Contrôle gestuel du modèle source/filtre de production de la voix

Christophe d'Alessandro, Sylvain Le Beux, Albert Rilliard

10ème Congrès Français d'Acoustique
Lyon, 12-16 Avril 2010

Contrôle gestuel du modèle source/filtre de production de la voix

Christophe d'Alessandro, Sylvain Le Beux, et Albert Rilliard

LIMSI-CNRS BP 133 F91403 Orsay {cda, slebeux, rilliard}@limsi.fr

En utilisant des cadres théoriques élaborés par Fant (le modèle source/filtre de production de la parole, le modèle LF de la source) cette communication présente nos travaux récents en contrôle gestuel temps réel de la synthèse vocale. Plusieurs interfaces de contrôle gestuel sont développées pour piloter un modèle de source glottique et de conduit vocal, avec ou sans retour d'effort (joystick, tablette graphique, méta-instrument, bras haptique). On montre que ces instruments permettent de développer une analogie entre gestes manuels et mouvements prosodiques. Les dimensions perceptives de la source vocale sont ensuite étudiées en utilisant un paradigme d'analyse par la synthèse : on recherche les associations de paramètres et les gestes les plus efficaces pour reproduire des exemples de vocalisations expressives. Enfin, ces instruments s'appliquent également à la synthèse de voix chantée en musique électronique.

1 Introduction

Les contributions de Gunnar Fant à la phonétique, à l'acoustique, au traitement de la parole et aux technologies vocales sont nombreuses et fondamentales [1]. Les travaux présentés dans cette communication explorent des chemins ouverts par les travaux de Fant pour la synthèse vocale, en particulier le modèle source/filtre [2], la source vocale [3][4][5][7] et l'intonation.

Le propos est d'aborder l'expression vocale, en voix chantée et en voix parlée. L'expression relève plus du contrôle et de la dynamique des paramètres vocaux que de la qualité intrinsèque de la synthèse segmentale. En ce sens, et à l'inverse des méthodes basées sur de grands corpus, il est avantageux d'utiliser les paramètres simples mais explicites du modèle source/filtre.

Les outils développés dans le cadre de l'informatique musicale permettent aujourd'hui de calculer le modèle source/filtre en temps réel, et donc d'aborder la problématique du contrôle et de la dynamique des paramètres. Plusieurs instruments ont été développés depuis le début ne nos recherches sur les instruments vocaux à contrôle gestuels, recherches initiées au workshop Enterface' 05 [8].

Le contrôle gestuel d'un modèle de synthèse de la source glottique a été développé en 2005, à la suite de nos études sur la représentation spectrale de la source. Ce modèle a d'abord été piloté par un gant numérique ou un clavier. Afin de permettre un contrôle très précis de la mélodie, par exemple pour la voix chantée, un instrument utilisant une tablette graphique a ensuite été développé en 2006. En 2007, cet instrument a été augmenté d'un joystick pour contrôler les voyelles et la qualité vocale. Il a permis d'étudier la précision de contrôle gestuel de l'intonation, dans des tâches de stylisation. En 2008-2009, le modèle de synthèse a été porté sur une interface riche de 54 capteurs, le méta-instrument (développé pour la musique électronique). Le contrôle haptique de la synthèse vocale (c'est à dire l'utilisation de retour d'effort actif pour signaler des passages et des limites dans le fonctionnement vocal) a également été mis en œuvre. Les recherches actuelles portent sur la synthèse de consonnes, sur l'intonation, et sur les effets de qualité de voix dans l'expression des attitudes.

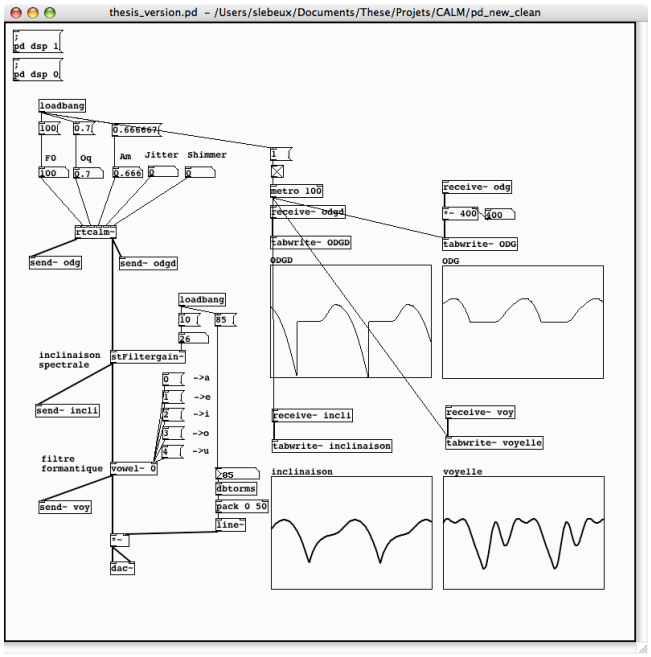

Figure 1 : Exemple de programme en Pure Data pour la synthèse de voyelles avec une source vocale élaborée. De gauche à droite et de haut en bas, sont affichées l'onde de débit glottique dérivée (ODGD) et l'onde de débit glottique, puis l'ODGD avec une inclinaison spectrale et enfin le résultat après filtrage par un filtre de conduit vocal.

2 Dimensions de la synthèse vocale

La production vocale peut en première approximation suivre un schéma source/filtre. La source, ou phonation est responsable pour une grande part de la qualité vocale. Le filtre permet d'articuler les voyelles, configurations relativement stables, et les consonnes, mouvements rapides des articulateurs.

Les paramètres à contrôler dans un modèle de synthèse peuvent se situer à plusieurs niveaux : les paramètres physiques ou acoustiques, comme la pression sub-glottique, la tension des plis vocaux, la position des articulateurs ; les paramètres acoustiques fonctionnels, comme l'onde de débit glottique, les formants vocaliques ; les paramètres phonétiques fonctionnels, comme l'effort vocal ou la proéminence. Les instruments décrits ici relèvent de ces deux dernières catégories.

Les paramètres du filtre sont représentés par les amplitudes et les fréquences centrales des premiers formants. De façon alternative, un ensemble de voyelles prédéfinies peut également être utilisé. Pour l'instant, la synthèse des transitions rapides (consonnes) n'est pas abordée.

Les paramètres de la source vocale comprennent la fréquence fondamentale, l'amplitude, le quotient ouvert, l'asymétrie et la vitesse de fermeture (ou pente spectrale) de la source. Une source de bruit additif permet de simuler l'aspiration. Les perturbations structurelles de la source (jitter et shimer) sont obtenues en rendant aléatoire les positions et amplitudes des périodes glottiques.

Ces paramètres de source peuvent de façon alternative être regroupés suivant plusieurs dimensions : la dimension d'effort vocal (regroupant amplitude, asymétrie et pente spectrale), la dimension de tension vocale (bruit additif et quotient ouvert), la dimension de bruit (additif et structurel), les mécanismes laryngés (ou « registres » de la voix) et bien entendu la dimension mélodique.

Suivant les instruments, ces diverses dimensions seront affectées aux différents capteurs et donc aux différents modes de jeu.

Les différents instruments sont programmés en Pure Data et/ou en Max/MSP. La figure 1 montre un exemple de programme pour la synthèse de voyelles.

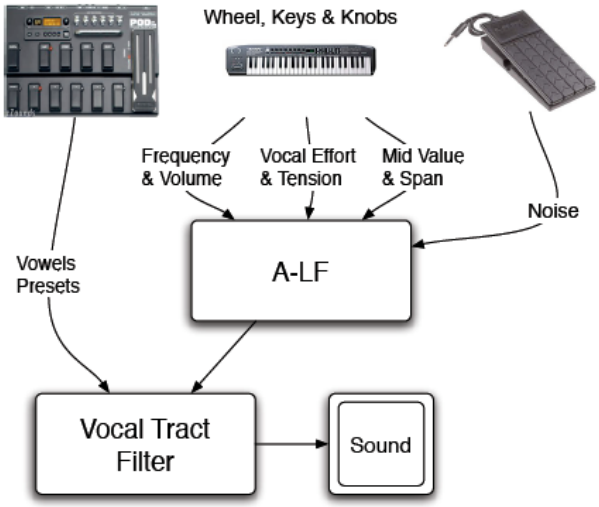

Figure 2 : orgue LF

3 Instruments vocaux

3.1 Instrument I : orgue LF

La première tentative utilise une version temps-réel du modèle de source LF, et un contrôle par des périphériques musicaux classiques, comme le clavier MIDI, la roulette MIDI, la pédale, d'où l'appellation « orgue LF » . Un tel instrument est décrit en Figure 2.

Ce type d'instrument fonctionne très mal pour la synthèse vocale [10], car les possibilités de contrôle ne sont pas du tout comparables au contrôle vocal : par exemple un contrôleur discret comme le clavier ne peut pas servir pour créer des mélodies vocales réalistes.

3.2 Instrument II : theremin vocal

Cet instrument permet le contrôle du modèle source/filtre par un gant numérique et un clavier. Le gant capte la contraction des doigts, une métaphore qui fonctionne bien pour la dimension de tension glottique, et par la distance à une borne dans l'espace pour la mélodie et l'intensité [10]. La synthèse est basée sur un modèle de source décrit dans le domaine spectral (CALM [6]). L'instrument se rapproche dans son principe d'un theremin (un des premiers instruments électroniques) augmenté de la possibilité de changer le timbre de la source et celui des voyelles. Tout comme le theremin, cet instrument n'est pas facile à contrôler mélodiquement. Il permet de jouer des râles, cris, plaintes, mais arriver à chanter juste est un véritable défi.

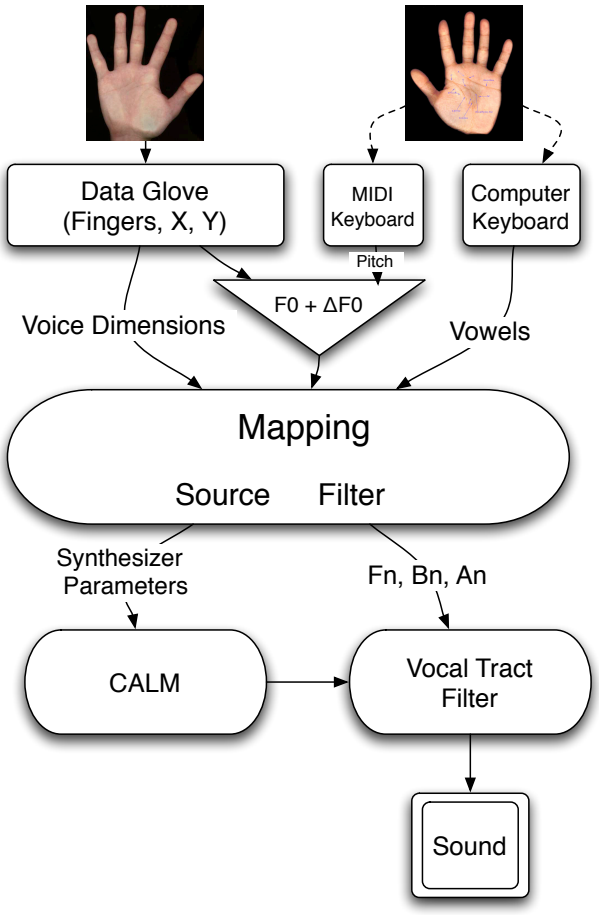

Figure 3 : Theremin vocal .

3.3 Instrument III : Calliphonie/synthèse

Pour pallier le problème de précision mélodique de l'instrument précédent, une tablette graphique est utilisée pour le contrôle mélodique, avec le même type de synthétiseur. Ainsi les capacités motrices développées lors de l'apprentissage de l'écriture, depuis l'enfance, sont mises à profit pour le contrôle intonatif.

Ce type d'instrument est d'emblée un succès, car le contrôle mélodique est à la fois intuitif et précis : on peut chanter juste avec peu d'entraînement [9][11].

Cet instrument peut être augmenté d'un joystick pour le contrôle des voyelles et des paramètres de qualité vocale. C'est alors un instrument très expressif, qui peut être employé pour imiter les vocalisations expressives, par un paradigme d'analyse par la synthèse.

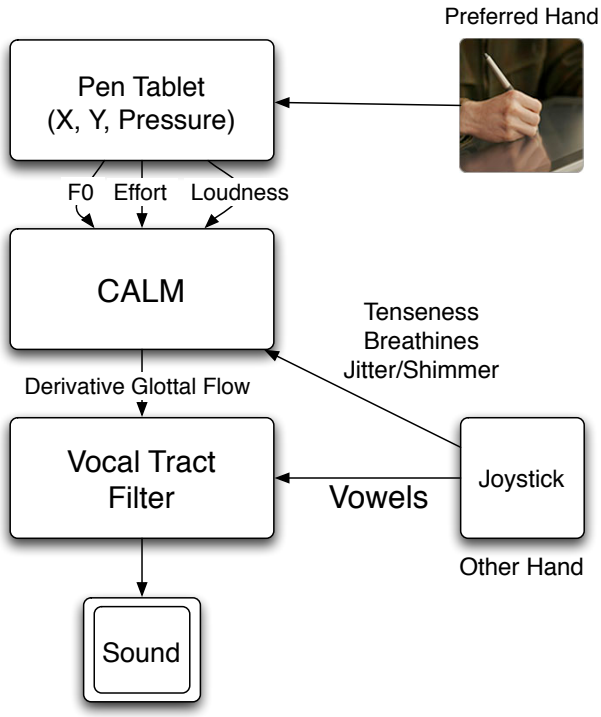

Figure 4 : Calliphonie/synthèseTheremin vocal .

3.3 Instrument IV: Calliphonie/chironomie

Cet instrument est utilisé pour le contrôle mélodique et rythmique de séquences préenregistrées, en utilisant une version temps réel de l'algorithme TD-PSOLA. Le terme « chironomie » est repris de la terminologie musicale, où il signifie le contrôle de la musique par les mouvements de la main (par exemple dans le chœur grégorien).

La tablette graphique permet de dessiner l'intonation d'une phrase, suivant les tracés de la main. Ce système permet donc de travailler sur l'analogie entre mouvements intonatifs et mouvements manuels.

Des études sur la stylisation de l'intonation par ce système indiquent que les mouvements manuels permettent de synthétiser une prosodie indiscernable de la prosodie vocale naturelle [12][13].

3.4 Instrument V: Meta-CALM

Le méta instrument est un dispositif de contrôle musical pourvu de 54 capteurs, potentiomètres linéaires et rotatifs et capteurs de force (FSR). Il permet de capturer les mouvements de pression des doigts, les rotations des poignets et des avants bras. La synthèse vocale a été portée sur le méta instrument [14], suivant le schéma de la figure 6.

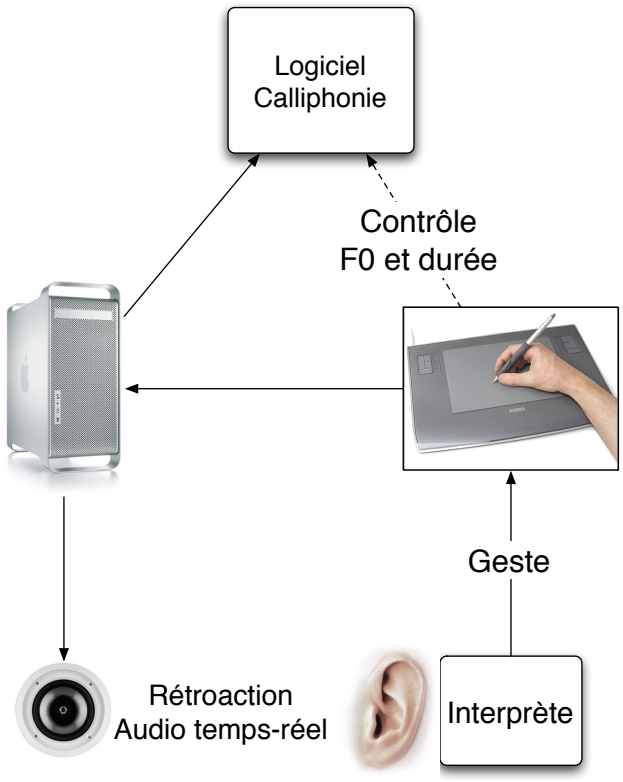

Figure 5 : Calliphonie/chironomie.

Ici encore, le contrôle mélodique précis est difficile. On se retrouve un peu dans la situation du gant numérique, mais avec des mouvements différents, plus amples, pour les capteurs de bras et de poignets, et des possibilité de déclenchement d'évènements, par les capteurs digitaux. Ainsi on peut utiliser le méta instrument pour jouer des séquences préenregistrées. Le contrôle lent de paramètres comme les formants ou la qualité vocale est également suffisamment précis.

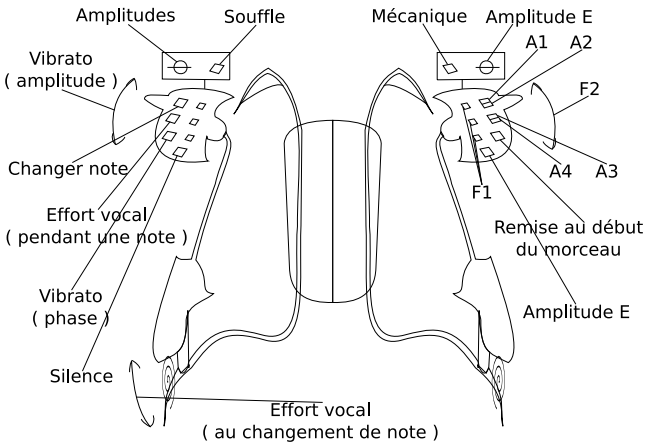

Figure 6 : Meta-calm : le méta instrument comme interface de pilotage du CALM.

3.5 Instrument VI : phonétogramme haptique

Pour augmenter le confort de jeu des instruments précédents, un bras haptique (avec un retour d'effort) est utilisé dans une interface pour la voix chantée. Il s'agit de rendre sensible les limites physiologiques et les domaines de variation de la voix grâce au retour haptique [14].

Pour la synthèse de source, cela se traduit par la sensation des régimes vibratoires de la glotte (mécanismes, passages), et par la reproduction des limites des domaines de variation. Les domaines de variation de la mélodie et de l'intensité vocale, connus sous le nom de phonétogramme (ou Voice Range Profile en anglais) sont rendus par l'intermédiaire de guides virtuels, afin d'intégrer les différentes contraintes physiologiques.

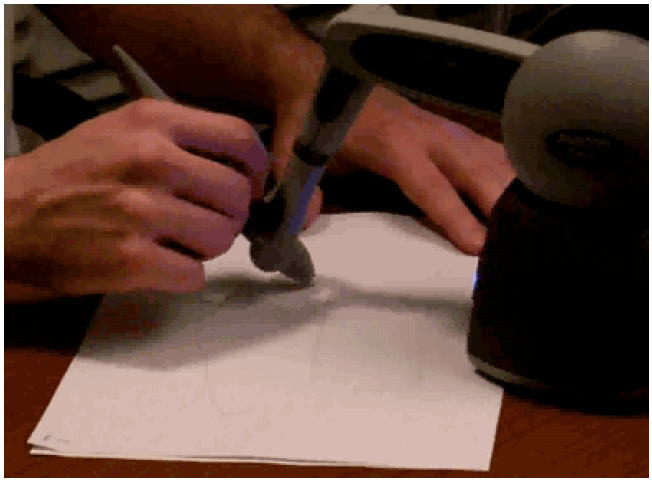

Figure 7 : phonétogramme haptique.

4 Discussion et perspectives

Cette communication présente une piste de recherche, la synthèse vocale temps réel contrôlée par le geste, que nous pensons féconde, et qui s'appuie fondamentalement sur les travaux de Gunnar Fant. Une série d'instruments a été développée, dont le plus récent comprend une dimension haptique. Ces instruments permettent d'aborder à nouveaux des questions anciennes, comme la modélisation intonative, et d'aborder des questions neuves comme le rôle de la qualité vocale dans l'expression, linguistique ou non linguistique en parole, ou dans le chant. Par l'utilisation d'instruments externes contrôlés par le geste, la chironomie permet de mettre en lumière l'affinité profonde entre mouvements ou gestes corporels et contenu expressif. Cette mise en mouvement de l'énonciation, cette mise en évidence d'un substrat gestuel signifiant, bien que non linguistique, s'inscrit dans les recherches sur ce que Fónagy appelait la « vive voix » et Certeau la « glossolalie ».

Parmi les perspectives certainement fructueuses de ce type de recherches, il faut mentionner le passage à la multimodalité (agent conversationnel expressif combinant graphique, haptique, et voix) et au contrôle collaboratif et réparti de la synthèse (choeur de voix virtuelles, instrument vocal réparti).

Références

[1] G. Fant, Half a century in phonetics and speech research, Fonetik 2000, Swedish phonetics meeting in Skövde, May 24-26, 2000

[2] G. Fant: Acoustic theory of speech production. Mouton, The Hague, 1960.

[3] G. Fant, J. Liljencrants, Q. Lin: A four-parameter model of glottal flow. STL-QPSR 4 (1985) 1–13.

[4] G. Fant: Glottal source and excitation analysis. STL-QPSR 1 (1979) 85–107.

[5] G. Fant: The LF-model revisited. transformations and frequency domain analysis. STL-QPSR 2-3 (1995) 119–156.

[6] B. Doval, C. d'Alessandro, N. Henrich: The voice source as a causal / anticausal linear filter. proc. Voqual'03, Voice Quality: Functions, analysis and synthesis, ISCA workshop, Geneva, Switzerland, Aug. 2003, 15–20.

[7] G. Fant: Preliminaries to analysis of the human voicesource. STL-QPSR 4 (1982) 1–27.

[8] C. d'Alessandro, N. D'Alessandro, S. Le Beux, J. Simko, F. Cetin, H. Pirker The Speech Conductor : Gestural Control of Speech Synthesis, Proc. of eNTERFACE 2005 Workshop, Mons, Belgium

[9] N. D'Alessandro, C. d'Alessandro, S. Le Beux, B. Doval Real-time CALM Synthesizer : New Approaches in Hands-Controlled Voice Synthesis, Proc. of New Interfaces for Musical Instruments Intl. Conference 2006, IRCAM, Paris

[10] C. d'Alessandro, N. D'Alessandro, S. Le Beux, B. Doval Comparing time domain and spectral domain voice source models for gesture controlled vocal instruments, Proc. 5th International Conference on Voice Phhysiology and Biomechanics (ICVPB'06), pp. 49-52, Tokyo, Japon.

[11] N. D'Alessandro, P.Woodruff, Y. Fabre, T. Dutoit, S. Le Beux, B. Doval, C. d'Alessandro Realtime and accurate musical control of expression in singing synthesis, in Journal on Multimodal User Interfaces, Vol. 1-1, mars 2007, pp. 31-39, Springer Berlin/Heidelberg

[12] C. d'Alessandro, A. Rilliard, S. Le Beux Computerized chironomy : evaluation of hand-controlled intonation reiteration, Proc. INTERSPEECH 2007, pp. 1270-1273, Anvers, Belgique.

[13] S. Le Beux, A. Rilliard, C. d'Alessandro Calliphony : a real-time intonation controller for expressive speech synthesis, Proc. 6th ISCA Workshop on Speech Synthesis, Bonn, Allemagne. (2007).

[14] S. Le Beux, Contrôle gestuel de la prosodie et de la qualité vocale, thèse de doctorat, Université Paris Sud- XI, décembre 2009.