sont apparues plus tard que les théories mélodiques, des théories rythmiques sont apparues très tôt pour la parole, la scansion des vers par exemple dans les métriques grecque ou latine. L'unité rythmique de base dans la musique est la note, avec toutes les nuances voulues (la question des ornements par exemples). Dans des langues comme le français ou le latin, l'unité rythmique est la syllabe, alors que dans des langues comme l'anglais l'unité rythmique est le pied (groupe de syllabes portant l'accent) et dans des langues comme le japonais, la more. Du point de vue perceptif, la syllabe est souvent considérée comme un unique appui rythmique. C'est l'antique ictus rythmique, aujourd'hui caractérisé par son centre perceptif ou P-Center. Pour la chironomie du rythme en synthèse vocale performative, le rythme doit être considéré du côté de la motricité plutôt que du côté de la perception. Un exemple de visualisation motrice du rythme est la chironomie grégorienne. Cette méthode de direction chorale du plain-chant utilise des gestes manuels pour dessiner les contours rythmiques sur la base de neumes, groupes accentuels d'une ou quelques syllabes. Les mouvements

de la main indiquent des successions d'arsis (levée) et thesis (posé), terminologie provenant de la métrique grecque. Par analogie avec la locomotion, chaque syllabe ou groupe e syllabe permet d'avancer d'un pas, puisque la marche est une succession d'appuis du pied au sol (thesis) et d'élan (arsis) vers le pas suivant. En musique se retrouve la même différence entre une notation rythmique monophasique, qui représente une note pour chaque valeur rythmique, et le mouvement rythmique biphasique nécessaire pour tout action, puisque jouer une note comprend toujours un appui et une levée. Le contrôle chironomique du rythme est donc biphasique dans nos instruments vocaux (Delalez et d'Alessandro, 2017-2). Il s'inspire de la structure de la syllabe, attaque, noyau vocalique, coda, et du mouvement mandibulaire sous-jacent. Comme dans la théorie *frame-content* de Mc Neilage, un énoncé est considéré comme une suite d'ouvertures et de fermetures de la mandibule. L'ouverture correspond aux noyaux vocaliques, et la fermeture aux espaces inter-vocaliques formés par le silence, l'attaque, la coda. Le contrôle rythmique utilise donc des dissyllabes. L'arsis correspond à la constriction, cluster consonantique (temps faible de la syllabe) et la thesis correspond à l'ouverture maximale ou noyau vocalique (temps fort de la syllabe). Ce contrôle biphasique des unités rythmiques est effectué généralement par la main non dominante à l'aide d'une interface à deux états, comme un bouton bistable. Des interfaces de contrôle continu comme le potentiomètre (par exemple une pédale) ou le double potentiomètre (deux pédales) peuvent être utilisées pour un contrôle temporel plus fin. La figure 6 donne un exemple de points de contrôle et de trace des geste chironomique (bouton bistable, potentiomètres).

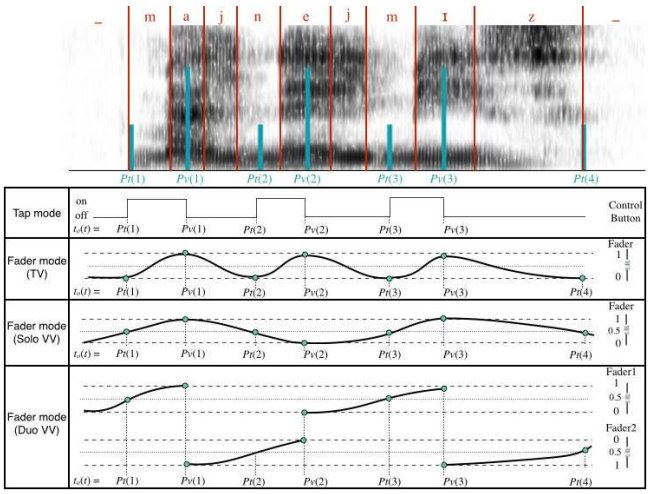

Figure 6. Chironomie du rythme par points de contrôle. Les phonèmes au-dessus du spectrogramme permettent de définir les syllabes. Les barres bleues représentent les points de contrôle, thesis (longues) et arsis (courtes). Le tableau montre le contrôle par bouton bistable, potentiomètre et double potentiomètre (Delalez et d'Alessandro, 2017)..

5.3 Points de contrôle rythmiques

Le placement des points de contrôle rythmique suit des règles empiriques : vers la fin du noyau vocalique pour la thesis, qui correspond à l'appui, la battue rythmique ; dans la tenue ou à la fin du cluster consonantique pour l'arsis, qui correspond à l'élan, anacrouse avant le noyau suivant, ou désinence pour la syllabe finale. Le calcul et la mise en place automatique des points de contrôle en

utilisant une segmentation par disyllabes donne un résultat satisfaisant. Toutes les langues partagent la notion de syllabe, mais pour la commodité du jeu chironomique, des points de contrôle accentuels ou moraïques peuvent être utilisés pour les langues dont les schémas rythmiques utilisent préférentiellement ces unités plus grandes ou plus petites que la syllabe.

La qualité du contrôle chironomique du rythme par tapotage a été formellement évaluée à l'aide d'un paradigme d'imitation prosodique (Delalez et d'Alessandro, 2017-2). La tâche des 8 sujets est de reproduire la prosodie d'un ensemble phrases (8 phrases de 2 à 9 syllabes, un locuteur et une locutrice) à l'aide d'une tablette graphique pour la mélodie et d'un bouton bistable pour le rythme syllabique. La mesure utilisée est la durée des unités inter-vocaliques, c'est-à-dire les durées entre les noyaux vocalique. En moyenne, pour tous les sujets et phrases, la différence de durée inter-vocalique entre l'original et la reproduction chironomique est de l'orgue de 20 ms. Cette méthode de contrôle rythmique montre donc une grande précision. Il faut remarquer que le débit syllabique produit avec un doigt sur un bouton bistable est limité à environ 6 ou 8 syllabes par seconde. C'est inférieur au débit d'une parole rapide. L'utilisation de plusieurs doigts ou de point de contrôle sur des unités plus grandes, comme les accents, permet un débit plus rapide. Enfin, contrairement au cas de l'intonation, des études antérieures ont montré que l'audition peut dominer la vision dans des tâches rythmiques.

6 Chironomie de l'effort vocal

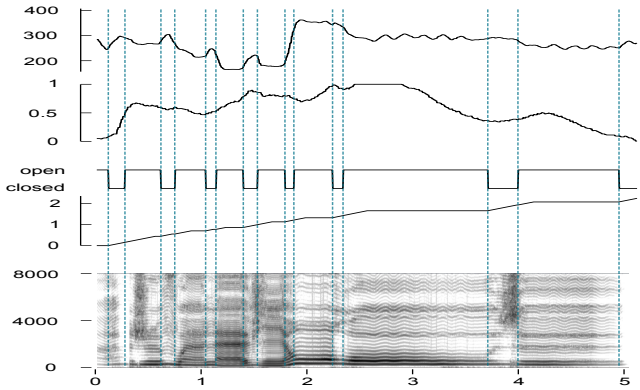

Figure 7. Les paramètres de jeu d'un énoncé : mélodie, effort vocal, rythme, déformation du temps et spectrogramme correspondant (Locqueville et al. 2020).

Une troisième dimension prosodique, la qualité vocale, est essentielle pour l'expression vocale. L'effort vocal est le paramètre dynamique employé à cet effet dans Cantor Digitalis et Voks. Une approche spectrale qui combine intensité, pente spectrale, position du formant glottique permet de rassembler dans une seule dimension le contrôle du geste laryngé (d'Alessandro, 2006, d'Alessandro et al. 2006). Le phonétogramme est utilisé pour prendre en compte les dépendances entre la hauteur mélodique et l'intensité vocale, afin de produire des variations réalistes d'effort vocal. La perception est étonnamment sensible à la coordination entre mélodie, rythme et effort vocal pour évaluer la qualité « humaine » de l'expression dans une performance. La figure 7 montre ce type de coordination : les trois courbes du haut correspondent aux traces gestuelles de l'intonation, de l'effort vocal et du rythme syllabique d'une phrase chantée. La quatrième courbe représente la déformation du temps apportée pas le tapotage : les 2 secondes de l'enregistrement original durent 5 secondes dans le signal de synthèse.

7 Conclusion

Le Cantor Digitalis et Voks sont de nouveaux instruments musicaux, qui ont été joués en concert à de nombreuses reprises (d'Alessandro et al., 2019). Ils ont reçu une reconnaissance internationale en remportant à deux reprise le prix Guthman pour les nouveaux instruments de musique (2015 et 2022). Ces instruments participent aux recherches et créations qui s'intéressent à la voix dans les musiques électroniques ou contemporaines et le théâtre. Cette famille d'instruments de musique numériques, permet d'envisager de nouvelles applications en pédagogie de l'intonation. La composante gestuelle, tant dans la production que dans la perception vocale, peut aider à l'apprentissage : faire ou percevoir un geste renforce l'acquisition du schéma intonatif correspondant. Pour cela une interface mobile (téléphone) qui permet le contrôle gestuel en temps réel de l'intonation des phrases produites a été développée (Xiao X. et al., 2021). Les mélodies et le timing d'une phrase cible peuvent être modifiés en traçant des contours mélodiques sur l'écran tactile d'une tablette mobile. Des locuteurs natifs et non natifs avaient pour tâche d'imiter la prononciation de phrases française ambiguës prosodiquement en utilisant leur voix et l'interface, avec et sans guide visuel, comme le montre la Figure 8. La comparaison des courbes mélodique résultantes avec le contour de référence et l'évaluation perceptive des énoncés synthétisés suggèrent que pour les locuteurs non natifs et natifs, l'imitation à l'aide d'un guide visuel est comparable en précision à l'imitation vocale. Le contrôle du timing semple par contre difficile. L'application de la chironomie à la suppléance vocale, avec l'espoir d'utiliser le geste manuel comme substitut du geste laryngé dans un larynx artificiel est actuellement en cours d'exploration.

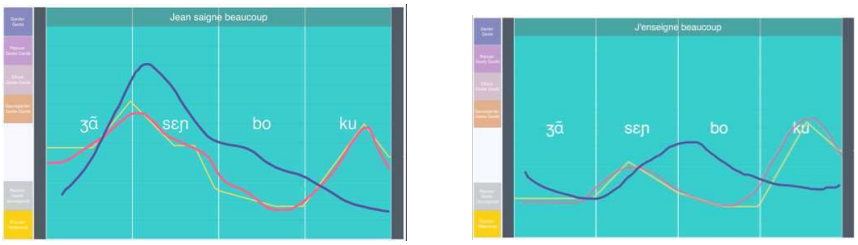

Figure 8. Interface de contrôle Gepeto pour l'apprentissage de l'intonation en L2. Phrases ambigües prosodiquement : j'enseigne beaucoup/Jean saigne beaucoup ((Xiao X. et al., 2021).

Pour les aspects segmentaux, la rapidité des gestes sous-jacents interdit pour le moment d'envisager des systèmes généraux qui soient effectivement utilisables. Le parcours chironomique dans un espace vocalique est compatible avec la vitesse des doigts et des mains : avec la Cantor Digitalis on joue de façon réaliste et précise les voyelles disposées sous forme trapèze sur une surface. Par contre l'ajout de consonnes dans ces trajets vocaliques est difficile, car les gestes consonantiques sont trop rapides. Jusqu'à présent, aucune solution pour le contrôle chironomique de l'articulation au niveau des phonème n'a été trouvée, malgré nos efforts. La qualité reste faible, en termes d'intelligibilité et de son, et le contrôle difficile. Le geste articulatoire demande la coordination complexe et précise de nombreux organes, avec des vitesses qui restent hors de portée du geste de membres comme la main ou le pied. D'autres pistes sont explorées, mais il est trop tôt pour en faire état.

Remerciements

Ces recherches ont bénéficié du soutien des projets ChaNTeR (ANR-13-CORD-0011), GEPETO (ANR-19-CE28-0018), ARS (ANR-19-CE38-0001) et SMAC. (FEDER IF0011085, Union Européenne et Région île de France).

Références

D'ALESSANDRO, C (2001) « 33 ans de synthèse de la parole à partir du texte : une promenade sonore (1968-2001) ». Traitement Automatique des Langues (TAL), Hermès, Vol. 42 No 1 (avec un disque compact audio de 62 mn), p. 297-321.

D'ALESSANDRO, C., D'ALESSANDRO, N., LE BEUX, S., SIMKO, S., ÇETIN F. ET PIRKER, H. (2005), « The Speech Conductor: Gestural Control of Speech Synthesis » proc. eNTERFACE 2005, Mons, Belgium, pp. 52-61.

D'ALESSANDRO, C. (2006) « Voice source parameters and prosodic analysis », in Method in Empirical Prosody Research, Edited by Stefan Sudhoff, Denisa Leternovà, Roland Meyer, Sandra Pappert, Petra Augurzky, Ina Mleinek, Nicoale Richter, Johannes Schliesser, Walter de Gruyter, Berlin, New York, , pp 63-87.

D'ALESSANDRO, N., D'ALESSANDRO, C., LE BEUX, S. ET Doval B. (2006) « Real-time CALM Synthesizer New Approaches in Hands-Controlled Voice Synthesis », Proc. NIME06, Paris, France, pp. 266-271.

LE BEUX, S., RILLIARD, A. ET D'ALESSANDRO, C. (2007) "Calliphony: A real-time intonation controller for expressive speech synthesis", 6th ISCA Workshop on Speech Synthesis (SSW-6), Bonn, Germany, August 22-24, 2007, p. 345-350.

D'ALESSANDRO, C., RILLIARD, A. ET LE BEUX, S. (2011) « Chironomic stylization of intonation » J. Acoust. Soc. Am., 129(3), 1594-1604.

D'ALESSANDRO, C., FEUGÈRE, L., LE BEUX, S., PERROTIN, O. ET RILLIARD, A. (2014) , « Drawing melodies : evaluation of chironomic singing synthesis », J. Acoust. Soc. Am. 135 (6), 3601-3612.

FEUGERE, L. ET D'ALESSANDRO, C. (2015) « Synthèse à partir du geste de la voix chantée: instruments Cantor Digitalis et Digitartic », Traitement du Signal, 32 (4), 417-442

PERROTIN, O. ET D'ALESSANDRO, C. (2016) « Seeing, listening, drawing: interferences between sensorimotor modalities in the use of a tablet musical interface », ACM Trans. on Applied Perception, 14(2).

FEUGÈRE, L., D'ALESSANDRO, C., DOVAL B. ET O. PERROTIN, O. (2017) « Cantor Digitalis: Chironomic Parametric Synthesis of Singing », J. Audio Speech Music Proc. (2017) 2017: 2. https://doi.org/10.1186/s13636-016-0098-5

DELALEZ, S. ET D'ALESSANDRO, C. (2017) « Vokinesis : syllabic control points for performative singing synthesis », Proc. NIME 2017, Copenhagen, Denmark 198-203.

DELALEZ, S. ET D'ALESSANDRO, C. (2017-2) « Adjusting the Frame: Biphasic Performative Control of Speech Rhythm », Proc. INTERSPEECH 2017, Stockholm, Sweden, , 864-868.

D'ALESSANDRO, C., DELALEZ, S., DOVAL, B., FEUGERE, L., PERROTIN, O. (2019) « Les instruments chanteurs », *Acoustique et Techniques*, vol 89, p. 36-43.

LOCQUEVILLE, G., D'ALESSANDRO, C., DELALEZ, S., DOVAL, B., XIAO XIAO (2020) « Voks: digital instruments for chironomic control of voice samples », Speech Communication, 120, p. 97-113.

XIAO, X., AUDIBERT, N., LOCQUEVILLE, G., D'ALESSANDRO, C., KUHNERT, B., PILLOT-LOISEAU, C. (2021) Prosodic Disambiguation Using Chironomic Stylization of Intonation with Native and Non-Native Speakers. Proc. Interspeech 2021, 516-520.

PERROTIN, O., FEUGÈRE, L. ET D'ALESSANDRO, C. (2021) « Perceptual equivalence of the Liljencrants-Fant and linear-filter glottal flow models », J. Acoust. Soc. Am. 150 (2), August 2021, p. 1273–1285.