Contrôle gestuel de la prosodie et de la qualité vocale

Sylvain Le Beux

N° D'ORDRE

THÈSE DE DOCTORAT

SPECIALITE : PHYSIQUE

Ecole Doctorale « Sciences et Technologies de l'Information des Télécommunications et des Systèmes »

Présentée par : Sylvain LE BEUX

Sujet :

CONTROLE GESTUEL DE LA PROSODIE ET DE LA QUALITE VOCALE

Soutenue le ……11 Décembre 2009……..devant les membres du jury :

M Christophe d'ALESSANDRO (DR au LIMSI-CNRS, Directeur de Thèse)

M Philippe DEPALLE (Pr. à l'Université McGill, Montréal, Rapporteur)

M Gaël RICHARD (Pr. à TELECOM Paris-Tech, Rapporteur)

M Gérard CHARBONNEAU (Pr. Université Paris-Sud XI, Président du Jury)

M Roberto BRESIN (Pr. à KTH, Stockholm, Examinateur)

Table des matières

1 Introduction **10**

1.1 Préambule . . . 11
1.2 Publications et Communications . . . 14
1.3 Quelques exemples de synthétiseurs "gestuels" . . . 16
1.3.1 Machines de von Kempelen et Faber . . . 16
1.3.2 Le Voder . . . 20
1.3.3 Glove Talk . . . 24
1.3.4 SPASM . . . 27
1.3.5 Le voicer . . . 30

2 Modification prosodique de la parole par contrôle gestuel **33**

2.1 Introduction . . . 35
2.2 L'algorithme PSOLA . . . 37
2.2.1 Les signaux d'analyse à court terme . . . 37
2.2.2 Les signaux de synthèse à court terme . . . 38
2.2.3 Le signal de synthèse final . . . 39
2.2.4 Le calcul des marqueurs de synthèse . . . 39
2.2.5 Le choix de la fenêtre d'analyse . . . 41
2.3 L'algorithme PSOLA en temps réel . . . 43
2.3.1 Les étapes . . . 43
2.3.2 Les contraintes temps-réel . . . 48
2.3.3 Le calcul des instants de synthèse . . . 52
2.4 Première expérience d'imitation mélodique . . . 54
2.4.1 Evaluation d'un système de réitération intonative contrôlé par la main . . . 54
2.4.2 Préambule . . . 58
2.4.3 Calliphonie : les premiers pas . . . 59
2.4.4 Résultats de l'expérience d'imitation . . . 67
2.4.5 Discussion et Conclusions Partielles . . . 68
2.5 Deuxième expérience d'imitation mélodique . . . 70
2.5.1 Le corpus . . . 70
2.5.2 Les sujets . . . 70

2.5.3 L'interface . . . 71
2.5.4 Le protocole . . . 71
2.5.5 Les résultats . . . 72
2.5.6 Analyse gestuelle . . . 79
2.6 Applications . . . 85
2.6.1 Enrichissement de base de données . . . 85
2.6.2 Voix chantée . . . 88
2.7 Conclusions du chapitre . . . 90

3 Synthèse de source glottique 92
3.1 Synthèse de Source Glottique et Qualité Vocale . . . 93
3.1.1 Modèle Linéaire Source/Filtre . . . 94
3.1.2 Les Principaux Modèles Signal de Source Glottique . . . 96
3.1.3 Le Modèle Linéaire Causal/Anticausal (CALM) . . . 106
3.2 Phonétique de la qualité vocale . . . 111
3.2.1 La notion de registre vocal . . . 112
3.2.2 La dimension de bruit . . . 120
3.2.3 L'effort vocal . . . 121
3.2.4 La dimension tendue/relâchée . . . 121
3.3 Le modèle CALM en temps réel ou RTCALM . . . 123
3.3.1 Les contraintes . . . 123
3.3.2 Les solutions . . . 123
3.3.3 Composantes non périodiques . . . 127
3.3.4 Description des fonctions de mapping . . . 131
3.4 Les différents instruments basés sur RTCALM . . . 140
3.4.1 Premier instrument . . . 140
3.4.2 Deuxième instrument . . . 142
3.4.3 Méta-instrument . . . 145
3.4.4 Exploration haptique du phonétogramme . . . 150
3.4.5 Réflexions sur l'adéquation interface/synthétiseur . . . 155
3.5 Applications . . . 156
3.5.1 Voix chantée . . . 156
3.5.2 Synthèse de qualité vocale et génération de stimuli . . . 157
3.5.3 Apprentissage phonétique . . . 157

4 Discussion et Perspectives 159
4.1 Discussion . . . 160
4.1.1 Modification prosodique . . . 160
4.1.2 Synthèse de source glottique . . . 162
4.2 Perspectives . . . 164

4.2.1 Evaluation objective de la modification de durée . . . 164
4.2.2 Evaluation des attitudes japonaises . . . 165
4.2.3 Intégration facilitée des interfaces . . . 165
4.2.4 Applications possibles . . . 166
4.2.5 Objectifs à plus long terme . . . 167

5 Références bibliographiques 169

6 Articles 179
Modification prosodique . . . 179
Interspeech 2007 . . . 179
Speech Synthesis Workshop 2007 . . . 184
Synthèse de source glottique . . . 191
eNTERFACE 2005 . . . 191
NIME 2006 . . . 202
ICVPB 2006 . . . 209
eNTERFACE 2006 . . . 214
JMUI 2008 . . . 225
Autres travaux : Le projet ORA . . . 235
ICMC 2009 . . . 235
Smart Graphics 2009 . . . 240

Table des figures

1.1 *Représentation schématique des différents tubes de Kratzenstein (d'après Dudley (Dudley and Tarnoczy, 1950))* . . . 17
1.2 *Première version des voyelles* b *et* d *par von Kempelen (d'après Dudley (Dudley and Tarnoczy, 1950))* . . . 17
1.3 *Premier synthétiseur de voyelles par von Kempelen (d'après Dudley (Dudley and Tarnoczy, 1950))* . . . 18
1.4 *La seconde machine parlante de von Kempelen (d'après Dudley (Dudley and Tarnoczy, 1950))* . . . 18
1.5 *Version finale de la machine parlante de von Kempelen (d'après Dudley (Dudley and Tarnoczy, 1950))* . . . 19
1.6 *Analogie entre l'appareil vocal humain et le fonctionnement du Voder (d'après (Dudley et al., 1939))* . . . 20
1.7 *La console du Voder en fonctionnement par une opératrice. (d'après (Dudley et al., 1939))* . . . 21
1.8 *Vue schématique des contrôleurs du Voder (d'après (Dudley et al., 1939))* . . . 22
1.9 *Vue schématique des contrôleurs du Voder (d'après (Dudley et al., 1939))* . . . 23
1.10 *Première version du Glove-Talk de Sidney Fels (Fels, 1991)* . . . 25
1.11 *Description schématique du Glove-Talk* II *de S. Fels & G. Hinton (Fels and Hinton, 1998)* . . . 26

1.12 *Description schématique du système de synthèse SPASM (Cook, 1992)* . . . 28
1.13 *L'instrument augmenté "SqueezeVox" (d'après (Cook and Leider, 2000))* . . . 29
1.14 *Le VOMID en action (d'après (Cook, 2005))* . . . 30
1.15 *Le Voicer en action (d'après (Kessous, 2004))* . . . 31

2.1 *Fichier audio original (à gauche) et fichier de GCIs correspondant (à droite).* . . . 45
2.2 *Fenêtrage et constitution du tampon audio* . . . 47
2.3 *Addition de deux signaux à court terme, lors d'une compression temporelle.* . . . 49
2.4 *Fonctionnement du tampon circulaire pour la synthèse.* . . . 51
2.5 *Diagramme générique du système de modification prosodique (Le Beux et al., 2007)* . . . 54
2.6 *Le système Calliphonie (Le Beux et al., 2007)* . . . 57
2.7 *Description technique du système* . . . 60
2.8 *Interface utilisée pour l'expérience. Les boutons permettent d'écouter la phrase originale, d'enregistrer sa voix ou la tablette graphique, d'écouter une performance et de l'enregistrer lorsqu'elle est satisfaisante. L'image représente la prosodie de la phrase courante. (Le Beux et al., 2007)* . . . 61
2.9 *Paramètres prosodiques d'une phrase à 7 syllabes ("Nous voulons manger le soir", focalisée) issue de notre corpus. (d'Alessandro et al., 2007)* . . . 64
2.10 *Valeurs de* F_0 *brutes (en demi-tons) issue d'une phrase originale (en gris) et deux imitations vocales d'un sujet (Les stimuli ne sont pas alignés temporellement, abscisses en secondes). (d'Alessandro et al., 2007)* . . . 66
2.11 F_0 *stylisé d'un phrase originale (la même que sur la Figure 2.10 – courbe verte, avec des valeurs lissées pour le segment vocalique exprimé en demi-tons, abscisses en secondes) et la valeur du paramètre de hauteur contrôlée par la tablette graphique pour toutes les imitations effectuées par un sujet. Les stimuli sont alignées temporellement. (d'Alessandro et al., 2007)* . . . 66
2.12 *Evolution des deux mesures de distance selon la longueur de la phrase. En abscisses : la longueur du stimulus en nombre de syllabes. En ordonnées, à droite : corrélation (ligne rouge), à gauche : différence* RMS *(ligne bleue). (d'Alessandro et al., 2007)* . . . 68
2.13 *Résultats de corrélation, pour chaque sujet (en abscisses), pour les imitations vocales (en vert) et gestuelles (en bleu)* . . . 73
2.14 *Résultats de corrélation suivant le genre du locuteur (Féminin, Masculin) pour les imitations gestuelles (en bleu) et vocales (en vert).* . . . 74
2.15 *Exemple d'imitation gestuelle (rouge) et vocale (vert), pour une voix cible masculine (gris), par un sujet féminin. Fréquence en demi-tons, temps en secondes.* . . . 75
2.16 *Effet de l'apprentissage musical sur les distances* RMS*. Imitations orales en rouge et avec la tablette en bleu* 75
2.17 *Effet de l'apprentissage musical sur les corrélations. Imitations orales en rouge et avec la tablette en bleu* . 76
2.18 *Résultats de corrélation suivant le type de phrase utilisée pour les imitations gestuelles (en bleu) et vocales (en vert).* . . . 77
2.19 *Résultats de moindres carrés suivant le type de phrase utilisée pour les imitations gestuelles (en bleu) et vocales (en vert).* . . . 78
2.20 *Visualisation, pour chaque sujet, des résultats de corrélation en fonction de la valeur de* β *obtenue par régression linéaire.* . . . 82

2.21 *Visualisation, pour chaque sujet, des résultats de corrélation en fonction de la valeur de la secousse.* . . . 83
2.22 *Distances* RMS *(losanges) et corrélations (carrés) suivant la longueur de la phrase. (Le Beux et al., 2007)* . 86
2.23 *Description du processus d'enrichissement d'une base de données. (Le Beux et al., 2007)* 88

3.1 *Représentation schématique du fonctionnement du modèle source-filtre (D'Alessandro et al., 2006)* 95
3.2 *Formes typiques de l'onde de débit glottique (en haut) et de sa dérivée (en bas), et leurs paramètres temporels respectifs (d'après (Doval et al., 2003))* 98
3.3 *Formes et expressions des différents modèles de Rosenberg (d'après Rosenberg (Rosenberg, 1971)).* . . 99
3.4 *Forme et expression de l'onde de débit glottique dérivée du modèle LF (d'après Fant (Fant et al., 1985))* . . 103
3.5 *Comparaison des quatre modèles de source glottique, pour un même jeu de paramètres, lors d'une fermeture abrupte et pour* E *constant (d'après (Doval et al., 2003))* 106
3.6 *Spectre en amplitude de l'onde de débit glottique dérivée : illustration du formant glottique (*$\mathrm{F_g}$, $\mathrm{A_g}$*) et du tilt spectral (*$\mathrm{F_a}$, $\mathrm{A_a}$*) (d'après (Doval et al., 2003))* 107
3.7 *Représentation dans le domaine temporel de l'onde de débit glottique dérivée : partie anticausale et partie causale. (d'après (D'Alessandro et al., 2006))* 108
3.8 *Configuration typique pour la voix modale, de la glotte (en haut) et, de l'ODG (en bas), d'après (Klatt and Klatt, 1990)* 115
3.9 *Les deux modes propres possibles du modèle à deux masses, avec en haut un mode propre où les deux masses vibrent en phase et, en bas, un mode où les deux masses sont en quadrature de phase (d'après (Titze and Strong, 1975)).* 116
3.10 *Spectres en amplitude typiques des mécanismes* M1, M2 *et de la voix soufflée, de gauche à droite (d'après (Klatt and Klatt, 1990))* 119
3.11 *Description schématique du fonctionnement de l'algorithme RT-CALM (d'après (D'Alessandro et al., 2006))* . 124
3.12 *Discontinuité de l'ODGD (à droite) due à la troncature de l'*ODG *au premier passage à* 0 *de la pulsation* CALM *(à gauche). (D'Alessandro et al., 2007)* 125
3.13 *Capture d'écran du patch Pure Data, pour la génération de l'*ODGD *et de son filtrage.* 127
3.14 *Comparaison du spectre idéal et mesuré de la pression glottique de la source(d'après (Hillman et al., 1983))* 130
3.15 *Phonétogramme moyen (*M1, M2*) pour une voix masculine (en haut) et féminine (en bas) (d'après (Roubeau et al., 2009))* 135
3.16 *Sauts en fréquence exprimés en demi-tons depuis le registre de poitrine (modal) vers le falsetto. En bleu, pour un saut à* 200 Hz*, en rouge pour un saut à* 300 Hz*, et à* 400 Hz *en jaune. (d'après (Bloothooft et al., 2001))* 136
3.17 *Sauts en fréquence en demi-tons depuis le falsetto vers le registre de poitrine (modal). (d'après (Bloothooft et al., 2001))* 137
3.18 *Enveloppes spectrales à long terme pour différents types de chanteurs (d'après (Sundberg, 2001))* 138
3.19 *Mapping des dimensions vocales de contrôle aux paramètres du modèle de source glottique. (d'Alessandro et al., 2006)* 142
3.20 *Les trois degrés de liberté de la tablette graphique (D'Alessandro et al., 2006)* 143
3.21 *Illustration du passage d'une voix relâchée à tendue (D'Alessandro et al., 2006)* 143
3.22 *Vue schématique de l'instrument CALM avec tablette et joystick. (d'Alessandro et al., 2006)* 144

3.23 *Le Méta-instrument* 145
3.24 *Détails des capteurs de la main droite* 145
3.25 *Les correspondances des capteurs du Méta-instrument avec les paramètres du synthétiseur CALM.* 150
3.26 *Le bras haptique PHANTOM Omni.* 152
3.27 *Visualisation dans Max/MSP des mécanismes* $\mathrm{M1}$ *(en bleu) et* $\mathrm{M2}$ *(en rouge), avec la position du stylet du bras haptique dans l'espace.* 154

Remerciements

Je tiens en premier lieu à remercier mon directeur de thèse Christophe d'Alessandro pour avoir su tout au long de mes années de doctorat faire part de pédagogie, de professionnalisme et d'exigence pour m'obliger à me dépasser et surtout comprendre les rouages du métier de chercheur. Au-delà des échanges scientifiques que nous avons pu avoir, je tiens également à souligner sa gentillesse, sa disponibilité et sa grande culture qui rendent chacune des discussions l'occasion d'un enrichissement intellectuel.

Je remercie amicalement tous les membres de mon jury : mes rapporteurs, Philippe Depalle et Gaël Richard pour avoir accepter de faire la critique constructive de mon travail et y avoir trouvé un intérêt non dissimulé. Je remercie Gérard Charbonneau, d'avoir accepté de faire partie mon jury, et dont j'avais eu l'honneur de reprendre le flambeau pour donner des cours de traitement du signal musical et de psychoacoustique. Enfin, je remercie Roberto Bresin, dont j'apprécie particulièrement le travail, et je tiens à souligner avoir accepté de relire un manuscrit dont la langue n'était pas sa langue maternelle.

Je souhaite également adresser des remerciements appuyés à Boris Doval et Albert Rilliard, collègues du LIMSI, qui ont pu suivre mon travail au jour le jour et surtout me permettre de mener à bien les développements présentés dans ce présent manuscrit. Mes connaissances actuelles en traitement du signal de parole, de phonétique, et d'analyse prosodique ne seraient pas aussi approfondies sans les explications qu'ils ont pu m'apporter au cours de mon doctorat.

J'en profite pour remercier toutes les personnes, enseignants, doctorants et chercheurs avec qui j'ai pu collaborer que ce soit lors de mon travail de thèse, du projet ORA, de mes enseignements à l'IEF, de ma formation complémentaire à l'ENS Cachan, ou de mes modules d'ouverture scientifique. A ce titre, j'adresse un remerciement particulier à Nicolas D'Alessandro, Christian Jacquemin, Samir Bouaziz, Alain Finkiel et Gilles Léothaud.

Evidemment, je n'aurai pu apprécier mon passage au LIMSI sans les nombreuses amitiés avec les doctorants et le personnel du LIMSI, et qui m'ont permis de rendre cette expérience plus agréable. Je sais reconnaître l'importance de l'atmosphère de travail permettant de se

sentir à l'aise, et l'ambience présente au LIMSI est à ce titre très agréable. Au risque d'oublier quelqu'un, je n'essaiera pas de dresser de liste exhaustive, mais je sais que chacun saura se reconnaître.

Enfin, je remercie tous mes proches, amis et famille pour m'avoir toujours apporté leur soutien et dont le recul suffisant leur permet de poser des questions telles que : "Alors, c'est pour quand ?", "Mais, à quoi ça sert exactement ?", "C'est bien beau d'étudier, mais quand est-ce que tu vas travailler ?" ou "Tu reprendras bien un peu de rôti ?". Un grand merci à eux, sincèrement.

Chapitre 1

Introduction

Sommaire

1.1 Préambule **11**
1.2 Publications et Communications **14**
1.3 Quelques exemples de synthétiseurs "gestuels" **16**
1.3.1 Machines de von Kempelen et Faber 16
1.3.2 Le Voder 20
1.3.3 Glove Talk 24
1.3.4 SPASM 27
1.3.5 Le voicer 30

1.1 Préambule

Depuis aussi longtemps que les ordinateurs existent, la possibilité de synthétiser la voix suscite un intérêt. On peut ainsi remonter à 1961 et aux expériences de Max Mathews aux Bell Labs pour retrouver la trace d'une chanson synthétisée grâce à un IBM 704. Il s'agit de la désormais fameuse chanson "Daisy Bells", composée par Harry Dacre en 1892, et qui fut notamment popularisée par le film 2001 : L'odyssée de l'espace.

Après plusieurs décennies de recherche et de développement en synthèse vocale assistée par ordinateur, nous ne sommes aujourd'hui pas encore parvenu à obtenir de synthèse vocale véritablement "transparente". Par transparente, j'entends une voix de synthèse qui ne soit pas distinguable d'une voix naturelle. Et cela, même grâce à l'aide de synthétiseurs par concaténation d'unités acoustiques non uniformes. Ces derniers synthétiseurs sont en effet reconnus de façon consensuelle comme les représentants de la meilleure synthèse vocale atteignable à l'heure actuelle. Si l'on y regarde de plus près, on peut effectivement admettre qu'ils produisent la meilleure intelligibilité, mais concernant la génération d'expressions, des progrès restent encore à accomplir. L'un des défis majeurs en synthèse vocale depuis maintenant près d'une dizaine d'années a donc consisté à adopter des techniques permettant de synthétiser une voix plus naturelle, plus expressive ou avec une émotion donnée. Il est possible de regrouper selon deux tendances principales ces différentes méthodes : les unes consistent à régler savamment les coûts de concaténation afin de ne sélectionner que les unités pouvant bien se raccorder en termes d'intonation et de durée, les autres travaillent plus en amont, en cherchant à construire des bases de données contenant au préalable certaines émotions caractéristiques (joie, peur, tristesse ...).

Aussi bonnes soient telles, ces techniques n'adressent pas le problème de manière fondamentale. Les premières méthodes ne pourront jamais synthétiser de meilleurs expressions que celles déjà présentes dans la base de données. En outre, elles nécessitent afin de régler les coûts de concaténation de pouvoir établir des règles fiables de modélisation prosodique. Or, il n'existe à l'heure actuelle pas de consensus clair sur la manière de modéliser la prosodie naturelle. Le problème de la seconde approche, est que l'on ne sera pas capable de reproduire une émotion donnée si celle-ci n'a pas été enregistrée. De plus, il faut alors enregistrer autant de fois la base de données que d'émotions que l'on cherche à synthétiser. Lorsque l'on sait qu'une base de données de travail minimale nécessite déjà plusieurs heures d'enregistrement, on peut facilement imaginer la difficulté pour construire une telle base de données. Ces bases de données sont pour la plupart enregistrées par un acteur, et la validité de ce corpus particulier peut intrinsèquement être remise en cause, sachant que certains auditeurs sont capables, avec des performances loin de la simple chance de distinguer entre une émotion actée et une émotion naturelle. Et que dire des émotions mixtes, entre la peur et la tristesse, entre la joie et la surprise. Une simple concaténation alternée d'unités de "peur" et de "tristesse", ne servirait qu'à construire un "monstre". Enfin, la grande majorité des systèmes actuels s'intéresse aux émotions et non pas à l'expression. Pourtant, aussi

provocant que cela puisse paraître de prime abord, une voix émotionnelle n'est pas naturelle. Nous passons la très grande majorité de notre temps à parler avec certes une certaine expression, mais sans doute pas avec émotion. Quand bien même, s'il on peut dire que les émotions ne sont que les configurations extrêmes de certaines expressions, ces dernières ne peuvent être représentées simplement en termes de mélanges d'émotions.

Il convient également de noter que ces méthodes, dans leur ensemble, ne s'intéressent principalement qu'à la prosodie, que nous pouvons définir en première approche comme les changements d'intonation et de durée de la parole. Comme nous le verrons par la suite, un domaine important concernant la modélisation de l'expressivité vocale est celui de l'étude de la source glottique, que nous engloberons sous le terme de qualité vocale. La qualité vocale contribue en effet de manière non négligeable à la production expressive d'une occurrence de parole.

Toutes ces constatations nous ont amené à réfléchir à la manière d'aborder le problème sous un nouvel angle, plus en adéquation avec la perspective de recherche que nous souhaitions adopter. Et pour y parvenir, le croisement fructueux de deux domaines souvent tenus à l'écart nous a permis d'avancer dans notre démarche. Le premier est celui des méthodes de modification de la parole, nous permettant de nous départir des bases de données, puisqu'il ne s'agit plus ici de chercher à opérer la meilleure sélection d'unités possible, mais d'obtenir l'expression ciblée selon un processus d'analyse puis resynthèse. Ce type de processus est notamment couramment utilisé dans le domaine de la conversion vocale. Le second domaine est celui du contrôle gestuel, qui consiste à utiliser des interfaces plus ou moins usuelles, allant de la souris jusqu'au gant de données en passant par la tablette graphique, afin de contrôler les paramètres de synthèse de manière interactive. L'avantage de cette approche est justement son aspect interactif. Nous sommes ainsi capable de nous abstraire de toute modélisation prosodique *a priori*. Cela nous permet alors de pouvoir redéfinir de manière plus fine cette même modélisation. Cette approche comporte toutefois une contrainte, qui est celle de la réalisation des processus en temps-réel. Aujourd'hui, la popularité et la robustesse d'environnements de programmation temps-réel, tels que Max/MSP, Pure Data ou SuperCollider, ont pu faire leurs preuves, en offrant la possibilité de développement de systèmes de synthèse audionumériques interactifs.

Notre étude s'est donc déroulée selon deux axes : d'une part la modification prosodique de la hauteur et de la durée sur de la parole enregistrée, naturelle ou synthétique, et d'autre part, la synthèse de voyelles, à partir d'un modèle de source glottique, dans l'intention d'approfondir l'étude sur la qualité vocale.

Concernant la modification prosodique, nous avons implémenté en temps réel l'algorithme d'addition-recouvrement de fenêtres synchrones à la période (PSOLA), pour permettre la modification conjointe de la hauteur et de la durée d'une phrase de parole enregistrée. Grâce à cet outil, nous avons mené différentes expériences, visant à la fois à valider et évaluer la possibilité de modifier

indépendamment et conjointement la hauteur et la durée de la parole grâce au geste manuel. Les résultats de nos expériences nous ont montré que l'on était capable, avec une précision proche de celle de la voix, de reproduire une intonation donnée avec le geste manuel, lui-même comparable à un geste d'écriture.

Le second axe de notre recherche s'est focalisé sur la réalisation d'un synthétiseur de voyelles, dont on peut modifier la qualité vocale, dans un espace perceptif. La première étape de la réalisation de notre synthétiseur de source glottique a donc consisté à implémenter une version en temps réel de l'onde de débit glottique, et de sa dérivée, ainsi que les apériodicités de la source vocale. Nous pouvons ainsi contrôler simultanément la fréquence fondamentale, l'effort vocal, la tension, le souffle, les apériodicités structurelles (jitter et shimmer), et les différents mécanismes laryngés. De multiples interfaces différentes ont été utilisées au cours du développement de notre synthétiseur, afin de réaliser le contrôle de ces dimensions vocales. Une attention particulière a été apportée à la réalisation du phonétogramme, ainsi qu'aux correspondances perceptives entre les paramètres du modèle et les dimensions vocales.

Notre étude permet d'envisager à terme la modification conjointe, en temps réel, des composantes prosodiques et de qualité vocale.

L'organisation de ce manuscrit est la suivante : dans ce présent chapitre 1, il sera question de machines parlantes contrôlées par le geste, dans le but de présenter leurs particularités, leurs avantages et leurs inconvénients. Le chapitre 2 [1] traitera de la modification prosodique, au sens de la modification de hauteur et de durée. Nous présenterons dans un premier temps l'implémentation logicielle en temps réel de l'algorithme PSOLA, puis les expériences ayant été menées pour l'évaluation et la validation de la modification gestuelle de l'intonation. Le chapitre 3 [2] traitera quant à lui de la qualité vocale et de la synthèse de voyelles expressives. Après avoir présenté les différents modèles de source glottique usuels, puis les relations existant entre ces modèles et la phonétique de la production vocale, nous détaillerons la réalisation d'un synthétiseur de voyelles tenant compte de la qualité vocale. A la fin de ce chapitre, nous présenterons les différents systèmes de contrôle gestuels utilisés avec ce synthétiseur. Enfin, nous résumerons dans le chapitre 4 les différents accomplissements obtenus et nous étudierons les perspectives et les applications possibles, à plus ou moins long terme.

A la fin de ce manuscrit, après les références bibliographiques, sont par ailleurs regroupés par thèmes les différentes articles publiés (voir chapitre 6 et sommaire).

1. Les articles faisant référence à ce chapitre sont repérés en fin de chapitre 1 par un signe $*$
2. Les articles faisant référence à ce chapitre sont repérés en fin de chapitre 1 par un signe $\pounds$

1.2 Publications et Communications

Actes de Revue avec relecture

[1]$^{£}$ 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

Actes de conférence avec relecture

[2]$^{£}$ 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

[3]$^{*£}$ S. Le Beux *Modification en temps réel des paramètres prosodiques et de qualité vocale de la parole grâce au contrôle gestuel et les relations entre l'expressivité de la voix et l'intention gestuelle*, Proc. Journées Jeunes Chercheurs en Audition, Acoustique Musicale et Signal Audio 2006, Lyon, France

[4]$^{£}$ 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.

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

[6]* 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.

[7] C. Jacquemin, R. Ajaj, S. Le Beux, C. d Alessandro, M. Noisternig, B. F. G. Katz, B. Planes. (2009). *The Glass Organ : Musical Instrument Augmentation for Enhanced Transparency*. In Proceedings 9th International Symposium on SmartGraphics 2009. pp. 179-190. Salamanca, Spain. May 28-30 2009.

[8] C. d'Alessando, M. Noisternig, S. Le Beux, L. Picinali, B. FG Katz, C. Jacquemin, R. Ajaj, B. Planes, N. Strurmel, N. Delprat. (2009). *The ORA Project : Audio-Visual Live Electronics and the Pipe Organ*. In Proceedings International Computer Music Conference (ICMC 2009). Montreal, Canada. August 16-21, 2009.

Actes de Workshops

[9]$^{£*}$ 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

[10]$^{£}$ N. D'Alessandro, B. Doval, S. Le Beux, P. Woodruff, Y. Fabre *RAMCESS : Realtime and Accurate Musical Control of Expression in Singing Synthesis* , Proc. of eNTERFACE 2006 Workshop, Dubrovnik, Croatia.

Rapports et Mémoires

[11] S. Le Beux, *Synthèse de la parole à partir du texte*, Rapport Interne IRCAM, 2004

[12] S. Le Beux, *Contrôle Gestuel de la synthèse de parole*, Rapport de Master Recherche SETI, LIMSI-CNRS, 2005

[13] S. Le Beux *Créativité et Education*, Rapport de Mémoire de formation ACTA, ENS Cachan, Septembre 2007.

1.3 Quelques exemples de synthétiseurs "gestuels"

Nous faisions référence en préambule à la première synthèse vocale réalisée par ordinateur, mais il ne faut pas oublier que certains pionniers n'avaient pas attendus l'arrivée de l'ordinateur pour concevoir des "instruments" d'un nouveau genre permettant de synthétiser la voix. Nous focaliserons cependant ici notre inventaire sur les machines parlantes disposant d'une interface gestuelle.

1.3.1 Machines de von Kempelen et Faber

On trouve une très belle description des machines de von Kempelen et de celle de Faber, dans un article de H. Dudley (Dudley and Tarnoczy, 1950), lui-même à l'origine du Voder et du Vocoder, que nous décrirons pour sa part dans la section suivante.

Vers la fin du $XVIII^e$ siècle, Wolfgang von Kempelen construisit avec succès l'une des plus fameuses machines parlantes de l'histoire. Commencé en 1769, l'accomplissement de ce travail l'occupa pendant près de deux décennies, au bout desquelles il obtenu, avec sa troisième version, un système qui le satisfasse pleinement. Cette machine était capable de produire mécaniquement certaines voyelles et consonnes, qui combinées les unes aux autres permettaient de créer quelques mots rudimentaires.

Jusqu'alors, seules existaient des machines parlantes, ou têtes parlantes, dont l'aspect extérieur était celui d'un visage antropomorphe, et qui étaient composées de mécanismes permettant de mouvoir la bouche. Cette dernière était alors reliée à des tubes au bout desquels un opérateur parlait et donnait ainsi au spectateur l'illusion d'une tête qui parle par un procédé acousmatique.

Jacques de Vaucanson avait d'ailleurs utilisé ce même procédé pour construire son automate flûtiste, dont Bernard Le Bouyer de Fontenelle dira à son propos lors de la présentation faite par Vaucanson devant l'académie des sciences :

> *" L'Académie a été témoin ; elle a jugé que cette machine étoit extrêmement ingénieuse, que l'Auteur avoit su employer des moyens simples et nouveaux, tant pour donner aux doigts de cette Figure, les mouvemens nécessaires, que pour modifier le vent qui entre dans la Flûte en augmentant ou diminuant sa vitesse, suivant les différents tons, en variant la disposition des lèvres, et faisant mouvoir une soupape qui fait les fonctions de la langue ; enfin, en imitant par art tout ce que l'homme est obligé de faire."*

A la même époque, le développement de la phonétique et donc de la compréhension de la production vocale a sans doute contribué à l'émergence des travaux de von Kempelen. A ce titre, la réalisation de l'un de ces contemporains, Christian Gottlieb Kratzenstein, a probablement inspiré von Kempelen.

Pour les besoins d'un concours lancé par l'Académie Impériale de Saint-Pétersbourg en 1779, et visant à l'explication physiologique et la réalisation d'un appareil permettant de comprendre le mécanisme de production des voyelles, Kratzenstein a imaginé et développé cinq tubes différents, dont les longueurs et les formes étaient censées reproduire la forme du conduit vocal lors de la production des différentes voyelles $\mathrm{a}, \mathrm{e}, \mathrm{i}, \mathrm{o}, \mathrm{u}$. Au bout de ces tubes était placée une anche vibrante, les tubes servant alors de "caisse de résonance" pour la production des voyelles. Une représentation schématique de ces tubes est présentée sur la figure 1.1.

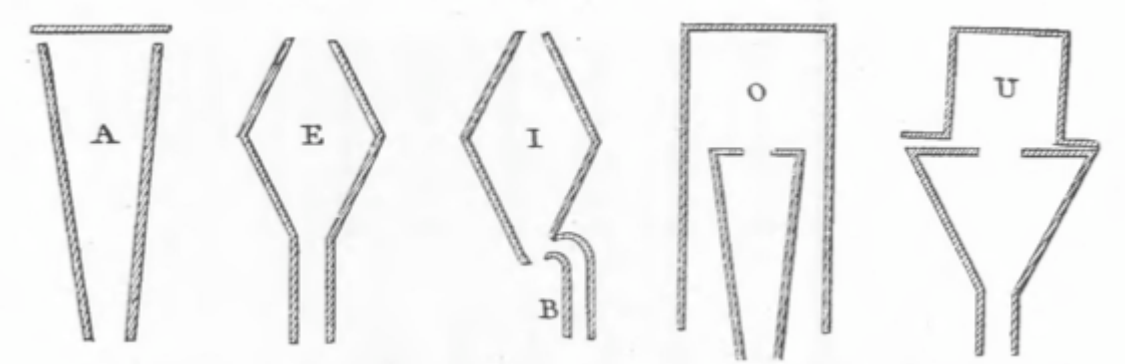


FIGURE 1.1 – *Représentation schématique des différents tubes de Kratzenstein (d'après Dudley (Dudley and Tarnoczy, 1950))*

L'idée de von Kempelen, a été, plutôt que d'avoir des formes fixes pour chaque voyelle ou consonne, de disposer d'une partie fixe, censée reproduire la bouche, et d'une partie mobile, symbolisant la langue et les lèvres. Sur la figure 1.2 est illustrée l'une des premières implémentations de von Kempelen pour la production des consonnes b et d. Ainsi, grâce au déplacement de la pièce mobile interne et à l'ouverture ou non de la cavité, les consonnes b et d pouvaient être produites.

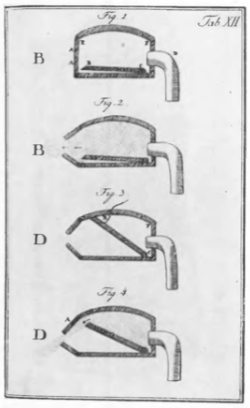


FIGURE 1.2 – *Première version des voyelles* b *et* d *par von Kempelen (d'après Dudley (Dudley and Tarnoczy, 1950))*

Pour la production des voyelles, il imagina, comme montré sur la figure 1.3, une pièce en forme de cloche, à laquelle était attachée une anche. La réalisation des voyelles pouvait alors se faire soit en déplaçant une pièce permettant d'ouvrir plus ou moins la partie évasée ou, plus simplement et efficacement en plaçant sa main sur cette ouverture selon différentes configurations pour chacune des voyelles.

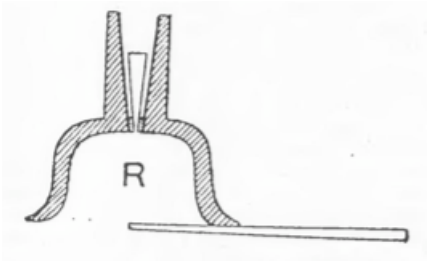


FIGURE 1.3 – *Premier synthétiseur de voyelles par von Kempelen (d'après Dudley (Dudley and Tarnoczy, 1950))*

Cette première méthode ne convenait cependant pas, selon von Kempelen pour pouvoir faire la distinction entre les différentes voyelles. Aussi, von Kempelen décida de construire une seconde machine, comme illustré sur la figure 1.4, pour laquelle, à l'instar de Kratzenstein, il utilisa un résonateur dédiée pour chacune des voyelles. On peut voir sur l'illustration, les différentes formes de résonateurs utilisés, ainsi que l'apparition d'un clavier pour pouvoir sélectionner le son à produire, et également la présence d'un soufflet à l'arrière jouant le rôle des poumons. Cette seconde machine lui permit de produire de manière assez convaincante les voyelles a, u et o, ainsi que les consonnes p, m et l.

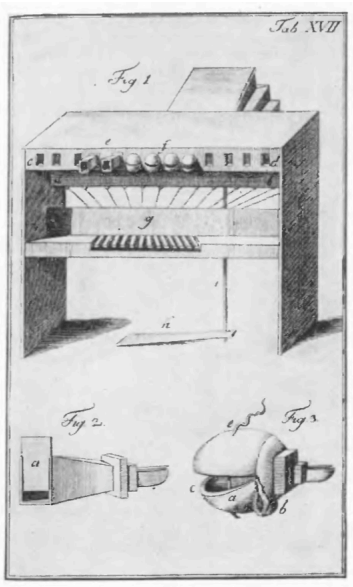


FIGURE 1.4 – *La seconde machine parlante de von Kempelen (d'après Dudley (Dudley and Tarnoczy, 1950))*

Toutefois, von Kempelen n'était pas suffisamment satisfait de la co-articulation entre les voyelles et les consonnes, et particulièrement par le caractère plosif des voyelles. Il entreprit alors de construire une troisième et dernière machine en cherchant à combler les défauts de ces deux premières réalisations. On peut voir sur la figure 1.5, une reproduction de cette machine.

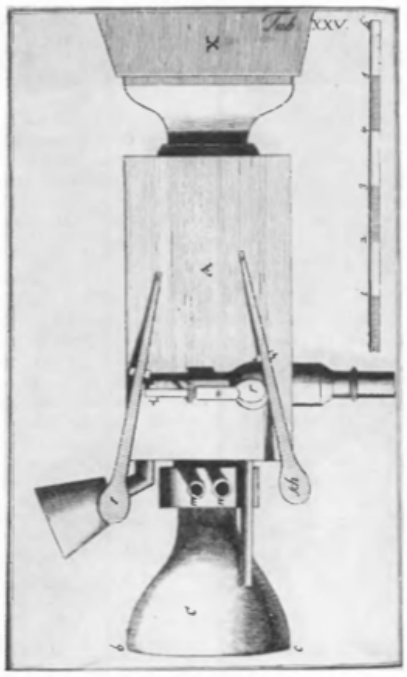


FIGURE 1.5 – *Version finale de la machine parlante de von Kempelen (d'après Dudley (Dudley and Tarnoczy, 1950))*

Au-dessus de la partie indiquée par un X, se trouvait le soufflet utilisé pour insuffler de l'air dans les parties basses A et C. Cette machine était alors utilisée à la manière d'une cornemuse : le soufflet était contrôlé par le mouvement d'un bras, l'une des mains déplaçait les différents actionneurs permettant la production des consonnes et l'autre main était placé sur l'extrémité C pour les voyelles, comme expliqué ci-dessus. De façon surprenante mais également convaincante, cette machine permettait de produire pratiquement toutes les consonnes de l'allemand (mis à part le son ŋg), comme indiqué dans la tableau 1.1 :

Classe	Son
Semi-voyelles	l,m,n,r
Plosives	p,b,d,t,k,g
Fricatives	f,v,s,z,sh
Transitions	h,w,y

TABLE 1.1 – Consonnes produites par la machine de von Kempelen

1.3.2 Le Voder

Pour les besoins, notamment, de l'exposition universelle de New-York en 1939, H. Dudley et ses collaborateurs des Laboratoires Bell ont permis la réalisation d'un synthétiseur vocal original, à savoir le Voder[3]. Dans un article , datant également de 1939, H. Dudley (Dudley et al., 1939) décrit le fonctionnement de cette machine parlante d'un nouveau genre.

Le principe de base, presque tautologique, sur lequel H. Dudley fait reposer le fonctionnement du Voder est celui de la comparaison avec le fonctionnement de la voix chez l'humain. Cette analogie est décrite sur la figure 1.6.

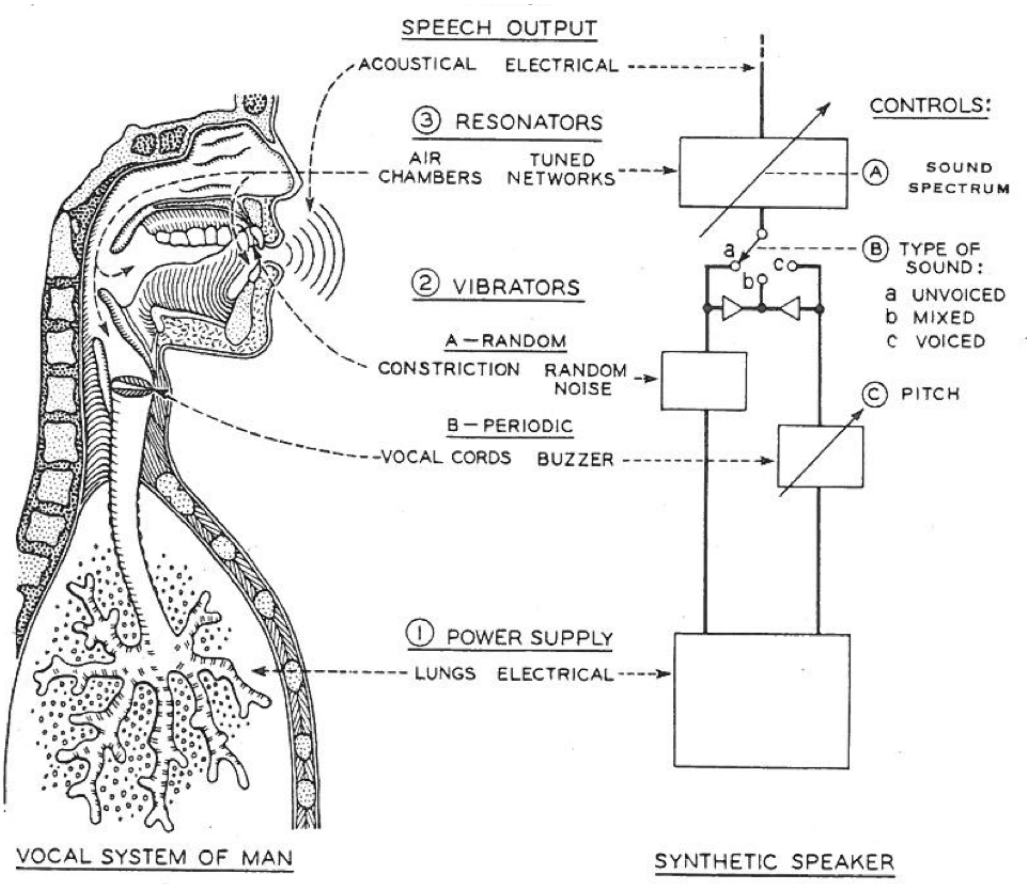


FIGURE 1.6 – *Analogie entre l'appareil vocal humain et le fonctionnement du Voder (d'après (Dudley et al., 1939))*

On retrouve alors sous forme électrique les trois fonctions principales nécessaires à la production vocale :

3. acronyme pour Voice Operation DEmonstratoR

1. Le rôle des poumons : l'énergie électrique fournie au système représente la force de l'air issu des poumons. Ainsi, l'énergie du courant électrique appliqué en entrée du système sert au contrôle de l'amplitude du signal transmis jusqu'aux haut-parleurs en sortie.
2. Le rôle du larynx : soit le signal est voisé et les cordes vocales sont en vibration, et alors le signal est généré grâce à un oscillateur électrique pour produire un son qualifié de "bourdonnement" (buzz), soit le son n'est pas voisé et il existe une constriction dans le conduit vocal, et le son est généré par un bruit aléatoire qualifié de "chuintement" (hiss). La fréquence fondamentale de l'oscillateur peut être modifiée pour changer la hauteur du son produit.
3. Le rôle du conduit vocal : afin de modéliser les résonances du conduit vocal, un réseau d'une dizaine de filtres résonants de différentes fréquences est utilisé.

Ce qui a constitué l'une des grandes forces et une particularité intéressante du Voder est le fait que ce synthétiseur était contrôlé par un opérateur, et donc par un moyen gestuel. On trouve sur les figures 1.7 et 1.8 respectivement une photographie du Voder en fonctionnement, et une description schématique des différents contrôles utilisés.

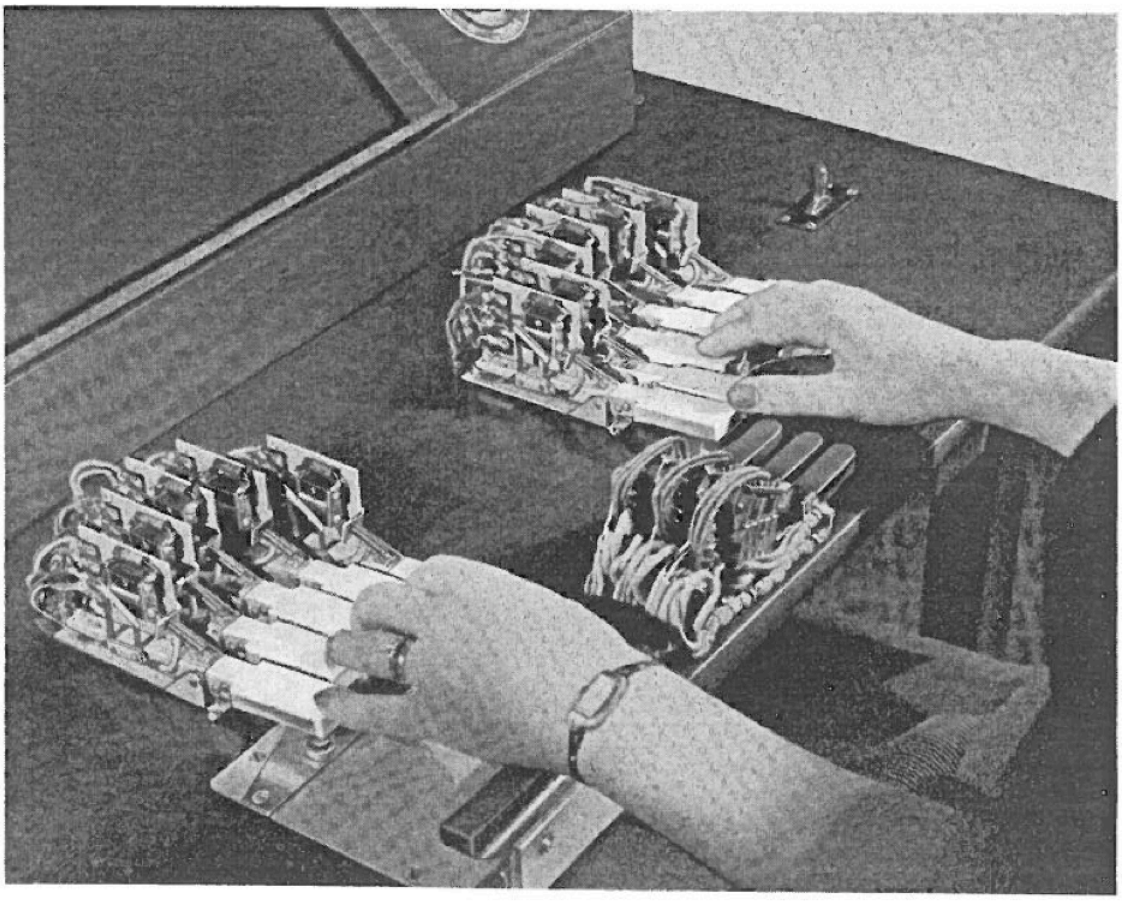


FIGURE 1.7 – *La console du Voder en fonctionnement par une opératrice. (d'après (Dudley et al., 1939))*

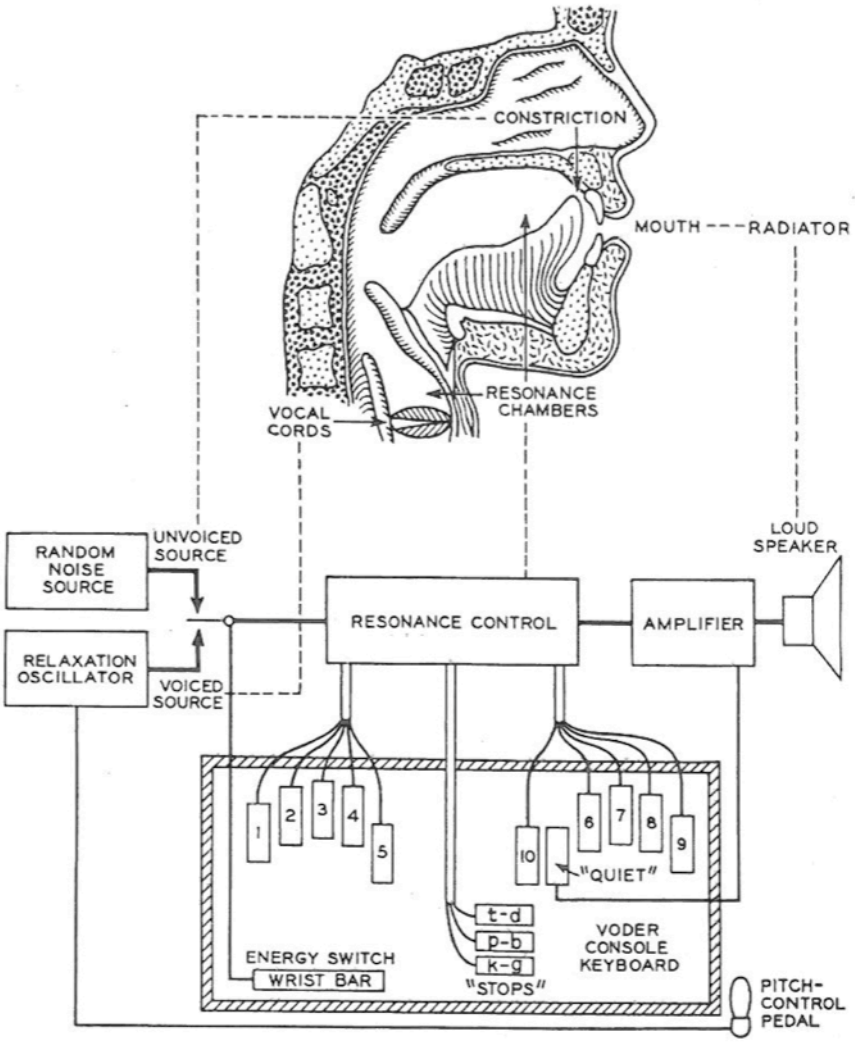


FIGURE 1.8 – *Vue schématique des contrôleurs du Voder (d'après (Dudley et al., 1939))*

Revenons un instant sur cette dernière figure 1.8 concernant le contrôle de la partie vibratoire du Voder. Comme on peut le voir, le basculement entre l'oscillation périodique et la source de bruit est réalisé grâce à une touche placée au niveau du poignet gauche de l'opératrice, ce que l'on retrouve illustré sur la figure 1.7. Le Voder dispose en outre d'un contrôle de la fréquence fondamentale à l'aide d'une pédale reliée à l'oscillateur.

La majeure partie du contrôle du Voder est dédiée à l'implémentation du conduit vocal. Sur la figure 1.9, on observe que les dix touches du clavier du Voder sont chacune liée à un filtre passe-bande dont la bande passante correspond à l'une des valeurs indiquées sur cette figure (numérotés de 1 à 10). L'enfoncement de la touche correspondante sur le clavier du Voder détermine l'amplitude du filtre, de manière logarithmique.

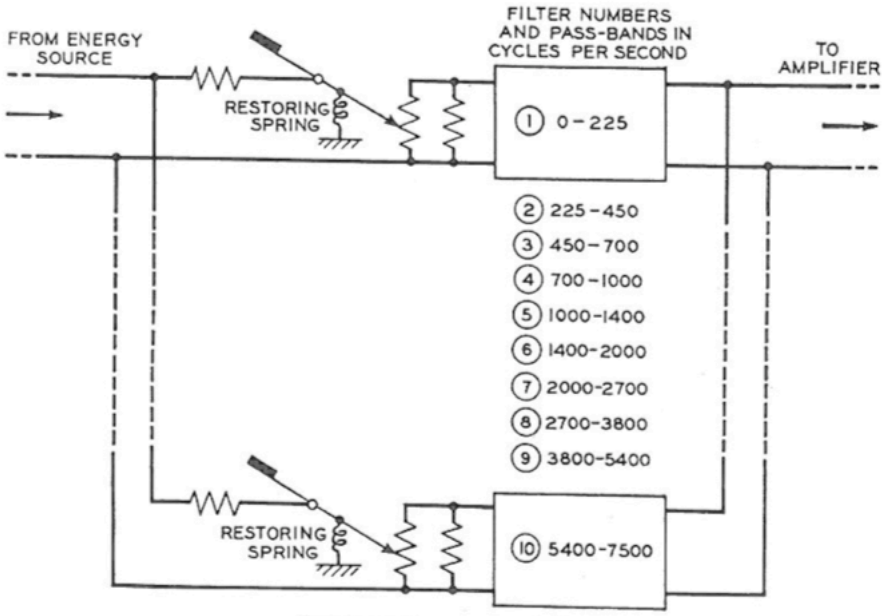


FIGURE 1.9 – *Vue schématique des contrôleurs du Voder (d'après (Dudley et al., 1939))*

Du fait de la complexité apparente de contrôle du Voder, un certain nombre de simplifications ont été apportées. Par exemple, lorsque la fréquence du Voder diminue, le volume est lui aussi abaissé, de la même manière que l'humain le fait naturellement. Le Voder disposait également d'une touche pour le chuchotement ou le silence. La première touche permettait de produire des sons non voisés, la seconde d'abaisser significativement le volume de sortie pour certains sons.

En outre, la production des consonnes plosives était simplifiée par l'utilisation de trois touches dont le rôle consistait à produire un son prototypique constitué d'une période d'attaque, d'une période de silence, d'un sursaut de bruit et enfin du filtrage de ce bruit par les résonances de la voyelle suivante (c'est-à-dire des formants de cette voyelle). Le fonctionnement de ces trois touches étaient le même, à la différence près que les durées des différentes étapes étaient modifiées, conformément aux caractéristiques des différentes plosives considérées. Cette simplification était indispensable, du fait de la vitesse élevée du mouvement des articulateurs de la bouche lors de la production de plosives, supérieure à la vitesse maximale atteignable avec un geste manuel.

Pour la réalisation du Voder, H. Dudley aura su faire preuve d'une grande ingéniosité pour repousser les limites de la synthèse vocale de son époque. Nombre des choix réalisés reposent en réalité sur l'empirisme, plus que sur une analyse approfondie de la production vocale, justement parce que le Voder était intrinsèquement lié à l'utilisateur de façon interactive, permettant ainsi de corriger les dysfonctionnements éventuels au fur et à mesure de son élaboration.

Les différents exemples de machines vocales présentées précédemment représentent non seulement un intérêt historique, car elles comptent sans nul doute parmi les plus célèbres machines vocales ayant été créées, pour la première à l'ère mécanique et pour la seconde à l'ère électrique, mais surtout, on s'aperçoit de manière frappante que le geste (ou tout du moins le contrôle) humain était largement présent. Et il est également surprenant de s'apercevoir que cette "tradition" ne se soit pas perpétuée à l'ére électronique ou numérique.

En effet, mis à part quelques rares exemples, que nous présentons ci-après, le nombre de synthétiseurs vocaux contrôlés par un opérateur est dramatiquement faible, alors même que la quantité de synthétiseurs a, quant à elle, significativement augmentée, tant dans leurs techniques (synthèse par formant, par modèle physique, par concaténation d'unités acoustiques ...) que dans leurs applications (serveurs vocaux, multi-langues, aide au handicap, recherche ...). Tout le monde dispose effectivement aujourd'hui d'un synthétiseur vocal (commercial ou non) installé sur son ordinateur, mais personne (mis à part pour des applications de recherche ou musicales) ne possèdent de moyen de contrôle sur cette synthèse, si ce n'est le texte à synthétiser.

Nous présentons donc dans la suite de ce chapitre quelques machines exemplaires de synthèse vocales numérique qui sont contrôlées par le geste, et ayant essayé de poursuivre la voie tracée par les illustres von Kempelen et Dudley.

1.3.3 Glove Talk

Le Glove-Talk [4] est un système qui a été développé par Sidney Fels, et dont la première version remonte à 1991 (Fels, 1991; Fels and Hinton, 1998). Le principe de base de ce système était d'utiliser un gant de donnée pour commander un synthétiseur à formants (en l'occurrence le synthétiseur DECTalk (Klatt, 1982) pour la première version). Le but de ce projet n'était donc pas de reconstruire un synthétiseur, mais bien d'adopter de nouvelles stratégies pour effectuer le contrôle et le réglage des différents paramètres d'un synthétiseur par règles, souvent fastidieux lorsqu'il est réalisé de façon automatisée.

La méthodologie privilégiée par Fels pour effectuer ce contrôle a consisté à utiliser des réseaux de neurones, qui fournissaient selon lui l'avantage de pouvoir s'adapter à n'importe quel utilisateur, après une courte phase d'apprentissage par le système. Un autre avantage souligné par l'auteur est celui de la robustesse des réseaux de neurones en terme de reproductibilité des mots à synthétiser.

Le principe du contrôle dans la première version du système était le suivant : à la manière de la langue des signes (toutes proportions gardées), l'utilisateur pouvait produire un mot selon la

4. "le gant parlant"

configuration de la main portant le gant. Cette configuration principale était celle de la forme de la main. Ensuite la direction de la main, sa vitesse et sa trajectoire définissaient respectivement la fin de mot, la vitesse d'articulation et l'accentuation du mot. Notons également que l'orientation de la main était utilisé pour décider, d'après la forme de la main, du mot à synthétiser, sur la base d'un antagonisme sémantique. Par exemple, les mots "aller" et "venir" étaient réalisés grâce au même signe mais étaient différenciés par le fait que la paume de la main était orientée soit vers le haut soit vers le bas. Une description schématique de ce système est donnée sur la figure 1.10

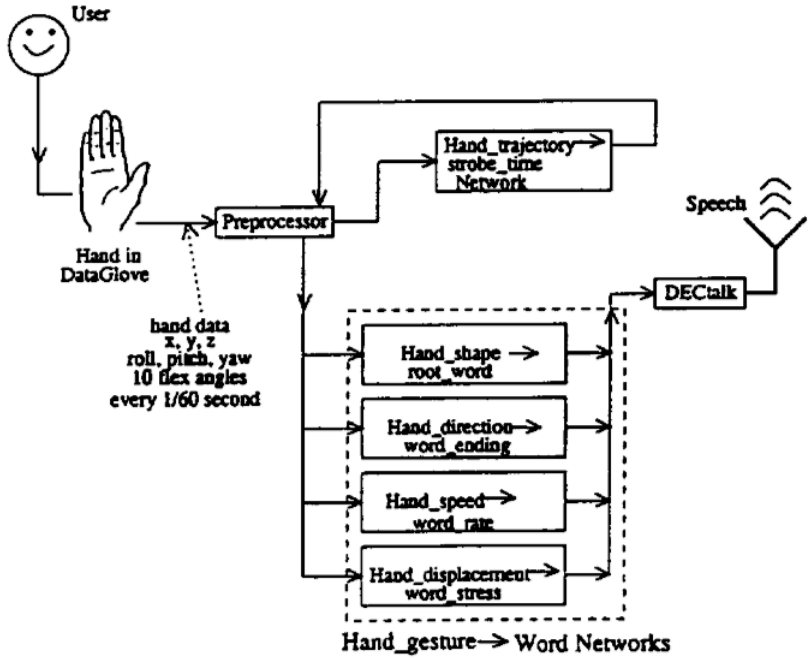


FIGURE 1.10 – *Première version du Glove-Talk de Sidney Fels (Fels, 1991)*

Une autre remarque importante soulevée par l'auteur est celle de la bande passante requise suivant la granularité souhaitée pour le synthétiseur. Tandis que la production d'un phonème se situe en moyenne autour de $100\ \mathrm{ms}$, celle d'un mot se trouve plutôt autour de $300\ \mathrm{ms}$. La production de parole au niveau des phonèmes (i.e. au niveau articulatoire) demande donc une bande passante plus large que pour des mots, et de fait, une plus grande dextérité de la part de l'utilisateur. En revanche, en se plaçant au niveau articulatoire, il est théoriquement possible de produire n'importe quel mot ou phrase, grâce à une articulation adéquate des phonèmes consécutifs. Comme la première version du gant parlant opérait au niveau des mots, il n'était alors possible de produire qu'environ 200 mots, peu suffisants pour imaginer converser avec un tel système.

La deuxième version du gant parlant "Glove-Talk II" visait ainsi à combler cette lacune, grâce

au développement d'un synthétiseur articulatoire *ad-hoc* et en permettant également la production d'une parole plus riche d'un point de vue du vocabulaire. D'autre part, la première version ne permettait pas non plus de modifier la hauteur et le volume de la parole générée.

Le Glove-Talk II, schématisé sur la figure 1.11, a été développé dans le but de répondre aux limitations imposées par la première version. On observe ainsi sur cette figure qu'il n'y plus que 3 réseaux de neurones au lieu de 4. Deux d'entre eux sont dédiés à la paramétrisation du synthétiseur à formants, pour les voyelles d'une part, pour les consonnes d'autre part. Le dernier réseau sert quant à lui à décider si l'utilisateur souhaite produire une consonne ou une voyelle.

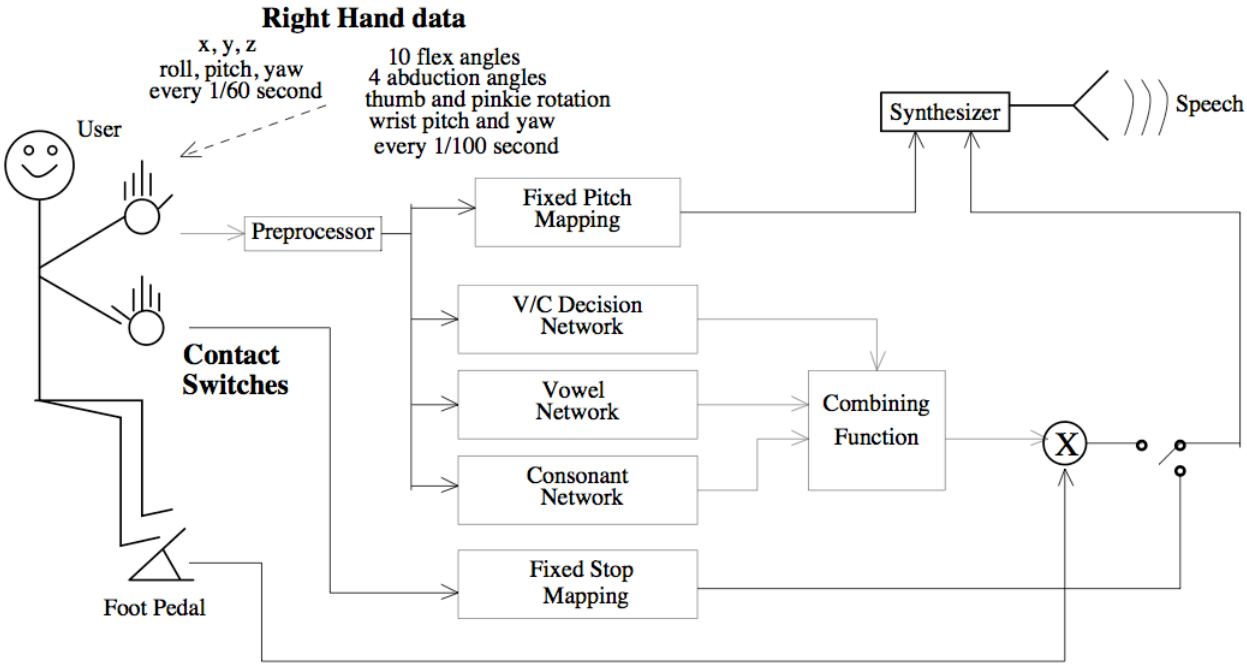


FIGURE 1.11 – *Description schématique du Glove-Talk* II *de S. Fels & G. Hinton (Fels and Hinton, 1998)*

En outre, ce synthétiseur dispose d'un contrôle de la hauteur grâce à une correspondance directe de la hauteur (dans l'espace) de la main de l'utilisateur avec la fréquence fondamentale du synthétiseur. Une pédale permet aussi de démarrer ou arrêter la synthèse.

Enfin, comme souligné par l'auteur, certains sons trop rapides pour être contrôlés par un humain (b,d,g,j,p,t,k,ch,ng) sont gérés par de simples points de contacts situés sur le gant[5]. En d'autres termes, un contrôle non linéaire est réalisé pour la production de ces sons.

5. par exemple le contact du pouce avec l'index ou le majeur.

Les choix et les différents développements effectués pour la réalisation des Glove-Talk nous révèle un problème loin d'être anodin, à savoir que la liberté d'expression de l'utilisateur se fait au détriment de sa facilité d'exécution. Car, au-delà de la qualité intrinsèque du synthétiseur (base sur laquelle il ne convient pas de juger ce système), cette investigation nous montre qu'il est tout simplement trop difficile de contrôler aisément des mouvements articulatoires avec le geste pour la simple et bonne raison que l'exécution séquentielle des mouvements articulatoires est trop rapide pour qu'ils puissent être réalisés par le geste manuel (ou par le geste d'un membre quelconque plus généralement).

Il convient ainsi, afin de réaliser le contrôle gestuel, d'extraire des paramètres variant sur une échelle temporelle suffisamment longue avant de pouvoir les contrôler par un mouvement donné.

Avant de passer à la description de deux derniers systèmes développés récemment, notons que les trois exemples précédents ont concentrés leurs efforts essentiellement (voire quasi-exclusivement) sur le contrôle par le geste des mouvements articulatoires, et très peu sur le contrôle de la source vocale. Ceci se comprend aisément de façon historique, par le fait que le premier souci des chercheurs en synthèse vocale était au départ de produire un son compréhensible avant de se soucier de produire un son naturel et/ou expressif. Cependant, avec l'apparition des synthétiseurs par concaténation d'unités acoustiques, mais également à l'heure où l'analyse des sons de parole est plus approfondie en termes d'analyse formantique, il paraît plus approprié de se confronter au naturel et à l'expressivité de la voix, c'est-à-dire, donc, plus simplement de s'attacher au contrôle de la source vocale.

1.3.4 SPASM

Le système SPASM[6] développé par Perry R. Cook, pour son travail de thèse (Cook, 1991), présente deux intérêts principaux pour notre étude. Le premier est que, ce système, en plus des paramètres articulatoires, dispose d'un contrôle avancé de la source glottique. D'autre part, ce système, sous la forme d'un logiciel à la base, a servi ensuite pour un certain nombre de contrôleurs pour la production de voix chantée.

Ce système, de part sa vocation à réaliser une synthèse de voix chantée, comporte nécessairement des caractéristiques absentes des systèmes présentés précédemment. Le système SPASM fournit également la possibilité de synthèse à partir du texte. Il repose sur une modélisation du conduit vocal par un modèle à guide d'onde numérique. Concernant les cavités de résonance, deux guides d'ondes fonctionnent en parallèle, l'un pour la modélisation de la forme du conduit vocal, le second pour la modélisation des anti-résonances créées au niveau du conduit nasal.

6. Singing Physical Articulatory Synthesis Model

La source glottique est générée grâce à des tables d'ondes contenant les différentes formes d'onde de débit glottique. Les turbulences d'air présentes dûes au passage à travers la glotte sont générées par une source de bruit modulée par l'onde de débit glottique.

Le système permet en outre d'enregistrer des voyelles grâce à un microphone, d'en faire l'analyse spectrale et ainsi de réinjecter les caractéristiques formantiques d'une voix particulière pour la resynthèse. La figure 1.12 décrit le fonctionnement détaillé du synthétiseur SPASM.

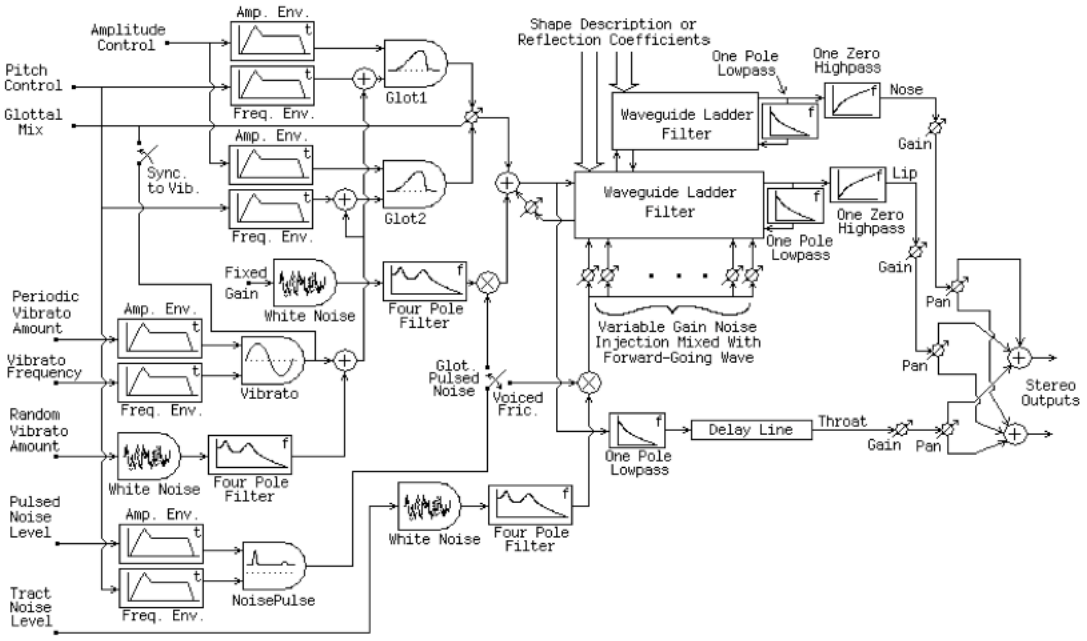


FIGURE 1.12 – *Description schématique du système de synthèse SPASM (Cook, 1992)*

Le modèle SPASM fonctionne en temps réel, mais sa complexité apparente oblige à devoir contrôler une quarantaine de paramètres de synthèse pour le faire fonctionner. Aussi, des études plus récentes menées par P. Cook l'ont conduit à utiliser des contrôleurs dédiés afin de permettre un contrôle facilité du système de synthèse.

L'une de ces implémentations (Cook and Leider, 2000) a consisté à utiliser un accordéon pour contrôler le souffle, la hauteur et l'articulation. La hauteur tempérée était ainsi contrôlée par le clavier de l'accordéon, tandis que le contrôle fin et le vibrato étaient gérés par une bande tactile linéaire avec l'autre main. Le souffle était naturellement contrôlé par le soufflet de l'accordéon. Enfin les différentes configurations de voyelles et de consonnes étaient activées par les nombreux boutons présents sur l'accordéon. On peut voir une photographie de l'instrument "SqueezeVox" sur la figure 1.13, qui contient sur l'instrument lui-même une paire d'enceintes conférant à l'instrument une certaine autonomie.

FIGURE 1.13 – *L'instrument augmenté "SqueezeVox" (d'après (Cook and Leider, 2000))*

Deux autres réalisations de contrôle gestuel de la synthèse vocale sont rapportés par P. Cook (Cook, 2005). La première, le COWE [7] (non représenté ici) consiste en une interface avec un capteur de pression du souffle, un capteur de pression linéaire, une glissière à pouce et un accéléromètre à deux dimensions. Ce dernier permet de se déplacer dans l'espace vocalique. Le capteur de pression linéaire, quant à lui, prend la main sur l'accéléromètre pour charger des séquences de phonèmes ou de mots. Etant donné la relative difficulté à produire un souffle assez long et soutenu, un ballon gonflable a ensuite été ajouté pour fournir un réservoir d'air suffisant, à la manière d'une cornemuse.

Enfin, la dernière interface utilisée rapportée par P. Cook est le VOMID [8] représenté sur la figure 1.14 suivante.

7. Controller, One With Everything
8. Voice-Oriented Melodica Interface Device

FIGURE 1.14 – *Le VOMID en action (d'après (Cook, 2005))*

Ce nouvel instrument avait pour but de pallier deux inconvénients des deux interfaces précédentes, à savoir l'encombrement réduit et la possibilité d'avoir un contrôle précis des notes par un clavier. Comme son nom le suggère, cet instrument est basé sur la facture du mélodica. Un capteur de pression du souffle a ainsi été ajouté à un clavier maître MIDI. Celui-ci fonctionne dans les deux sens pour fournir à la fois la phonation lorsque l'on souffle ou la respiration lorsque l'on aspire. Comme pour le SqueezeVox, un capteur de pression linéaire complète les touches du clavier MIDI.

1.3.5 Le voicer

Le voicer est un instrument de synthèse de voyelles chantées développé par L. Kessous pendant sa thèse de doctorat (Kessous, 2004). Trois implémentations différentes du Voicer ont été mises en place, la première avec une source en forme de dent de scie, la seconde utilisant 5 FOFs [9] (Rodet, 1984) en parallèle et la dernière avec un modèle de source glottique.

Nous ne décrirons ici que la dernière version, car c'est celle la plus proche de l'implémentation que nous avons réalisé, et nous concentrerons particulièrement notre attention sur le contrôle des paramètres du modèle. Cette version particulière du Voicer utilise le modèle de source glottique R++ de Veldhuis (Veldhuis, 1996) que nous aurons l'occasion de décrire en détails dans le chapitre 3 consacré à la description des différents modèles de source glottique.

Concernant le contrôle de la hauteur, celui-ci est réalisée selon une disposition circulaire découpée en douze demi-tons qui représentent une octave. Il est ensuite possible de changer d'octave soit de façon continue en poursuivant le mouvement circulaire dans un sens ou dans l'autre pour monter

9. Forme d'Onde Formantique

ou descendre les octaves, soit de façon discontinue grâce à un bouton situé sur le joystick. On peut voir une illustration de ce système sur la figure 1.15, ici dans la configuration tablette graphique plus joystick.

FIGURE 1.15 – *Le Voicer en action (d'après (Kessous, 2004))*

Concernant le contrôle des voyelles, ce dernier est effectué par le joystick, en se déplaçant dans un plan, où seront placées les voyelles sous la forme d'un vecteur de paramètres. Une interpolation est alors effectuée pour se déplacer dans l'espace vocalique, pouvant prendre trois formes différentes : linéaire, attractif ou répulsif. La classification des voyelles est réalisée selon le lieu d'articulation et le degré de constriction du conduit vocal. Cette méthode revient finalement à se déplacer dans l'espace $(\mathrm{F}_1, \mathrm{F}_2)$ définissant le triangle vocalique.

Pour ce qui est du contrôle de la source glottique, L. Kessous nous rapporte qu'il ne modifie vraiment que l'intensité, grâce à une correspondance entre la pression du stylet et le quotient ouvert, l'asymétrie et l'amplitude de l'onde de débit glottique [10]. Cette correspondance a été réalisée par tâtonnement, et ne repose pas sur une étude formelle des plages de co-variations de ces paramètres. En outre, on pourra regretter qu'aucune manipulation d'une autre dimension de qualité vocale ne soit présente.

Pour le choix des périphériques utilisés, on notera que le joystick ne sera pas utilisé pour le contrôle de la hauteur tonale, car jugé trop imprécis. En revanche, il sera privilégié pour le contrôle

10. se reporter au chapitre 3 pour une description détaillée des notions de qualité vocale

des voyelles, du fait qu'il retourne à une position médiane lorsqu'il n'est pas manipulé. On notera également, comme on le voit sur l'illustration précédente, que ce système dispose d'un retour visuel, qui sera déporté ou non suivant que l'on utilise une tablette graphique simple ou un écran tactile. La pression exercée sur l'interface est alors soit celle du stylet, soit celle du doigt sur l'écran.

Chapitre 2

Modification prosodique de la parole par contrôle gestuel

Sommaire

2.1 Introduction **35**
2.2 L'algorithme PSOLA **37**
2.2.1 Les signaux d'analyse à court terme 37
2.2.2 Les signaux de synthèse à court terme 38
2.2.3 Le signal de synthèse final 39
2.2.4 Le calcul des marqueurs de synthèse 39
2.2.5 Le choix de la fenêtre d'analyse 41
2.3 L'algorithme PSOLA en temps réel **43**
2.3.1 Les étapes 43
2.3.2 Les contraintes temps-réel 48
2.3.3 Le calcul des instants de synthèse 52
2.4 Première expérience d'imitation mélodique **54**
2.4.1 Evaluation d'un système de réitération intonative contrôlé par la main . . . 54
2.4.2 Préambule 58
2.4.3 Calliphonie : les premiers pas 59
2.4.4 Résultats de l'expérience d'imitation 67
2.4.5 Discussion et Conclusions Partielles 68
2.5 Deuxième expérience d'imitation mélodique **70**
2.5.1 Le corpus 70
2.5.2 Les sujets 70
2.5.3 L'interface 71
2.5.4 Le protocole 71
2.5.5 Les résultats 72

2.5.6 Analyse gestuelle 79
2.6 Applications **85**
2.6.1 Enrichissement de base de données 85
2.6.2 Voix chantée 88
2.7 Conclusions du chapitre **90**

2.1 Introduction

Ce premier chapitre a pour but de présenter les développements et les expériences que nous avons pu mener dans le domaine de la modification prosodique, et plus particulièrement sur la modification en temps réel de l'intonation et du rythme de la parole.

Rappelons, à toutes fins utiles, que la prosodie est par essence de nature multidimensionnelle et perceptive. Le plus souvent, la prosodie est définie comme le résultat des variations de la fréquence (l'intonation), de la durée (le rythme) et de l'énergie (l'intensité) du signal acoustique de parole. D'autre part, récemment, certains raffinements ont été proposées à cette notion, afin de prendre également en compte les variations de qualité vocale (Campbell and Mokhtari, 2003), et du degré de réduction des voyelles (Pfitzinger, 2006). En résumé, il est possible d'attribuer, en première approximation, cinq dimensions distinctes à la prosodie. Pourquoi la prosodie comporte-t-elle un intérêt pour l'étude de l'expressivité de la voix ? Tout simplement, parce que la prosodie représente ce caractère de la voix, portant l'expression vocale. En d'autres termes, l'expression vocale (attitude, humeur, émotion ...) est exprimée et reconnue selon l'évolution des caractéristiques prosodiques précitées.

Néanmoins, dans le cadre de la synthèse vocale, toutes les dimensions prosodiques ne sont pas d'égales accessibilité et modification. En effet, encore aujourd'hui, l'analyse des différentes composantes constituant la qualité vocale, ne sont pas aisément extraites d'un signal de parole. En revanche, plusieurs modèles de source glottique existent, et il est possible de synthétiser un signal vocal à partir de ces modèles et d'en faire varier les paramètres afin d'obtenir différentes "qualités vocales". Mais nous sortons ici du cadre de l'analyse pour la synthèse pour entrer dans celui de l'analyse par la synthèse. Nous aurons l'occasion de revenir plus en détails sur ces notions dans le chapitre 3, justement consacré à la synthèse de source glottique.

Concernant l'énergie du signal, on pourrait se dire naïvement qu'une simple augmentation du volume du signal pourrait suffire à rendre la voix plus "forte". Il n'en est rien : un enregistrement de voix chuchotée joué à un volume sonore élevé ne produit pas une voix criée, et vice-versa. L'intensité de la parole, est fortement liée à la dimension d'effort vocal, faisant elle-même partie de la qualité vocale. Nous verrons également dans le prochain chapitre, le lien existant entre la hauteur d'un signal de parole et son intensité, d'après ce que l'on dénomme couramment un phonétogramme (Henrich, 2006).

Le degré de réduction des voyelles fait quant à lui référence au phénomène de coarticulation et correspond prosaïquement à la réalisation approximative des formants d'une voyelle donnée. Ainsi, lorsqu'un locuteur augmente ostensiblement son débit de parole, les valeurs standards des formants pour ce locuteur ne sont pas atteintes (ou seulement partiellement) et l'on observe une réduction du triangle vocalique et donc une dérive des différentes classes phonétiques de voyelles associées.

Pris individuellement, le signal d'une voyelle peut ne pas être perçu comme étant la voyelle auquel il fait référence. Ce n'est alors qu'en présence du contexte phonétique environnant qu'un auditeur peut en déduire la voyelle qui a été prononcée. Cette notion ne sera cependant traitée que de façon subsidiaire dans ce présent manuscrit, de même que les phénomènes d'articulation du conduit vocal de façon plus générale.

Enfin il nous reste l'intonation et le rythme, et c'est de ces deux dimensions dont il sera question dans ce chapitre. Depuis quelques années déjà, et surtout avec l'apparition des méthodes d'addition-recouvrement de fenêtres temporelles (Hamon et al., 1989; Moulines and Charpentier, 1990; Moulines and Laroche, 1995), est apparu un certain nombre d'outils de traitement du signal permettant de modifier la hauteur et/ou la durée d'un signal enregistré. Ces techniques de modification ont été largement utilisées (et le sont encore) au sein de la communauté de recherche en synthèse de parole ou en analyse de la prosodie. Cela dit, parmi ces exemples, peu d'entre eux sont capables de réaliser de telles modifications en temps réel. Et, à notre connaissance, aucun système n'utilise de contrôle gestuel comme modalité en entrée.

Nous essaierons donc ici de nous attacher à décrire les difficultés rencontrées par une implémentation en temps réel, et les avancées expérimentales que nous avons pu accomplir grâce à ce nouvel outil. Ce chapitre sera organisé de la manière suivante : une première partie théorique et technique, puis une seconde partie dédiée à l'expérimentation et la validation de notre travail. Nous décrirons ainsi les bases théoriques de l'algorithme PSOLA, de manière ensuite à décrire les détails de l'implémentation de notre version de cette algorithme en temps réel. Puis, nous présenterons deux expériences menées sur la modification de l'intonation de parole enregistrée, et enfin une validation perceptive de la modification combinée de la hauteur et de la durée de phrases synthétiques (i.e. générées par un système de synthèse par concaténation d'unités acoustiques).

2.2 L'algorithme PSOLA

Comme nous le verrons dans les sections suivantes de ce chapitre, pour les besoins de nos expériences, nous avons utilisé l'algorithme TD-PSOLA afin de pouvoir réaliser certaines modifications prosodiques (intonation et durée). Nous avons ainsi été amené à réaliser une version en temps réel de cet algorithme, que nous décrirons dans la section 2.3.

Le but de cette présente section est donc de décrire le fonctionnement de l'algorithme original TD-PSOLA, tel que le décrivent E. Moulines & J. Laroche dans leur article de 1995 (Moulines and Laroche, 1995). Cette technique d'addition-recouvrement de fenêtres synchrones à la période, ou PSOLA (Pitch-Synchronous OverLap-Add), doit être réalisée selon trois étapes, que nous allons reprendre et décrire ici, à savoir :

1. Le passage du signal de parole aux signaux d'analyse à court terme
2. Le passage des signaux d'analyse à court terme aux signaux de synthèse à court terme, et
3. Le passage des signaux de synthèse à court terme au signal de parole de sortie.

Notons que pour simplifier et s'accorder sur les multiples notations (parfois contradictoires) relatives à l'algorithme PSOLA, nous reprendrons ici le formalisme de cet article pour éviter toute confusion éventuelle.

2.2.1 Les signaux d'analyse à court terme

Si l'on note le signal de parole original par $x(s)$, les signaux d'analyse à court terme vont correspondre à la simple multiplication du signal de parole original par des fenêtres d'analyse adéquates, centrées chacune autour de marqueurs de période d'analyse. Ces marqueurs ou instants d'analyse sont choisis de manière à correspondre, pour une période du signal de parole donnée, au maximum énergétique pour cette période. Nous verrons dans la section 2.3, les choix que nous avons réalisés quant à la détermination de ces instants d'analyse.

Ces instants d'analyse seront notés $t_a(s)$ par la suite, et correspondent à des instants synchrones à la fréquence fondamentale instantanée pour les parties voisées et à des instants régulièrement espacés (taux constant) pour les parties non voisées. Les signaux d'analyse à court terme sont alors exprimés de la manière suivante :

$$x(s,n) = h_s(n) \times x(n - t_a(s)) \tag{2.1}$$

où $h_s(n)$ est la fenêtre d'analyse centrée sur l'instant d'analyse $t_a(s)$ correspondant, et $x(s,n)$ sera le signal d'analyse à court terme. La longueur T de la fenêtre d'analyse est proportionnelle à la période instantanée $P(s)$, c'est-à-dire que $T = \mu P(s)$. Les caractéristiques standards pour la version temporelle de l'algorithme PSOLA sont $\mu = 2$, avec des fenêtres de Hann pour l'analyse.

2.2.2 Les signaux de synthèse à court terme

Une fois que les signaux d'analyse à court terme ont été extraits, nous disposons d'une série de signaux fenêtrés, dont les longueurs respectives dépendent de leurs périodes instantanées. Il convient alors ensuite de "copier" ces fenêtres aux instants de synthèse adéquats, qui eux sont dépendants à la fois de la valeur de modification de hauteur et de la valeur de modification temporelle souhaitées.

Toutefois, ces deux valeurs de modification peuvent dans l'idéal prendre des valeurs réelles et ainsi, les instants de synthèse calculées à partir des ces valeurs ne correspondront pas exactement, dans la grande majorité des cas, à un instant d'analyse précis. Ces instants de synthèse bruts sont ainsi appelés marqueurs de hauteur virtuels ou instants de synthèse virtuels et seront notés $t_s^{'}(u)$.

Ces instants virtuels vont en effet être compris la plupart du temps entre deux instants d'analyse, c'est-à-dire appartenir à l'intervalle $[t_a(s), t_a(s+1)]$. Ainsi, il convient pour calculer les instants de synthèse réels $t_s(u)$, de réaliser un choix de correspondance entre les instants de synthèse réels et les instants d'analyse. Ce choix peut être réalisé de différentes manières, la plus simple consiste à choisir, pour un instant virtuel $t_s^{'}(u)$ donné, l'instant d'analyse le plus proche $t_a(s) \oplus t_a(s+1)$, et ainsi à chaque instant de synthèse correspondra un instant d'analyse unique.

Une autre solution plus élaborée consiste à réaliser une pondération entre les deux signaux d'analyse à court terme encadrant l'instant de synthèse virtuel. Alors, le calcul du signal de synthèse $y(u,n)$ sera effectué selon la formule suivante :

$$y(u,n) = (1-\alpha_u)x(s,n) + \alpha_u x(s+1,n) \tag{2.2}$$

où $y(u,n)$ correspond au signal de synthèse à court terme associé à l'instant de synthèse $t_s(u)$ et le facteur α_u représente le facteur de pondération entre les deux fenêtres d'analyse successives. α_u est calculé à partir de l'instant de synthèse virtuel, de la manière suivante :

$$\alpha_u = \frac{t_s^{'}(u) - t_a(s)}{t_a(s+1) - t_a(s)} \tag{2.3}$$

Prenons un exemple pour mieux expliciter cette étape. Supposons que l'on souhaite modifier la hauteur du signal original d'un facteur $1/2$ et sa durée d'un facteur $4/3$. Supposons que l'on se trouve à un instant de synthèse réel $t_s(u)$ pour lequel on vient juste de copier une fenêtre d'analyse $x(s,n)$. Le prochain instant de synthèse sera alors situé à une durée égale à $1/2 \times 4/3 = 2/3$ de la période instantanée d'analyse $P(s) = t_a(s+1) - t_a(s)$. L'instant de synthèse $t_s(u+1)$ sera donc situé à l'instant $t_s(u) + 2/3 \times P(s)$ et l'on devra alors réaliser le choix de la fenêtre d'analyse à court terme à copier en lieu et place.

L'instant de synthèse virtuel $t_s'(u)$ sera donc situé plus proche de $t_a(s+1)$ que de $t_a(s)$. On peut alors selon la première méthode copier à l'instant de synthèse réel $t_s(u+1)$ le signal d'analyse à court terme de l'instant $t_a(s+1)$ ou alors, selon l'équation 2.2, réaliser une pondération linéaire entre les deux signaux d'analyse à court terme consécutifs, situés aux instants respectifs $t_a(s)$ et $t_a(s+1)$, avec une valeur $\alpha_u = 2/3$.

Selon la première méthode, pour l'exemple donné, chaque signal d'analyse à court terme sera copié deux fois, à des instants de synthèse réels plus resserrés. Selon la seconde approche, les contributions successives pour les signaux d'analyse à court terme, seront les suivantes $((1-\alpha_u), \alpha_u) = (1,0), (1/3, 2/3), (2/3, 1/3), (0,1)$ et ainsi de suite.

L'avantage de la seconde méthode est de fournir un signal de synthèse "plus doux". En effet, du fait des facteurs de modifications temporel et fréquentiel, il arrive des situations où l'on doive soit sauter, soit répéter certains signaux d'analyse à court terme. Dans le premier cas, la deuxième méthode fournit au signal de synthèse une certaine "mémoire" plus réaliste, le signal de parole étant un signal certes aléatoire globalement mais fortement déterministe localement (deux périodes fondamentales voisées consécutives du signal n'auront jamais des valeurs grandement différentes). Dans le second cas, cette méthode permet d'éviter une trop grande redondance, en conservant un certain jitter naturel dans le signal de synthèse (bien qu'il puisse être différent du jitter local du signal d'analyse).

2.2.3 Le signal de synthèse final

Après l'obtention de $y(u,n)$, on dispose d'une série de signaux de synthèse à court terme, disposés selon les instants de synthèse d'après les valeurs de modification de hauteur et de durée. Il ne reste plus alors pour obtenir $y(n)$ qu'à ajouter tous ces signaux pour obtenir le signal de synthèse final. Si l'on fait le choix de réaliser une pondération, il faut prendre en compte les contributions respectives des deux signaux d'analyse à court terme consécutifs. Après cette étape, le signal de synthèse final $y(n)$ est calculé, dont la hauteur et la durée sont modifiées par rapport au signal original, selon les valeurs de modification, d'après le calcul que nous allons spécifier dans le paragraphe suivant.

2.2.4 Le calcul des marqueurs de synthèse

Modifications de hauteur et de durée combinées

Nous noterons, en accord avec la notation de Moulines & Laroche (Moulines and Laroche, 1995), $\beta_s = \beta(t_a(s))$ le facteur de modification de hauteur, et $\alpha_s > 0$ le facteur de modification temporelle, pour les parties voisées du signal (nous traiterons des parties non voisées dans le paragraphe

suivant).

En notant $P(t_a(s)) = t_a(s+1) - t_a(s)$ la période fondamentale entre les instants d'analyse $t_a(s)$ et $t_a(s+1)$, la fonction qui, à tout instant t associe $P(t)$, est une fonction de contour mélodique constante par morceaux, qui est définie de la manière suivante :

$$P(t) = P(t_a(s)), \quad avec \quad t_a(s) \leq t < t_a(s+1) \tag{2.4}$$

Les marques de synthèse doivent également être positionnées de manière synchrone à la hauteur, selon la fonction de contour mélodique de synthèse $t \mapsto P'(t)$. Le problème consiste alors à trouver une série de marques de synthèse $t_s(u)$ telle que $t_s(u+1) = t_s(u) + P'(t_s(u))$ et où $P'(t_s(u))$ est approximativement égale à $1/\beta(t_s(u))$ fois la hauteur du signal original autour de l'instant de synthèse $ts_(u)$, soit :

$$P'(t_s(u)) \approx \frac{P(t_s(u))}{\beta(t_s(u))} \tag{2.5}$$

Les modifications de durée sont légèrement plus complexes que les modifications de hauteur en ce sens que le signal original et le signal de synthèse final calibré ne partage pas le même axe temporel. La modification de durée est spécifiée en associant à chaque marque d'analyse un facteur de modification temporelle α_s, à partir duquel la fonction de correspondance temporelle, qui à tout instant t associe $D(t)$, peut être déduite, de la manière suivante :

$$\begin{aligned} D(t_a(1)) &= 0 \\ D(t) &= D(t_a(s)) + \alpha_s(t - t_a(s)), \quad lorsque \quad t_a(s) \leq t < t_a(s+1) \end{aligned} \tag{2.6}$$

Il est clair que la fonction $D(t)$ est une fonction linéaire par morceaux strictement croissante, car α_s est strictement positif. On souhaite, tout en modifiant la durée du signal original conserver malgré tout le contour mélodique. Le contour mélodique sera alors défini comme la fonction, qui à tout instant t associe la valeur $P'(t) = P(D^{-1}(t))$: la hauteur du signal modifié temporellement à l'instant t doit être proche de la hauteur dans le signal original à l'instant $D^{-1}(t)$. Nous devons ensuite trouver une série de marques de synthèse $t_s(u)$ telle que $t_s(u+1) = t_s(u) + P'(t_s(u))$. Pour résoudre ce problème, il est utile de considérer la série de marques de synthèse virtuelles $t_s'(u)$ dans le signal original reliées aux marques de synthèse réelles de la manière suivante :

$$t_s(u) = D(t_s'(u)) \quad soit \quad t_s'(u) = D^{-1}(t_s(u)) \tag{2.7}$$

En supposant que $t_s(u)$ et $t_s'(u)$ sont connus, nous cherchons à déterminer $t_s(u+1)$ et $t_s'(u+1)$, tels que la quantité $t_s(u+1) - t_s(u)$ est approximativement égale à la période instantanée dans le signal original au temps $t_s'(u)$.

La valeur $P'(t_s(u) = t_s(u+1) - t_s(u)$ tenant compte à la fois de la valeur de modification de

hauteur et de durée, est ensuite calculée grâce à l'équation intégrale suivante :

$$\begin{aligned} t_s(u+1) - t_s(u) &= \frac{1}{t'_s(u+1) - t'_s(u)} \int_{t'_s(u)}^{t'_s(u+1)} \frac{P(t)}{\beta(t)} dt \\ avec \quad \beta(t) &= \beta_s, \quad pour \quad t_a(s) \leq t < t_a(s+1) \end{aligned} \tag{2.8}$$

Selon cette dernière équation, la période de synthèse $t_s(u+1) - t_s(u)$ à l'instant $t_s(u)$ est égale à la valeur moyenne de la période dans le signal original calculée pendant l'intervalle $t'_s(u+1) - t'_s(u)$. Notons que cet intervalle temporel $t'_s(u+1) - t'_s(u)$ est relié à $t_s(u+1) - t_s(u)$ grâce à la fonction de correspondance $D(t)$. Les fonctions $P(t)$, $\beta(t)$ et $D(t)$ étant toutes continues par morceaux, l'équation intégrale ci-dessus peut être calculée simplement.

Modification temporelle des sons non voisées

Dans les sections précédéntes, nous nous sommes focalisés uniquement sur la modification de sons voisés. Une attention particulière doit être apportée aux segments non voisés. En particulier, lors d'allongements de parties non voisées du signal de parole (telles que des fricatives), le processus introduit un bruit tonal parce que la répétition de segments de "type bruité" engendre une autocorrélation à long terme artificielle sur le signal de sortie, perçue comme une sorte de périodicité. Une solution simple consiste à renverser l'axe temporel à chaque fois que l'algorithme nécesssite la répétition d'un signal à court terme, c'est-à-dire que $x_s(m)$ est remplacé par $x_s(-m)$. Une telle opération préserve le spectre en amplitude à court terme mais change le signe du spectre de phase, réduisant ainsi la corrélation indésirable dans le signal de sortie. Cette approche élimine efficacement le bruit tonal lorsque le facteur de modification temporelle est inférieur à 2 (ce qui est couramment suffisant pour les types de modification rencontrées en pratique). Des précautions doivent également être prises pour les plosives non voisées, afin d'éviter de dégrader les portions transitoires du signal (éclats).

2.2.5 Le choix de la fenêtre d'analyse

La fenêtre d'analyse utilisée par l'algorithme TD-PSOLA doit être telle que la transformée de Fourier à court terme $X(\omega)$ représente une estimée raisonnable de l'enveloppe spectrale du signal. Deux facteurs influencent significativement les caractéristiques de $X(\omega)$: le type et la longueur de la fenêtre d'analyse $h(n)$. Rappelons à ce titre que la troncature du signal par une fenêtre, quelle qu'elle soit, va nécessairement introduire des effets antagonistes d'étalement spectral (lié au lobe principal) et de fuites spectrales (liées aux lobes secondaires). Plus la fenêtre possédera de discontinuités (la fenêtre rectangulaire en étant le cas extrême), plus les fuites spectrales seront importantes et l'étalement faible. Comme mentionné précédemment, la longueur de la fenêtre d'analyse T est proportionnelle à la période fondamentale instantanée $P(s)$ (i.e. $T = \mu P(s)$). Pour une fenêtre standard, la résolution spectrale (largeur du lobe principal) est inversement proportionnel

à la longueur de la fenêtre. La largeur du lobe principal, en fréquence normalisée est égale à $8\pi/T$ pour les fenêtres de Hamming et de Hann, et à $12\pi/T$ pour la fenêtre de Blackman.

Pour $\mu = 2$ (analyse à bande large), la fréquence de coupure de la fenêtre ($4\pi/P(s)$ pour une fenêtre de Hann) est plus élevée que l'intervalle séparant les harmoniques ($2\pi/P(s)$) : la fenêtre d'analyse ne peut donc pas résoudre les harmoniques individuellement. Le spectre d'analyse à court terme $X(\omega)$ est une estimée lissée de l'enveloppe spectrale du signal de parole : la largeur du lobe principal fournit un moyen d'interpolation entre les harmoniques. Inversement, des valeurs plus larges de μ augmentent la résolution de la fenêtre et révèlent la structure harmonique de $X(\omega)$, une propriété néanmoins indésirable pour la méthode TD-PSOLA : le rééchantillonnage de $X(\omega)$ à la hauteur de synthèse $2\pi k/\beta P$ est encline à produire des artefacts audibles dus à l'annulation/atténuation harmonique.

Le type de fenêtre d'analyse utilisée est également important : une fuite spectrale excessive introduit des modification de timbre indésirables en lissant les détails fins de la structure formantique. Cette remarque exclut les fenêtres particulières introduites lors de contributions précédentes telles que la fenêtre trapézoïdale proposée par Lukascewickz et Karjalainen (Lukaszewickz and Kajalainen, 1987), ou dans une moindre mesure, la fenêtre en cosinus asymétrique proposée par Hamon (Hamon et al., 1989) dans la version originale de l'algorithme TD-PSOLA.

Ces artefacts deviennent plus prononcés pour des voix plus aiguës, pour lesquelles les fenêtres d'analyse sont de durée plus courtes et comportent alors des lobes principaux plus larges. Les erreurs typiques sont (i) un élargissement de la largeur de bande formantique, spécialement pour le premier formant, et (ii) une fusion des formants faiblement espacés, bien que cela ne se produise que moins fréquemment. Heureusement, en général, ces effets n'engendrent pas de dégradations de la qualité de la voix de sortie trop sévères, tout du moins lorsque des modifications de hauteurs modérées sont appliquées.

2.3 L'algorithme PSOLA en temps réel

Pour les besoins de nos expériences et de notre application, nous avons été amenés à réaliser une version en temps réel de l'algorithme PSOLA, afin d'obtenir simultanément une modification de la hauteur et de la durée du signal de parole. Cette implémentation a également été pensé en amont dans la perspective du traitement du signal de parole. Aussi, un certain nombre des étapes que nous allons décrire ci-après sont spécifiques au signal de parole et l'extension à d'autres types de signaux, musicaux ou non, demanderait une adaptation adéquate (notamment concernant les attaques de notes pour les sons musicaux).

2.3.1 Les étapes

L'implémentation de l'algorithme PSOLA en temps réel pour la modification d'un signal de parole a requis la réalisation méthodique de plusieurs étapes d'analyse et de traitement du signal de parole, dont certaines lors d'un pré-traitement dans le but d'obtenir la meilleure modification possible en termes de qualité. Nous allons donc décrire ici les étapes de pré-traitement, puis celle de l'algorithme temps réel proprement dit.

Calcul des instants de fermeture glottique

L'algorithme PSOLA, comme nous avons pu le voir dans la section 2.2, repose sur un fenêtrage des signaux à court terme. Afin d'obtenir une synthèse la meilleure possible, il est nécessaire lors de l'étape d'addition-recouvrement que les fenêtres s'additionnent de la façon la plus douce possible. En d'autres termes, il est judicieux de centrer les fenêtres à court terme sur les instants d'énergie maximale, ce qui, pour un signal de parole correspond à l'instant de fermeture glottique ou GCI[1]. Il convient donc de détecter le plus précisément possible ces GCIs, et de les indexer dans le signal de parole original.

Il n'existe pas, à notre connaissance, à l'heure actuelle, de méthode permettant de calculer les GCIs d'un signal de parole en temps réel. Aussi, dans le souci de pas alourdir outre mesure notre algorithme, nous avons choisi de réaliser le calcul des GCIs en amont et en temps différé de l'algorithme temps réel proprement dit.

L'algorithme DYPSA (Kounoudes et al., 2002) est sans doute actuellement l'une des méthodes les plus robustes et précises connues pour le calcul des GCIs. En outre, cet algorithme est disponible sous l'environnement de programmation Matlab, ce qui facilite son implémentation. Nous avons donc opté pour l'utilisation des fonctions Matlab de l'algorithme DYPSA[2]. Nous avons ainsi la

1. Glottal Closure Instant
2. Disponible dans la toolbox Voicebox

possibilité de calculer de manière automatique les GCIs d'un signal de parole et de sauvegarder les résultats dans un fichier texte sous la forme d'une suite d'indices d'échantillons. Deux modifications ont toutefois été apportées au calcul des GCIs afin de pouvoir s'adapter à notre application : (i) le calcul des zones non voisées et (ii) l'application d'un taux constant pour ces zones non voisées, conformément aux spécificités de l'algorithme PSOLA.

Evidemment, le calcul de GCIs pour des zones de parole non voisée n'a pas de sens, puisque par essence la glotte n'entre pas en vibration. Cela dit, d'une part l'algorithme DYPSA nous renvoie tout de même des valeurs pour les périodes non voisées et d'autre part, l'algorithme PSOLA a besoin, comme nous l'avons vu dans la section 2.2 d'un taux constant pour les zones non voisées.

Nous avons ainsi réalisé deux manipulations à partir des résultats de l'algorithme DYPSA originel. La première a consisté à obtenir une mesure suffisamment fiable des périodes de voisement, et donc par complémentarité, de celles de non voisement. L'un des développements récents et robustes de ce type d'algorithme est celui de A. Camacho (Camacho, 2008), également disponible sous la forme d'une fonction Matlab. Pour l'obtention de la meilleure détection possible, nous avons réglé les paramètres de façon à n'étudier uniquement que des fréquences comprises entre 75 et 600 Hz, correspondant typiquement aux valeurs limites de la fréquence fondamentale pour une voix humaine. Le reste des paramètres étaient fixés à leurs valeurs par défaut.

Pour l'obtention de marqueurs à taux constant dans les périodes non voisées, nous analysons en sortie des résultats de DYPSA si deux marques sont espacées d'une distance supérieure à la période fondamentale minimale fixée, à savoir 75 Hz, soit environ 13.5 ms. Si c'est le cas, alors on place une ou plusieurs marques d'analyse de manière isochrone. Par exemple, si deux marques sont espacées de 30 ms, alors on place deux nouvelles marques d'analyse, placées respectivement à 10 et 20 ms de la première marque.

Enfin, pour différencier d'un point de vue algorithmique les marques d'analyses correspondant à des périodes voisées de celles correspondant à des périodes non voisées, on place arbitrairement les valeurs d'échantillons des périodes non voisées à leur valeur opposée (i.e. négative). On se retrouve alors avec un fichier texte composé d'une série d'index dont les valeurs absolues seront nos marques d'analyse (sous la forme d'un index d'échantillon à 44.1 kHz) et dont le signe détermine si cette marque correspond à une période instantanée voisée ou non.

Sur la figure 2.1 suivante, on peut observer le résultat du calcul des GCIs sur un fichier de parole original (en bleu, à gauche) sur la figure de droite (en rouge). On observe ainsi que ce signal de parole dispose de plus de 300 GCIs, dont les valeurs sont leurs indice d'échantillonnage dans la fichier original (échantillonné à 44.1 kHz). Le signe de ces valeurs sur la figure de droite correspond bien aux parties voisées/non voisées du signal de parole original à gauche.

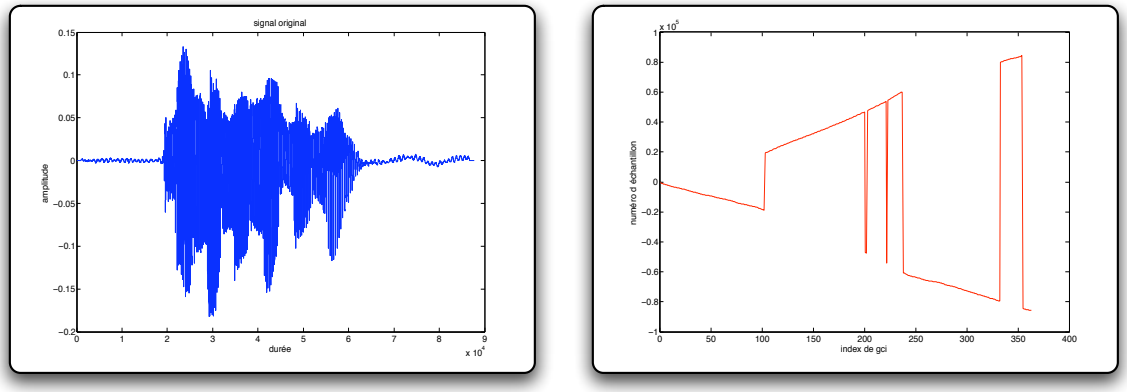


FIGURE 2.1 – *Fichier audio original (à gauche) et fichier de GCIs correspondant (à droite).*

Intégration dans Max/MSP

Nous disposons désormais en entrée de notre algorithme, implémenté sous l'environnement Max/MSP (Max/MSP, 2008), de deux fichiers distincts : (i) notre fichier de parole original et (ii) un fichier texte correspondant, avec les numéros d'échantillons de toutes les marques d'analyse du signal de parole.

A cette étape de notre implémentation, deux précisions doivent être apportées. D'une part, plutôt que de travailler directement sur le fichier de parole original, nous avons décidé de créer un tampon audio interne [3] contenant tous les signaux à court terme du signal original préalablement fenêtrés. Ce choix n'aide en rien d'un point de vue calculatoire ou algorithmique, mais permet en revanche de faciliter grandement l'aspect conceptuel. Dans ce tampon audio, chaque signal à court terme est identifié uniquement par un index et une période. L'index correspond au GCI et la période à la durée allant du GCI précédent au GCI suivant, autrement dit à la période instantanée.

La seconde précision est la suivante : contrairement à la plupart des algorithmes d'addition-recouvrement usuels, le fenêtrage utilisé n'est pas ici symétrique. Nous avons vu, dans la section précédente, que Moulines & Laroche, recommandent l'utilisation de fenêtres de Hann symétriques pour les signaux d'analyse à court terme, car elles fournissent une meilleure reconstruction spectrale. Cependant, pour cette solution, la durée du signal à court terme la plus grande sera égale à deux fois la période instantanée maximale, tandis que pour la solution asymétrique la durée du signal à court terme la plus grande sera nécessairement plus petite. Cette dernière solution fournit donc une latence plus faible au système global, comme nous aurons l'occasion de la détailler un peu plus loin. Nous avons malgré tout implémenté les deux solutions, permettant à l'utilisateur de choisir la solution qu'il préfère. Nous décrivons, ci-après les détails de l'implémentation du fenêtrage asymétrique ;

3. en ce sens qu'il n'est pas visible de l'extérieur.

celle du fenêtrage symétrique suit globalement le même processus.

Le fenêtrage asymétrique consiste à avoir, pour un GCI donné, à la fois une fenêtre maximale unitaire exactement au GCI et un fenêtrage minimal nul pour les deux GCIs précédent et suivant. Si l'on note $t_a(s) = t(\mathrm{GCI})$ l'instant de fermeture glottique s, on s'aperçoit clairement que, a priori,

$$P(s) = t_a(s-1) - t_a(s) \neq P(s+1) = t_a(s+1) - t_a(s) \tag{2.9}$$

La durée de la fenêtre $h(s)$ est alors égale à $P(s) + P(s+1) = t_a(s+1) - t_a(s-1)$. La durée totale de ce tampon audio interne est donc nécessairement deux fois plus grande que le fichier original. Le calcul de ce tampon audio interne est résumé sur la figure 2.2 suivante [4]. Sur cette figure, est détaillé le fonctionnement pour le fenêtrage du signal original et l'obtention du tampon audio de travail, qui nous servira pour le reste du calcul de modification. Sur la figure du haut, on peut voir une portion voisée du signal de parole originale (i.e. non fenêtrée), en bleu, superposée avec les positions de GCIs correspondant (en rouge) et les fenêtres (en vert).

4. Les valeurs d'amplitude des fenêtres ne sont pas significatives, elles ont été normalisées pour le besoin de visualisation.

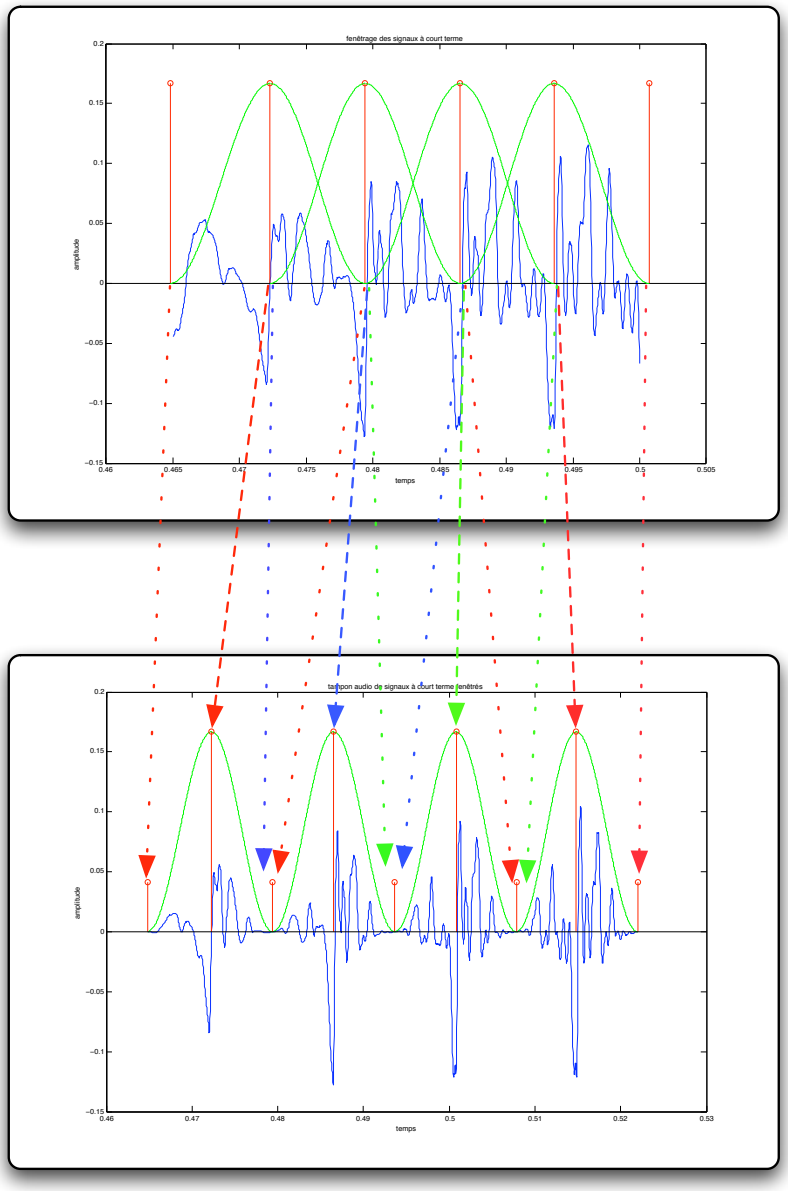


FIGURE 2.2 – *Fenêtrage et constitution du tampon audio*

Les fenêtres utilisées ici sont des fenêtres de Hanning *asymétriques*. C'est-à-dire que pour une fenêtre allant de $t_a(s-1)$ à $t_a(s+1)$, la portion de courbe allant de $t_a(s-1)$ à $t_a(s)$ n'est pas symétrique par rapport à l'axe vertical passant par $t_a(s)$ avec la portion de courbe allant de $t_a(s)$ à $t_a(s+1)$. Toutefois, la continuité est conservée avec des valeurs pour les fenêtres telles que :

$$h(t_a(s-1)) = h(t_a(s+1)) = 0 \quad et \quad h(t_a(s)) = 1 \tag{2.10}$$

Sur un signal de parole commun non pathologique, comme c'est le cas sur notre figure, il n'est pas réellement possible de se rendre compte à l'oeil nu de la variation d'une période à l'autre, étant donné que la gigue fréquentielle (jitter) naturelle est assez faible. Pour s'en convaincre, donc, est reporté sur la figure également les correspondances entre les GCIs originaux et ceux du tampon audio interne.

Avec des traits en tirets sont représentées les correspondances de GCIs pour la période considérée. Autrement dit, le GCI $t_a(s)$ donnera naissance au fenêtrage $h(s)$. Et l'on retrouve alors en bordure de fenêtrage, là où la fenêtre de Hanning est nulle, les positions des GCIs $t_a(s-1)$ et $t_a(s+1)$ respectivement. Quatre fenêtrages consécutifs sont ainsi représentés ici respectivement par des flèches de couleurs rouge, bleu, vert, puis rouge. On voit ainsi que dans le tampon audio interne, le GCI $t_a(s+1)$ d'une fenêtre donnée va correspondre avec le GCI $t_a(s-1)$ suivant (c'est à dire le GCI $t_a(s)$ courant, puisque $s \leftarrow s+1$). Comme les valeurs de fenêtres sont nulles en ces points, cela ne pose pas de problème.

La formule pour le calcul d'une fenêtre de Hanning donnée est la suivante :

$$h(n) = \begin{cases} \frac{1}{2}(1+\cos(2\pi\frac{Pg+n+1}{Pg+1})) & \text{pour} \quad -Pg \leq n \leq 0 \\ \\ \frac{1}{2}(1-\cos(2\pi\frac{Pd+n+1}{Pd+1})) & \text{pour} \quad 1 \leq n \leq Pd \end{cases} \tag{2.11}$$

avec $Pg = t_a(s) - t_a(s-1)$ et $Pd = t_a(s+1) - t_a(s)$.

Concernant le fenêtrage symétrique, le calcul précédent reste identique, à la seule différence que $P_g = P_d$. Deux choix équivalents sont alors possibles : soit la durée de la fenêtre est égale à $2 \times P_g$, soit $2 \times P_d$. Nous n'avons alors plus d'identité entre les valeurs nulles de la fenêtre courante, et les GCIs précédent ou suivant. Dans le premier cas, le GCI précédent correspond à une valeur nulle mais pas le GCI suivant. Dans le second cas, c'est l'inverse. Etant donné, que pour un fichier de parole original donné, on dispose de périodes de silence au début et à la fin du fichier audio, c'est-à-dire des parties nécessairement non voisées, l'une ou l'autre des solutions engendre une latence équivalente.

Parallèlement à ce tampon audio, on conserve dans deux vecteurs les positions des GCIs d'une part et les périodes associées d'autre part. Cette méthode possède l'avantage de ne manipuler uniquement que des indices au moment de la recopie vers le signal de synthèse de sortie, et s'avère ainsi plus simple à manipuler.

2.3.2 Les contraintes temps-réel

Le principal problème rencontré lors de notre implémentation a été de pouvoir disposer d'un outil de modification de la hauteur et de la durée du signal de parole, tout en respectant les contraintes

imposées par une application temps réel.

En effet, de telles manipulations impliquent nécessairement une compression de l'axe temporel, lors d'une accélération ou d'une augmentation de la hauteur du signal. Or, selon le cadre de calcul PSOLA, l'addition-recouvrement des signaux à court terme est réalisé de manière discrète à chaque marque de synthèse selon les valeurs de modification de hauteur et de durée. Et, si l'on considère une forte compression du signal, avec deux signaux à court terme avec des périodes de valeurs suffisamment différentes, on se retrouve à calculer un signal à court terme donné nécessitant une addition dont les échantillons de synthèse sont déjà sortis sur la carte audio. Cette situation est illustrée sur la figure 2.3 ci-dessous.

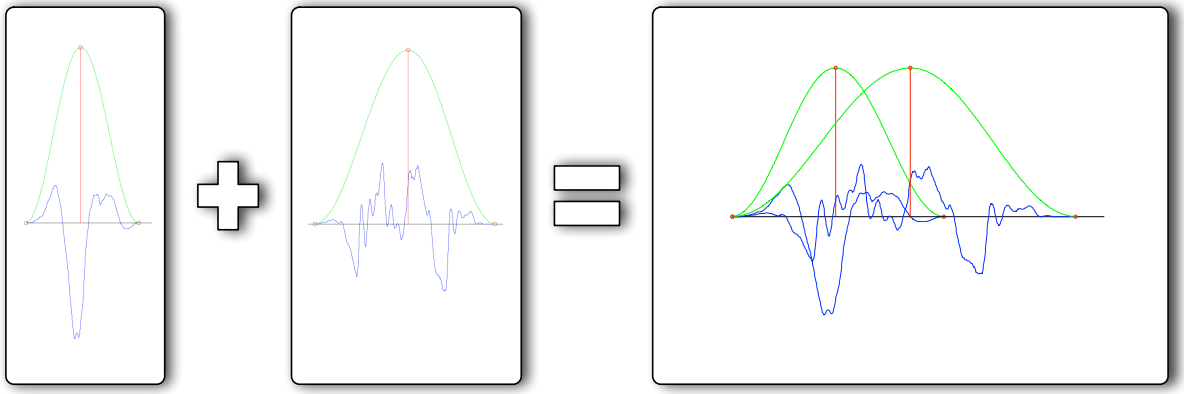


FIGURE 2.3 – *Addition de deux signaux à court terme, lors d'une compression temporelle.*

Ce que l'on observe lors d'une telle compression de l'axe temporel, est que lorsque l'on arrive à la seconde marque de synthèse, on doit ajouter en lieu et place, le signal à court terme représenté ici. Mais, comme chaque fenêtrage court le long deux périodes fondamentales instantanées, on se retrouve à additionner avec le signal à court terme précédent, des échantillons antérieurs à la marque de synthèse précédente. Et, si l'on considère d'autre part que ce traitement est effectué en temps réel, on se rend bien compte que ces échantillons sont déjà "sortis" sur la carte audio de l'ordinateur. Il est donc pratiquement impossible de réaliser un tel traitement avec une latence nulle.

La solution consiste donc à créer un nouveau tampon audio, beaucoup plus petit, contenant la mémoire des quelques derniers signaux de synthèse à court terme. Quelle est alors la taille optimale de ce nouveau tampon pour une réalisation efficace ?

Rappelons nous tout d'abord, que le principe de l'algorithme PSOLA est de dupliquer ou ignorer les signaux d'analyse à court terme. La duplication la plus large que l'on soit théoriquement amené

à effectuer est une répétition du signal à court terme dont la période instantanée est la plus grande [5]. La taille suffisante de ce tampon sera donc de deux fois la taille de la période maximale. Comme on s'est assuré lors de l'étape de pré-traitement de limiter l'espacement entre deux GCIs, la taille de ce tampon ne sera jamais exagérément grande. Sachant que la fréquence minimale d'analyse est de $75\ \mathrm{Hz}$, correspondant à une période instantanée maximale de $13,5\ \mathrm{ms}$ environ, et étant donné que le fenêtrage est effectuée sur deux périodes, notre tampon audio aura une taille maximale de $54\ \mathrm{ms}$. Mais ceci ne constitue pas toutefois directement notre latence.

Le tampon où est effectué l'addition-recouvrement est implémenté sous la forme d'un tampon circulaire, c'est-à-dire que puisque l'on est sûr que notre addition-recouvrement ne dépassera jamais plus de deux fois la période maximale $\mathrm{P_{max}}$, les échantillons plus anciens que $2 \times \mathrm{P_{max}}$ peuvent désormais être écrasés sans souci. Nous dénommerons maintenant ce tampon par tampon circulaire, par opposition au tampon interne de fenêtrage. Le fonctionnement de ce tampon circulaire est détaillé sur la figure 2.4 suivante.

5. il n'en existe qu'un seul !

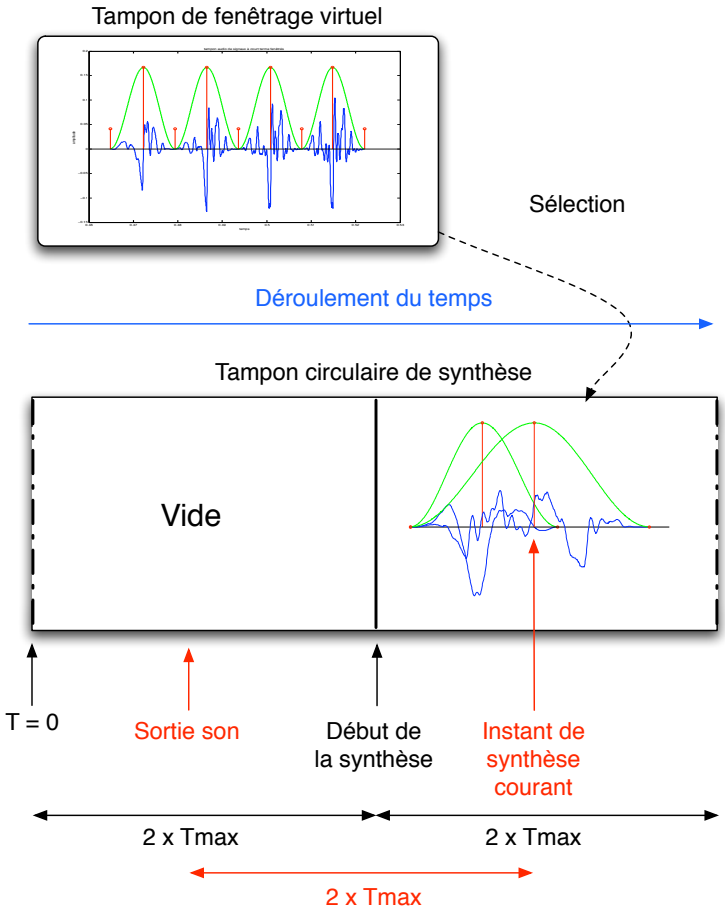


FIGURE 2.4 – *Fonctionnement du tampon circulaire pour la synthèse.*

A l'instant initial $T = 0$, le tampon circulaire est entièrement vide, mais rien ne nous empêche de commencer la synthèse $2 \times P_{max}$ "plus loin". En effet, le problème évoqué ci-dessus sera nécessairement rattrapé au moment où les échantillons sortent sur la carte audio. En d'autre termes, les $2 \times P_{max}$ premiers échantillons seront du silence. Aussi, la latence totale fixe de notre système est de $2 \times P_{max} < 27\ ms$. Cette latence audio est acceptable pour notre application de modification de parole.

Concrètement, d'un point de vue informatique, un pointeur va se déplacer le long de notre tampon circulaire (d'où son nom) pour savoir où effectuer la copie du signal à court terme considéré. Ce pointeur se situera toujours $2 \times P_{max}$ plus tard que l'échantillon de sortie. Une fois que ce pointeur arrive au bout du tampon, il retourne au début du tampon et commence à écraser les anciennes valeurs. Seulement, cette fois, ces nouveaux échantillons vont remplacer des valeurs antérieures à la sortie audio et qui n'ont surtout plus besoin d'être modifiées, puisque l'on s'en est assuré en amont.

L'unique limitation de notre méthode, outre la latence, est que l'on ne pourra pas ralentir notre signal de parole d'un facteur supérieur à 2, ou diviser notre hauteur par plus de 2. Cela étant, il est connu que l'algorithme PSOLA a tendance à introduire des artefacts audibles au-delà de ces valeurs. En outre, dans le cadre de la modification d'un signal de parole, cela nous permet déjà de couvrir une large majorité des situations naturelles rencontrées.

2.3.3 Le calcul des instants de synthèse

Notons bien, avant de décrire le déroulement du calcul des instants de synthèse, que tous les calculs relatifs à la copie des signaux à court terme ne sont réalisés qu'une seule fois par période. Pendant tout le reste du temps, l'algorithme ne se charge uniquement que de sortir sur la carte audionumérique les N échantillons venant d'être calculés, somme pondérées des échantillons de différents signaux à court terme, selon l'approche théorique ayant été expliquée. Cela signifie donc, qu'une fois par période, on récupère les valeurs courantes de α_s et β_s de manière : d'une part, à savoir quelle sera la durée, en nombre d'échantillons, de la prochaine période (en d'autre termes, cela revient à déterminer à quel instant sera situé le prochain instant de synthèse réel) et, d'autre part, cela permet de savoir quels seront les signaux (étant entendu que l'on effectue une pondération entre deux signaux consécutifs) d'analyse à court terme à copier à cet instant précis.

Deux variables sont particulièrement importantes lors du calcul des instants de synthèse, l'index du GCI du signal à court terme d'analyse courant, et celui du signal à court terme d'analyse précédemment copié [6]. Pour le calcul des instants de synthèse, deux cas de figures peuvent se présenter : soit le prochain signal d'analyse à court terme fait partie d'une zone voisée, soit d'une zone non voisée. Dans le premier cas, on doit calculer la prochaine période de synthèse en fonction des valeurs de α_s et β_s, dans le second cas, il suffit juste de *sauter* au prochain GCI, quelque soient les valeurs de modification. Dans ce ce cas cependant, un drapeau est modifié de façon à inverser l'axe temporel, afin d'éviter le bruit tonal. En fait, il s'agit juste lors de la recopie du signal à court terme, de copier les échantillons en partant de la fin de la période.

C'est alors ici qu'intervient la fonction D, d'appariement des instants de synthèse et d'analyse ; Rappelons, que cette fonction d'appariement est cumulative, donc dans le cas où la prochaine période est voisée, la prochaine période de synthèse vaudra $\mathrm{P_s} = \mathrm{P}/\beta_s$, dans le cas où l'on souhaite manipuler la hauteur grâce à un facteur multiplicatif, ou $\mathrm{P_s} = \mathrm{SR}/fréquence$ avec SR taux d'échantillonnage, dans le cas où l'on souhaite modifier la hauteur de façon absolue. Dans les deux cas, il suffit ensuite d'ajouter à la fonction D cette valeur, en tenant compte du facteur de modification temporelle, comme suit : $\mathrm{D} \rightarrow \mathrm{D} + \alpha_s \times \mathrm{P_s}$.

Une fois ce calcul effectué, il ne reste plus qu'à parcourir le tableau des index d'échantillons du

6. et qui ne sont pas forcément consécutifs, du fait de la répétition/suppression

fichier original, pour savoir *où l'on se trouve*, c'est à dire quelle sera le prochain signal à court terme à copier. On peut alors calculer la valeur de α_u, coefficient de pondération des signaux à court terme consécutifs, comme expliqué précédemment.

Le déroulement du cœur de notre algorihme PSOLA en temps réel, peut être résumé, selon le pseudo-code suivant :

---

si *Période non terminée* **alors**
  Sortie d'un échantillon
sinon
  ancien index $\leftarrow$ nouvel index
  si *prochain GCI non voisé* **alors**
    période $\leftarrow$ nouvelle période de synthèse
    $D \leftarrow D+$ période
    inversion $\leftarrow$ - inversion
  sinon
    sr $\leftarrow$ fréquence d'échantillonnage
    période $\leftarrow$ sr/fréquence
    $D \leftarrow D + \alpha_s \times$ période
    inversion $\leftarrow 1$
  tant que $D \leq t_a(s)$ **faire**
    $k \leftarrow k + 1$
  nouvel index $\leftarrow k$
  $\alpha_u \leftarrow (D - t_a s(s))/(t_a(s+1) - t_a(s))$
  ancien $t_s(u) \leftarrow$ nouveau $t_s(u)$
  nouveau $t_s(u) \leftarrow (t_s(u)+$ période $)$ % [taille du tampon]
  copie du signal à court terme à l'instant $t_s(u)$

2.4 Première expérience d'imitation mélodique

La suite de ce chapitre s'attache à décrire les différentes expériences que nous avons menées dans le domaine de la modification de la hauteur et de la durée de parole. La première expérience décrite ci-dessous a constitué le point de départ de nos explorations et de leurs approfondissements. Néanmoins, cette première expérience n'utilisait pas encore le système de modification de hauteur et de durée combinée. Pour cette expérience, nous avons utilisé un autre outil de modification de hauteur, et c'est à l'aune des résultats assez convaincants obtenus, que nos avons ensuite décidé de développer notre propre outil de modification PSOLA en temps réel. Nous avons par la suite utilisé ce dernier outil pour les deux expériences suivantes, qui feront l'objet des deux prochaines sections de ce présent chapitre.

2.4.1 Evaluation d'un système de réitération intonative contrôlé par la main

La synthèse intonative utilisant une interface contrôlée par les gestes manuels représente une nouvelle approche pour la synthèse efficace de prosodie expressive. C'est ce que nous appellerons dans la suite de ce manuscrit *chironomie*, après avoir défini l'origine historique de ce terme. Une expérience de réitération intonative contrôlée par les gestes manuels est décrite. Celle-ci a été menée grâce à un système de modification de l'intonation en temps-réel contrôlé par une tablette graphique et dont le principe général est présenté sur la figure 2.5. Nous appelerons par la suite *Calliphonie* la réalisation logicielle et matérielle du principe de la chironomie.

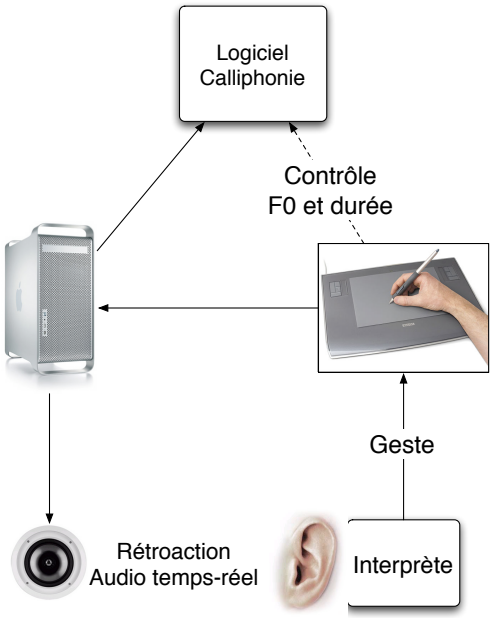


FIGURE 2.5 – *Diagramme générique du système de modification prosodique (Le Beux et al., 2007)*

Dans un ouvrage récent (Tatham and Morton, 2004) (p.167), M. Tatham dresse une liste des lacunes des systèmes de synthèse de parole standards, par rapport aux caractéristiques de la communication humaine.

Il cite ainsi trois problèmes majeurs :

- **La sensibilité perceptive de l'auditeur** : lors d'une conversation, le locuteur est conscient des réactions de l'auditeur à son propos. Les systèmes actuels manqueraient ainsi de *rétroaction* quant à la manière dont le message est interprété.

- **Le contenu émotionnel adéquat** : la majorité de la parole quotidienne concerne nos sentiments et nos croyances, véhiculées par notre ton de voix, ne pouvant difficilement être transmis par le message seul. Or les systèmes de synthèse ne sont pas encore capables de déclencher ou d'induire chez le locuteur une émotion désirée.

- **L'usage précis du langage** : dans quelle mesure est-on capable de prédire l'interaction appropriée entre le choix des mots et de la syntaxe avec l'intonation de la voix synthétique ? [7]

Le système de modification de hauteur et de rythme que nous proposons dans la suite, permet de résoudre (ou tout du moins de pouvoir aborder) au moins deux des trois problèmes rencontrés dans les systèmes de synthèse jusqu'alors. Précisons ici deux caractéristiques importantes de notre système pour la bonne compréhension des implications en termes de modification prosodique. Notre système est un système fonctionnant en temps réel et dont le contrôle est assuré par le biais d'une interface gestuelle contrôlée par un utilisateur.

Prenons les problèmes précités à rebours et reprenons l'exemple de M. Tatham, dans sa version originale. Etant donné que notre système ne réalise aucune inférence sur ce que doit être la prosodie cible selon le vocabulaire et la syntaxe utilisée, il est tout à fait possible de faire prononcer la phrase

7. traduction libre :

   We can see that current speech synthesis systems lack several features that are basic to human communication :
   - Sensitivity to the listener's perception. Speakers are aware of the effect of their utterances on a listener. Current voice output systems lack a means of feedback as to how the message is being interpreted.
   - Appropriate emotive content. Much of daily speech is about our feelings and beliefs, conveying by tone of voice the views that cannot easily be transmitted by the plain message. Employing emotive content associated with the plain message can convey this information, and can trigger an attitude or feeling the synthesis system wishes to encourage the user to have.
   - Precise use of language. Researchers still need to determine the etent to which the choice of words and syntactic constructions interact together with tone of voice in synthetic speech. For example, speakers can say *I am happy* with an intonation that normally triggers sadness. In this case, the emotional tone detecteed may be interpreted as sardonic, or simply generate confusion if inappropriately used.

Je suis content avec une intonation et un rythme qui ne soit pas celui habituellement attendu pour ce type de phrase.

Par ailleurs, le fait que le système soit contrôlé par un utilisateur humain permet, moyennant les artefacts potentiels inhérents à la synthèse, de produire selon le même processus, une expression donnée aussi bien que nous en sommes capables lorsque nous parlons. L'utilisateur peut immédiatement et *rétroactivement* ajuster la modification prosodique de la même manière qu'il le ferait en parlant.

Enfin, la sensibilité à la perception de l'auditeur est peut-être le point le plus discutable, car il n'est sans doute pas encore possible dans la pratique de réaliser une véritable conversation avec un auditeur humain. Néanmoins, il est tout à fait imaginable que cela soit atteignable et la situation dans laquelle deux personnes discutent par téléphone (voire par visioconférence) séparées d'une distance créant une latence non négligeable entre les interlocuteurs peut se révéler un point de départ de comparaison. Précisons, en effet, pour ce point que notre système s'intéresse principalement à la modification de la prosodie, en tant que post-traitement à la synthèse proprement dite (voire à un enregistrement de parole naturelle, suivant l'application souhaitée), et qu'il convient de considérer également, afin d'obtenir un système complet, la génération des phrases "à la volée", qui est possible mais nécessitant de la part de l'utilisateur de saisir les phrases à synthétiser, au fur et à mesure. Pour prendre une analogie musicale, cela consisterait pour un musicien à écrire sa propre partition avant de pouvoir la jouer.

Toutes choses égales par ailleurs, la configuration utilisée par notre système, nécessite l'intervention d'un utilisateur, à la manière d'un instrument musical. Cette intervention humaine est par ailleurs l'un des postulats de base de notre étude, permettant d'introduire cette *humanité* qui fait cruellement défaut dans la plupart des synthétiseurs actuels. L'inconvénient d'une telle méthode, dans la perspective de génération de prosodie, est que cette procédure n'est évidemment pas *automatique* (puisque par essence manuelle). Mais ce n'est pas pour cela qu'elle n'est pas *automatisable*. L'utilisation d'interfaces gestuelles permet en effet de conserver une trace (à une fréquence suffisamment élevée) de la totalité des gestes exécutés par l'utilisateur et donc de pouvoir trivialement, rejouer les modifications produites après enregistrements (à la manière d'un Disklavier (Disklavier, 2009)) ou, de manière plus approfondie, d'analyser les mouvements de l'utilisateur afin d'en déduire une modélisation plus fine des règles prosodiques. La Figure 2.6 détaille le processus d'enregistrements de ces données.

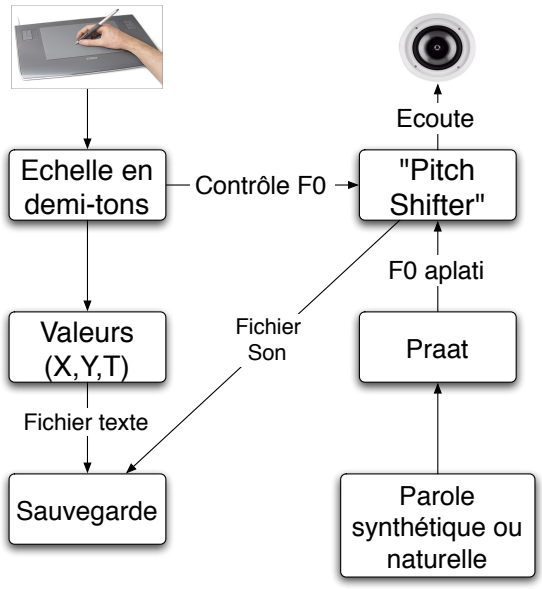


FIGURE 2.6 – *Le système Calliphonie (Le Beux et al., 2007)*

Nous avons choisi d'appeler notre système "Calliphonie", simplement parce que la nouvelle prosodie générée est produite grâce à des gestes manuels, pouvant être comparés à une réalisation calligraphique. Schématiquement, l'utilisateur contrôle la prosodie en temps-réel en dessinant des contours sur une tablette graphique tout en écoutant la parole modifiée.

Dans un premier temps, nous avons utilisé ce système pour une tâche de réitération d'un corpus (i.e. des phrases allant de 1 à 9 syllabes, de parole naturelle et réitérée). Les sujets ont également produit des imitations vocales de ce même corpus. Les corrélations et les distances entre les contours intonatifs naturels et réitérés ont été mesurées. Ces mesures objectives nous ont permis de montrer que la réitération gestuelle et la réitération vocale donnent des résultats comparables, de bonne qualité. Cette étude ouvre ainsi la voie à plusieurs applications de synthèse d'intonation expressive, ainsi que pour la modélisation d'un nouveau paradigme intonatif en termes de mouvements.

Concernant les applications, celles-ci peuvent se traduire selon deux approches :

1. La parole synthétique produite par un système de synthèse à partir du texte peut être efficacement réglée par le geste manuel pour l'obtention d'une parole expressive.
2. Plusieurs styles prosodiques peuvent être appliqués aux phrases de la base de données sans nécessiter d'enregistrer de nouvelles phrases.

Nous discuterons de ces deux possibilités, après avoir présenté la validation de notre cadre de travail.

2.4.2 Préambule

Bien que de nombreux modèles d'intonation aient été proposés pour une quantité de langues, la question de la représentation expressive de l'intonation reste encore ouverte. Les modèles phonologiques de l'intonation se focalisent sur des structures contrastées (souvent tonales) (Pierrehumbert, 1980) : ils ne se préoccupent pas de la description expressive des variations d'intonation. La description phonétique et la stylisation de l'intonation décrivent souvent les patrons mélodiques comme des "mouvements", des "contours" ou des "points cibles". L'approche défendue dans ce manuscrit repose sur l'hypothèse que l'intonation possède de nombreux points communs avec les caractéristiques d'autres types de mouvements ou gestes expressifs humains (comme les mimiques faciales ou les gestes manuels).

Poser la question des représentations intonatives en termes de mouvements, comme, par exemple, les mouvements gestuels, pourrait donner naissance à de nouveaux points de vue dans le domaine de la recherche en intonation. Et l'analogie entre l'intonation et les mouvements manuels semble prometteuse. Une première application est celle de la synthèse expressive directement contrôlée par le geste : cette dernière peut alors être utilisée pour l'enrichissement de corpora de synthèse concaténative de parole, ou pour la génération de stimuli en analyse de parole expressive. Une seconde application plus fondamentale est celle de la modélisation de l'intonation en termes de représentation de mouvements (vitesse du mouvement, direction, envergure). Le principal avantage d'une telle modélisation repose sur le fait que l'intonation et le rythme sont ici traités dans un seul et même cadre de travail, et qu'aucune inférence n'est réalisée, a priori, sur la forme que doivent prendre ceux-ci.

La description de l'intonation expressive en termes de mouvements manuels est connue depuis l'antiquité sous le terme de "chironomie" (Mocquereau, 1927) (chap. 1, p. 103) [8]. Ce terme provient du grec "chiro" (main) et "gnomos" (régle). Le terme apparaît pour la première fois dans le champ de la rhétorique pour décrire les mouvements manuels co-verbaux renforçant l'expression d'un discours (Tarling, 2004). Une autre signification est présente dans la musique médiévale, où la chironomie sert au chef de chœur pour indiquer les tonalités dans le chant grégorien (Mocquereau, 1927) (chap. 2, p. 683).

Concernant la description rythmique, on trouve également, durant l'antiquité romaine, la formalisation d'un système de règles concernant la bonne métrique à adopter pour la déclamation d'un texte (principalement en poésie). On retrouve également ces règles en grec ancien ou en sanskrit. Ce système de règles strictes est appelé *scansion* chez les romains, et concerne le bon rythme à adopter pour la déclamation de vers. On retrouve d'ailleurs le même type de règles (quoique plus souples) dans la poésie classique. Ce système de règles en synthèse de *vocable* a également servi d'inspiration récemment pour des applications musicales (McLean and Wiggins, 2008)

8. Chironomie sur Wikipedia (en anglais)

La musique et la parole sont toutes deux des formes de communication humaine par le biais d'un contrôle expressif de la production acoustique. La musique, contrairement à la parole, a développé l'usage "d'instruments" externes pour la production sonore et son contrôle. La musique instrumentale est produite par des "interfaces" contrôlées par les mains, le souffle, les pieds. Au vu du fort intérêt récent pour les nouvelles interfaces pour l'expression musicale, des ressources comme des langages de programmation sonore temps-réel, des appareils de contrôle, des algorithmes de modification sont désormais couramment disponibles au sein de la communauté d'informatique musicale (Cook, 2005; Kessous, 2004). Suivant cette direction de recherche, un système de chironomie informatique, c'est-à-dire de contrôle mélodique temps-réel par les mouvements manuels, est présenté et évalué ci-dessous. Parmi les appareils disponibles pour le contrôle manuel, ceux d'écriture (tablette graphique) ont été privilégiés. En grande partie parce que l'écriture permet un contrôle très précis et intuitif de l'intonation, mais aussi et nous le verrons par la suite, parce que le type de mouvements dynamiques possibles par l'écriture se rapprochent de ceux de la prosodie (cf. section 2.5.6). En outre, une grande majorité de la population a appris a écrire depuis la plus jeune enfance, ce qui facilite la mise en place des expériences.

Les trois principales problématiques abordées dans l'expérience suivante sont :
- De quelle manière les mouvements d'écriture reproduisent les mouvements intonatifs ?
- Jusqu'à quel point l'écriture et la stylisation de l'intonation vocale sont-elles comparables ?
- Dans les deux cas, quel est le degré de proximité entre les intonations naturelles et les contours stylisés ?

Ces questions sont traitées dans le cadre du paradigme de réitération de l'intonation (Larkey, 1983). La tâche des sujets consistait ici à reproduire les contours intonatifs de phrases cibles par imitation vocale d'une part, et par des mouvements manuels d'autre part. Des stimuli de parole naturelle et réitérée (i.e. de type "mamama", délexicalisée) de plusieurs longueurs étaient proposés. Des mesures de distances entre la phrase originale et la parole réitérée ont été utilisés comme référence de performance.

Nous allons désormais décrire les aspects techniques de notre expérience tels que l'appareillage expérimental, le paradigme de test et les procédures d'analyse. Puis nous présenterons les résultats en termes de performance pour l'imitation intonative. Enfin, nous discuterons des résultats obtenus jusqu'alors afin de pouvoir tirer de nouvelles conclusions et envisager la suite de nos expériences.

2.4.3 Calliphonie : les premiers pas

Modificateur de hauteur gestuel

Le système Calliphonie a été développé afin de contrôler le pitch de la parole grâce à a des mouvements d'écriture. Le système, similaire dans une certaine mesure à celui décrit dans

(D'Alessandro et al., 2006), est basé sur l'environnement de programmation Max/MSP (Max/MSP, 2008), et utilise une version temps-réel de l'algorithme TD-PSOLA. Le système de modification de hauteur que nous avons utilisé est celui développé par Tristan Jehan (Jehan, 2008), car il fournit une qualité audio satisfaisante. C'est ce même programme que nous avions utilisé lors du workshop eNTERFACE à Mons (d'Alessandro et al., 2005a), avec le synthétiseur MaxMBROLA. L'interface développée avait pour but de permettre une étude perceptive, avec des actions simples de la part de l'utilisateur que nous détaillerons dans la suite.

Pour les besoins de notre expérience, deux types de données étaient nécessaires :

1. Une phrase de parole enregistrée dont la fréquence fondamentale a été aplatie à la fréquence moyenne du locuteur considéré (par exemple, à 120 Hz pour une voix d'homme), et
2. Les données de sortie d'un appareil de contrôle gestuel tel qu'une tablette graphique (contrôlées par des mouvements de type écriture) connectée à la valeur de la hauteur de la phrase, résultant en un contrôle direct par le geste de la hauteur en sortie.

Ainsi, le système permet à un utilisateur de contrôler précisément la hauteur d'une phrase précédemment enregistrée, uniquement grâce au stylet d'une tablette graphique. La Figure 2.7 résume le fonctionnement qui vient d'être décrit.

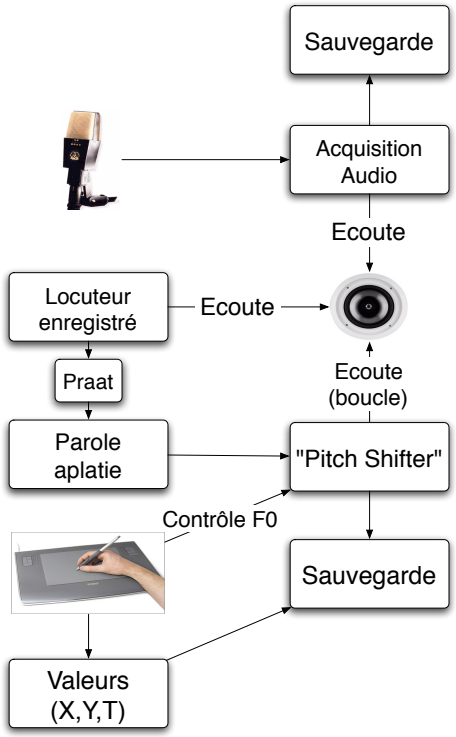


FIGURE 2.7 – *Description technique du système*

Interface d'imitation prosodique

Afin de tester si le contrôle de la prosodie par l'écriture permet de reproduire fidèlement la prosodie naturelle, une interface logicielle informatique *ad-hoc* a été développé (cf. figure 2.8) grâce à la plate-forme de programmation graphique temps-réel Max/MSP (Max/MSP, 2008). Le but est de permettre aux sujets de l'expérience d'imiter la prosodie de la parole naturelle soit vocalement, soit par les mouvements d'écriture. Chaque sujet écoute la phrase naturelle en cliquant sur un bouton avec le pointeur de la souris, et doit ainsi imiter la prosodie qu'il vient d'entendre par deux moyens : vocalement en enregistrant sa propre voix, et en utilisant le contrôleur gestuel de prosodie.

L'interface affiche quelques boutons de contrôle :

1. Pour enregistrer sa voix ou l'imitation avec la tablette graphique,
2. Pour rejouer les imitations déjà enregistrées et
3. Pour sauvegarder celles jugées satisfaisantes.

L'interface affiche également une représentation graphique des paramètres prosodiques du son original, comme illustré sur la figure 2.8.

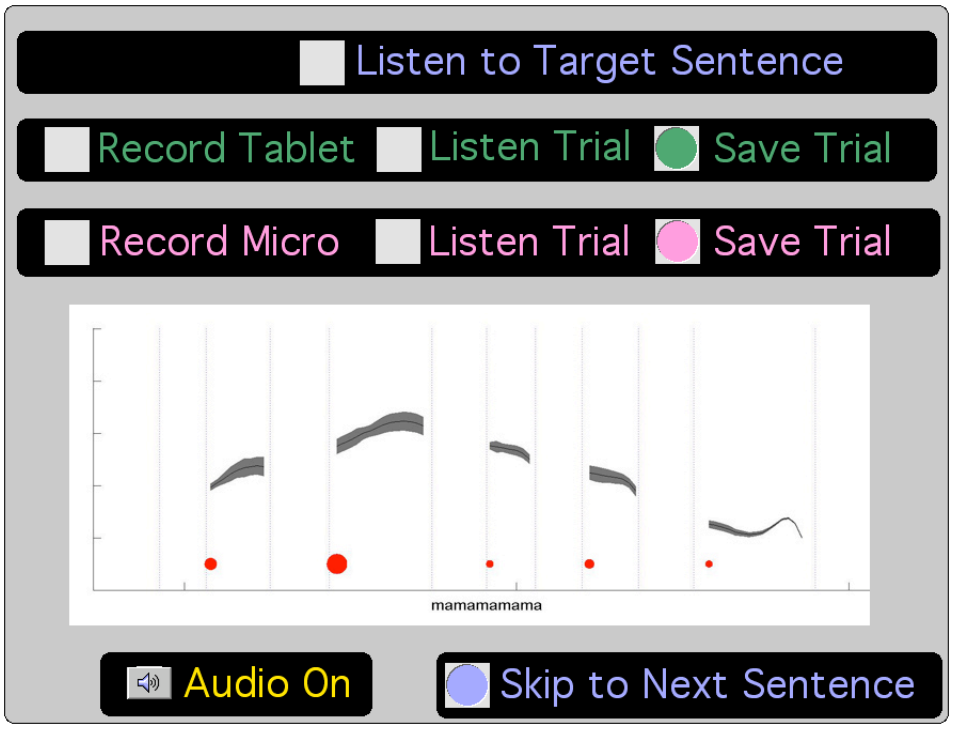


FIGURE 2.8 – *Interface utilisée pour l'expérience. Les boutons permettent d'écouter la phrase originale, d'enregistrer sa voix ou la tablette graphique, d'écouter une performance et de l'enregistrer lorsqu'elle est satisfaisante. L'image représente la prosodie de la phrase courante. (Le Beux et al., 2007)*

Le but de cette expérience étant d'observer jusqu'à quel point il est possible d'imiter la voix naturelle, les sujets avaient la possibilité d'écouter le son original, et pouvaient réaliser des imitations jusqu'à ce qu'elles soient jugées satifsfaisantes. Plusieurs enregistrements pouvaient être réalisés pour chaque son original.

Enfin, les sujets passent à la phrase suivante, grâce au bouton situé en bas à droite de la fenêtre. Cette action a pour effet de charger la nouvelle phrase cible, sa version aplatie et d'afficher le nouveau contour prosodique de la phrase cible en question. Un nouveau fichier texte est créé pour recueillir les données de la tablette, de même un nouveau fichier audio d'enregistrement de l'imitation vocale est créé. Ainsi, chaque phrase est labelisée de façon unique de manière à faciliter la phase d'analyse des données. Comme le test dure couramment plusieurs minutes par phrase, les sujets étaient invités à faire des pauses de temps en temps.

Corpus

Ces expériences sont basées sur un corpus dédié construit autour de 18 phrases, allant de 1 à 9 syllabes (cf . Tableau 2.1). Chaque phrase était enregistrée dans sa version lexicalisée et dans sa version délexicalisée en remplaçant chaque syllabe par une syllabe /ma/, afin d'obtenir une parole réitérée (Larkey, 1983). Lors de l'établissement du corpus, les mots étaient choisis de façon à respecter deux critères (l'utilisation d'une structure syllabique consonne-voyelle CV et l'absence de consonnes plosives au début des mots), dans le but d'obtenir des formes prosodiques facilement comparables parmi les phrases et d'éviter des effets trop importants de microprosodie dûs aux explosions plosives.

Nb. Syllabes	Phrase	Phonétique
1	Non	[nɔ̃]
2	Salut	[saly]
3	Répétons	[ʁepetɔ̃]
4	Marie chantait	[maʁi ʃãtɛ]
5	Marie s'ennuyait	[maʁi sãnɥijɛ]
6	Marie chantait souvent	[maʁi ʃãtɛ suvã]
7	Nous voulons manger le soir	[nu vulɔ̃ mãʒe lə swaʁ]
8	Sophie mangeait des fruits confits	[sofi mãʒɛ de fʁɥi kɔ̃fi]
9	Sophie mangeait du melon confit	[sofi mãʒɛ dy məlɔ̃ kɔ̃fi]

Nb. Syllabes	Phrase	Phonétique
1	L'eau	[lo]
2	J'y vais	[ʒi vɛ]
3	Nous chantons	[nu ʃãtɔ̃]
4	Vous rigolez	[vu ʁigole]
5	Nous voulons manger	[nu vulɔ̃ mãʒe]
6	Nicolas revenait	[nikola ʁəvənɛ]
7	Nicolas revenait souvent	[nikola ʁəvnɛ suvã]
8	Nicolas lisait le journal	[nikola lizɛ lə ʒuʁnal]
9	Nous regardons un joli tableau	[nu ʁəgaʁdɔ̃ ɛ̃ ʒoli tablo]

TABLE 2.1 – *Les 18 phrases du corpus, de 1 à 9 syllabes, avec leurs transcriptions phonétiques (d'Alessandro et al., 2007)*

Deux locuteurs (une femme et un homme, de langue maternelle française) ont enregistré le corpus. Ils devaient produire toutes les phrases, présentées dans un ordre aléatoire, chacune selon les trois différentes consignes suivantes :

1. En utilisant une intonation déclarative
2. En créant une emphase sur un mot spécifique au sein de la phrase (généralement le verbe)
3. En utilisant une intonation interrogative.

Lors de la phase d'enregistrement, les locuteurs devaient lire la phrase puis la produire en utilisant le style intonatif adéquat. Une fois la phrase enregistrée dans sa version lexicalisée, ils devaient la reproduire avec la même prosodie, mais dans sa version réitérée. Les locuteurs pouvaient faire autant d'essais que nécessaire pour obtenir une paire satisfaisante de phrases (lexicalisée et réitérée).

108 phrases ont ainsi été enregistrées et directement numérisées par un ordinateur (44.1 kHz, 16 bits) pour chaque locuteur, en utilisant un amplificateur USBPre®connecté à un microphone omnidirectionnel AKG™C414B situé à 40 cm de la bouche du locuteur, et en utilisant un filtrage passe-haut des fréquences au-dessus de 40 Hz et une réduction du bruit de 6 dB.

Les sujets

Pour cette expérience, qui nous a également servi de pré-test à l'expérience décrite dans la prochaine section 2.5, seuls 4 sujets ont complété la tâche requise sur un total de 9 phrases allant de 1 à 9 syllabes, à la fois lexicalisées et réitérées, et selon les trois conditions prosodiques (déclaratif, emphase, interrogation), pour le locuteur masculin uniquement. Tous les sujets étaient impliqués dans ce travail et parfaitement au courant de ses objectifs et en outre familier avec la recherche en prosodie. Trois des quatre sujets ont suivi un apprentissage musical. L'un des quatre sujets était le locuteur masculin du corpus original, ce dernier ayant dû ainsi imiter sa propre voix ; vocalement et par les mouvements d'écriture.

Mesures de contours prosodiques

Toutes les phrases de ce corpus ont été analysées manuellement afin d'extraire les paramètres prosodiques, à savoir la fréquence fondamentale (en demi-tons), la durée syllabique et l'intensité grâce à Matlab®(script yin (de Cheveigne and Kawahara, 2002)) et Praat (Boersma and Weenink, 2008).

Pour toutes les phrases, des courbes intonatives étaient affichées afin de décrire la prosodie du son original et ainsi faciliter la tâche des sujets lors de l'expérience. Un tel contour intonatif est illustré sur la figure 2.9. Ces courbes représentent le F_0 d'analyse lissé des segments vocaliques, dont la largeur de trait représente la force de voisement. La force de voisement était calculée à partir de l'intensité (en dB) du signal au point de l'analyse de F_0. Les lieux des Centres Perceptifs (p-centers (Scott, 1998)) sont représentés par des cercles rouges, dont le diamètre est relié à l'intensité moyenne du segment vocalique. Exprimé simplement, un centre perceptif traduit le moment d'apparition perceptive d'une syllabe. Les traits pointillés verticaux sur la figure traduisent les frontières de phonèmes.

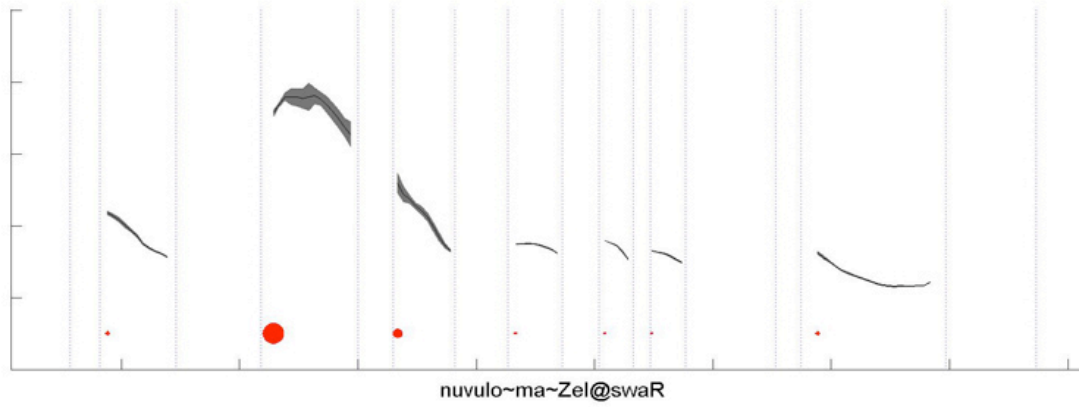


FIGURE 2.9 – *Paramètres prosodiques d'une phrase à 7 syllabes ("Nous voulons manger le soir", focalisée) issue de notre corpus. (d'Alessandro et al., 2007)*

Distances prosodiques et corrélation

Afin d'évaluer les performances d'imitation (vocale ou gestuelle), deux mesures physiques de distance entre la fréquence fondamentale extraite de l'imitation et celle de la phrase originale ont été appliquées sur la base des mesures de dissimilarité physique introduites par Hermès (Hermes, 1998), à savoir :

1. la corrélation entre les deux courbes de F_0, et
2. les différences au sens des moindres carrés entre les deux courbes.

Comme noté précédemment par Hermès (Hermes, 1998), la corrélation est une mesure de similarité entre deux courbes de F_0 tandis que la différence au sens des moindres carrés ou distance RMS [9] est une mesure de dissimilarité, mais chacune d'entre elles fournit une idée de la similarité entre les deux courbes de F_0 comparées. Cependant, la corrélation teste la similitude entre les formes de deux courbes, sans prendre en compte leurs distances moyennes : par exemple, il est possible de reproduire une courbe de F_0 une octave plus basse que l'original, mais si la forme reste la même alors la mesure de corrélation sera très élevée. Au contraire, la distance RMS donnera une idée de l'aire présente entre les deux courbes, et reste alors sensible aux différences entre les deux valeurs moyennes de F_0.

Selon un procédé similaire à celui décrit dans (Hermes, 1998), les deux distances prosodiques sont appliquées avec un facteur pondéré de façon à fournir plus d'importance aux phonèmes possédant une intensité sonore plus forte. Le facteur de pondération utilisé est donc l'intensité, en tant que mesure locale de la force de voisement. Ces deux mesures de dissimilarité étaient calculées automatiquement pour toutes les imitations gestuelles enregistrées par les quatre sujets, c'est-à-dire pour chacune des 54 phrases. Ensuite, seule l'imitation gestuelle la plus proche (d'après, premièrement, la corrélation pondérée, puis la différence RMS pondérée) était conservée pour les résultats d'analyse.

Cette partie du travail a pu être complètement automatisée, car il n'y avait pas de changement dans la durée de la sortie du contrôleur gestuel (seul F_0 est modifié). Ceci n'était cependant pas le cas pour les imitations vocales, qui, elles, devaient être marquées manuellement afin de calculer ces distances. De fait, seules les distances entre les phrases originales et les imitations gestuelles ont été calculées lors de cette première validation.

Les courbes avec les F_0 brutes, à la fois des stimuli originaux et des imitations vocales, ont été produites pour permettre de comparer au moins visuellement les performances gestuelles en regard des imitations vocales. Les courbes avec les F_0 stylisées des phrases originales (i.e. les F_0 lissées pour les segments vocaliques) superposées avec les tracés du stylet sur la tablette ont également été produits afin de comparer les deux modalités d'imitations (cf. figures 2.10 & 2.11)

9. Root-Mean Square

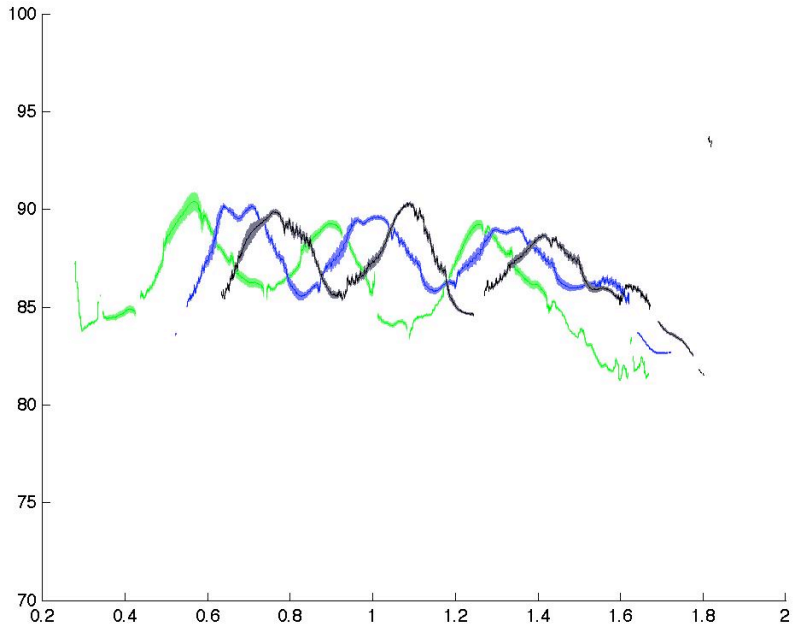


FIGURE 2.10 – *Valeurs de F_0 brutes (en demi-tons) issue d'une phrase originale (en gris) et deux imitations vocales d'un sujet (Les stimuli ne sont pas alignés temporellement, abscisses en secondes). (d'Alessandro et al., 2007)*

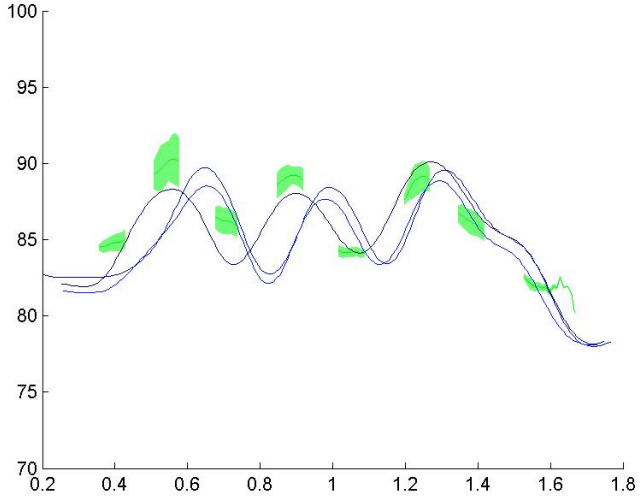


FIGURE 2.11 – *F_0 stylisé d'un phrase originale (la même que sur la Figure 2.10 – courbe verte, avec des valeurs lissées pour le segment vocalique exprimé en demi-tons, abscisses en secondes) et la valeur du paramètre de hauteur contrôlée par la tablette graphique pour toutes les imitations effectuées par un sujet. Les stimuli sont alignées temporellement. (d'Alessandro et al., 2007)*

2.4.4 Résultats de l'expérience d'imitation

Distances prosodiques et corrélation

Les distances physiques (RMS et corrélation R) entre les stimuli originaux et les phrases produites par l'imitation gestuelle sont résumées dans le tableau 2.2. En analysant les résultats de l'expérience, l'influence relative de chacun des paramètres de contrôle est détaillée ci-après.

Sujet	R	RMS
CDA	0.866	3.108
BD	0.9	3.079
SLB	0.901	3.091
AR	0.898	4.728
Total	0.891	3.502

TABLE 2.2 – *Distances moyennes pour chaque sujet et pour les 54 phrases imitées par les mouvements manuels. (d'Alessandro et al., 2007)*

Effet des sujets : Il n'y a pas de différence significative entre les résultats obtenus par tous les sujets : toutes les corrélations sont comparables et autour de 0.9, montrant que les sujets sont capables de percevoir et de reproduire la forme intonative grâce aux mouvements scriptifs. La seule différence notable est celle de la distance RMS obtenue par le sujet AR (4.7) comparativement aux résultats obtenus par les autres sujets (autour de 3.1). Cette différence se traduit par une courbe de F_0 plus proche de l'original pour les trois sujets autres que AR. Cette différence pourrait être expliquée par le fait que le sujet AR est le seul sujet sans éducation musicale, et ainsi qu'il ne soit pas enclin à reproduire une mélodie donnée aussi bien que les autres sujets. Cependant, au vu des corrélations assez proches, cela n'implique pas une difficulté à reproduire des variations de hauteur, mais seulement la hauteur absolue (hauteur tonale).

Effet de la longueur de la phrase : La longueur de la phrase induit une différence plus notable sur les distances. Comme montré sur la figure 2.12, les mesures de dissimilarité augmentent avec la longueur de la phrase : la corrélation diminue continuement quand la longueur de la phrase augmente, et hormis quelques accidents pour les phrases de 3 et 7 syllabes, la différence RMS augmente avec la longueur de la phrase. Ces deux accidents peuvent être expliqués par les hautes valeurs RMS obtenues par deux des sujets pour ces stimuli, et également par le fait que cette mesure est particulièrement sensible à de faibles différences entre les courbes. L'effet de la longueur de la phrase pourrait être un artefact, car le calcul de corrélation ne prend pas en compte une quelconque pondération pour compenser la longueur. Des analyses plus poussées seraient éventuellement nécessaires pour pouvoir conclure sur un effet de la longueur de phrase.

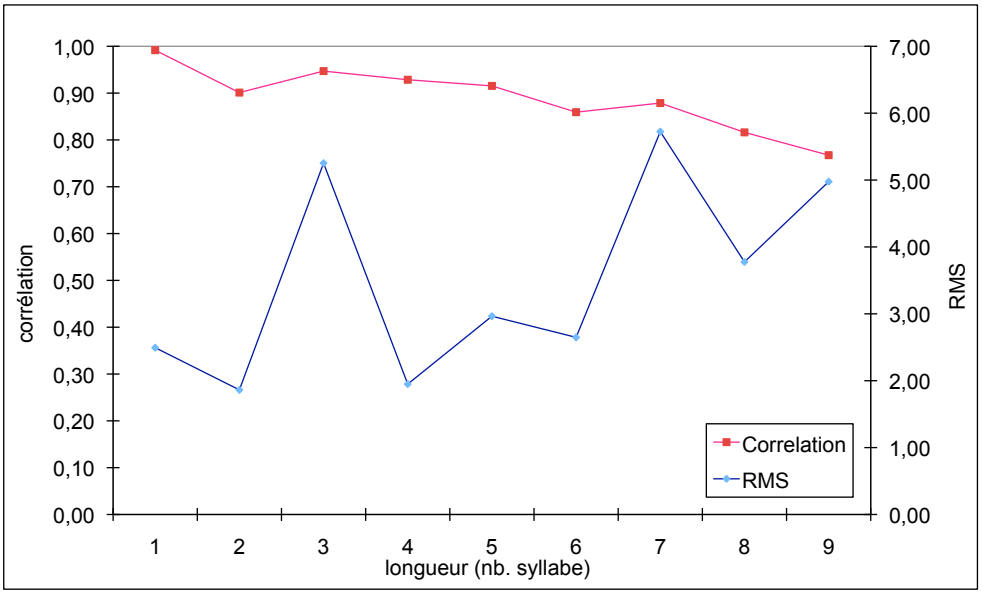


FIGURE 2.12 – *Evolution des deux mesures de distance selon la longueur de la phrase. En abscisses : la longueur du stimulus en nombre de syllabes. En ordonnées, à droite : corrélation (ligne rouge), à gauche : différence* RMS *(ligne bleue). (d'Alessandro et al., 2007)*

Effet du style prosodique et du type de stimulus : Les modalités de phrases déclaratives, avec emphase ou interrogatives donnent des résultats similaires selon la mesure de corrélation, mais la distance RMS est plus faible pour les courbes déclaratives (2.0) que pour les courbes avec emphase ou interrogatives (resp. 4.2 et 4.4). Ceci peut être relié au résultat précédent : les sujets sont capables de reproduire la forme générale de tous les contours intonatifs, mais la perception précise de la hauteur est plus difficile lorsque la courbe présente un glissando (par exemple, lors d'une emphase ou d'une interrogation) que pour une courbe relativement plate, comme dans le cas de l'intonation déclarative.

2.4.5 Discussion et Conclusions Partielles

Niveau de performance et faisabilité

De façon globale, de bons niveaux de performances sont atteints en terme de corrélation et de distance entre l'original et les contours intonatifs réitérés. Bien sûr, il doit être noté que la meilleure occurence réitérée a été sélectionnée pour chaque phrase. Malgré tout, la charge d'entraînement de chaque sujet n'était pas très lourde, chaque sujet ayant pu réaliser la tâche requise en moins d'une après-midi pour la totalité du test. La tâche n'a pas parue particulièrement difficile aux différents sujets, tout du moins comparativement à d'autres tâches de reconnaissance d'intonation, telles que les dictées musicales par exemple.

Modalités vocale et gestuelle

Un résultat à la fois remarquable et étonnant est le fait que les niveaux de performances obtenus par l'écriture manuelle et les intonations réitérées vocales sont tout à fait comparables. Ceci permet de suggérer que l'intonation, à la fois sur les aspects perceptifs et de production motrice, est soit traitée à un niveau cognitif relativement abstrait, ou que le système biophysique mis en jeu pour sa réalisation est comparable au geste manuel du point de vue dynamique, puisqu'elle est plutôt indépendante de la modalité utilisée. Ce phénomène avait déjà été inféré par les orateurs antiques, dans leur description de l'effet expressif des gestes co-verbaux (i.e. de la parole multimodale expressive) rapportée dans les premiers traités de rhétorique romains (Tarling, 2004). Ainsi, il paraît raisonnable d'émettre l'hypothèse que le contrôle de l'intonation peut être accompli par d'autres gestes que ceux de l'appareil vocal avec une précision comparable.

Gestes et intonation

Il semble que les variations micro-prosodiques aient été négligés pour pratiquement toutes les phrases. L'écriture est généralement plus lente que la parole, et donc les gestes manuels ne permettent pas de suivre les détails fins de l'intonation jusqu'au niveau de la micro-prosodie (Fels and Hinton, 1998). En outre, les résultats pour la parole délexicalisée et la parole naturelle sont comparables, bien que la micro-prosodie soit quasiment absente dans la parole délexicalisée. Les gestes manuels correspondent alors plutôt aux mouvements d'intonation prosodique. Les gestes spécifiques utilisés par les différents sujets pour accomplir la tâche d'imitation gestuelle n'ont pas été analysés en détails au cours de cette expérience. Des observations informelles révèlent que certains sujets ont plutôt utilisé des mouvements circulaires, d'autres des mouvements plutôt linéaires. Des analyses de formes plus approfondies seraient nécessaires pour statuer sur l'efficacité des différentes stratégies adoptées.

Conclusion et perspectives

Cette expérience présente une première évaluation de la chironomie par ordinateur, c'est-à-dire du contrôle de l'intonation dirigée par la main. Les résultats montrent que les réitérations intonatives vocales et les réitérations intonatives chironomiques produisent des contours d'intonation comparables en termes de corrélation et de distance RMS. Les applications et les implications de ces découvertes sont nombreuses. Le contrôle chironomique peut ainsi être appliqué à la synthèse vocale expressive de manière performante. Il peut également être appliqué à l'analyse de parole expressive, étant donné que les contours expressifs peuvent être produits et représentés par des tracés manuels.

2.5 Deuxième expérience d'imitation mélodique

Les objectifs principaux de cette expérience était d'une part de permettre la validation de notre outil *Calliphonie* sur un plus grand nombre de sujets, et d'autre part de mesurer la corrélation éventuelle des performances d'imitation gestuelle avec la pratique instrumentale musicale. En outre, contrairement à la première expérience d'imitation, cette dernière était réalisée avec notre nouvel algorithme de modification prosodique PSOLA temps-réel. A la suite de la première expérience, pour laquelle le facteur lexicalisé/délexicalisé ne s'est pas révélé significatif, nous avons décidé d'étudier des phrases lexicalisées, selon deux modalités différentes : des phrases interrogatives et des phrases focalisées sur un mot de la phrase. Un dernier facteur pris en ligne de compte était celui du genre, tant au niveau des stimuli présentés que des sujets sélectionnés pour la réalisation de l'expérience.

2.5.1 Le corpus

Contrairement à l'expérience préliminaire présentée dans la section 2.4.1, le corpus était ici composé de phrases uniquement lexicalisées. En effet, la première étude nous a révélé que les performances d'imitation ne sont pas liées, de quelque manière que ce soit, avec le fait que la phrase soit lexicalisée ou réitérée. Par ailleurs, afin de pouvoir statuer sur la possibilité de reproduction d'expressions plus larges que la simple déclaration, nous avons choisi de concentrer notre corpus sur des phrases interrogatives et focalisées sur un mot de la phrase. Ceci nous permet d'étudier l'effet de la difficulté relative à l'éventuelle complexité mélodique, a priori plus faible pour les phrases focalisées. La réduction substantielle du nombre de stimuli liée à la lexicalisation nous a également permis d'inclure des phrases produites par homme et par une femme. Toutefois, afin de réduire la lourdeur de la tâche demandée aux sujets, nous avons réduit le corpus à des phrases de 2 à 8 syllabes.

Au final, le corpus était constitué de 7 phrases interrogatives, avec leurs homologues focalisées, prononcées à la fois par un homme (le même locuteur que pour l'expérience préliminaire) et une femme, pour un total de 28 phrases.

2.5.2 Les sujets

Afin d'augmenter la significativité statistique du panel, nous avons choisi 10 sujets, 5 femmes et 5 hommes, parmi le personnel de notre laboratoire, dont aucun n'était impliqué dans l'étude en cours, mais dont une minorité pouvait avoir des connaissances relatives à la prosodie de la parole. Contrairement à l'expérience précédente, aucun sujet n'était en tout cas partie prenante dans l'étude en cours.

Grâce à un questionnaire réalisé en début d'expérience, nous demandions au sujet s'il avait suivi une quelconque formation musicale et instrumentale, et si oui, le nombre d'années pratiquées.

2.5.3 L'interface

Dans le but de simplifier l'interface et de guider le déroulement de l'expérience, l'interface utilisée pour l'expérience préliminaire a été quelque peu modifiée.

En premier lieu, afin de faciliter la tâche, les répétitions aplaties des stimuli étaient espacées temporellement de $500\ \mathrm{ms}$, permettant ainsi aux sujets de disposer du temps nécessaire pour repositionner le stylet pour la répétition suivante.

D'autre part, les boutons et les déclencheurs avec lesquels les sujets pouvaient interagir apparaissaient au fur et à mesure, afin d'éviter que les sujets sautent ou oublient de réaliser une étape. Au départ donc, n'apparaissait uniquement que le bouton permettant d'écouter la phrase originale à imiter. Une fois écouté, le sujet pouvait soit réécouter le stimuli ou commencer directement à réaliser des essais avec la tablette graphique. L'allure générale de cette interface était semblable à celle de la première expérience, présentée sur la figure 2.8.

Une fois ses essais à la tablette terminés, le sujet pouvait alors réécouter, soit la phrase originale, soit les imitations qu'il venait de réaliser. Ces différentes étapes pouvaient être réalisées à l'infini, ou plus raisonnablement jusqu'à ce que le sujet soit suffisamment satisfait de ses imitations, ou lorsqu'il jugeait qu'il ne pourrait pas faire mieux. C'est d'ailleurs cette dernière consigne qui était donné aux sujets au début de l'expérience, sans contraindre les sujets sur une quelconque durée maximale de réalisation.

Une fois l'imitation à la tablette terminée, le même processus (i.e. écoute, imitation, réécoute) était effectué pour l'imitation vocale, grâce au micro prévu à cet effet. Les sujets étaient conviés à ne pas faire de pause entre les imitations gestuelles et vocales d'un même stimulus.

2.5.4 Le protocole

L'expérience était composée de deux parties distinctes, une première phase d'entraînement permettant au sujet de se familisariser avec l'outil et la seconde partie, constituant l'expérience à proprement parler, et pour laquelle les données ont été analysées. Les données de la phase d'entraînement n'ont ici pas été prises en compte lors de l'analyse.

La phase d'entraînement

La phase d'entraînement était constituée de 6 phrases déclaratives choisies au hasard parmi les phrases utilisées pour la première expérience, dont la moitié était prononcée par une femme et l'autre par un homme. Le choix d'utiliser des phrases déclaratives était motivé par le souci de ne pas induire de stratégie a priori chez les sujets pour la suite de l'expérience.

L'essentielle particularité de cette phase était que les trois premières phrases était accompagnées des images respectives issues de l'analyse de contours de F_0, de la même manière que lors de l'expérience précédente, tandis que les trois dernières phrases ne disposaient pas d'images, pareillement à la phase d'enregistrement. Nous reviendrons sur ce point dans le paragraphe suivant.

Enfin, cette phase permettait principalement au sujet de se familiariser avec l'interface, sans la pression de devoir faire le mieux possible. Il pouvait alors à loisir essayer, autant qu'il le souhaitait, de réaliser des essais de modification de la hauteur de la phrase aplatie. D'autre part, aucun sujet n'avait eu d'entraînement préalable avec l'interface.

La phase d'enregistrement

Comme nous venons de le mentionner à l'instant, et contrairement au protocole choisi pour l'expérience préliminaire, aucune image du contour mélodique analysé n'était présenté. Ce choix a été motivé par l'intention de pouvoir tester l'hypothèse d'imitation mélodique, sans que l'on puisse reprocher aux sujets d'avoir essayé de reproduire le contour qui leur était présenté. Le sujet ne pouvait ici, afin de réaliser la tâche d'imitation, que se fier à sa propre audition.

Pour la phase d'enregistrement, les stimuli étaient présentés de façon aléatoire, différente pour chaque sujet, à la condition près que les stimuli "homme" et "femme" n'étaient pas mélangés entre eux. On présentait ainsi aux sujets, 14 stimuli aléatoires prononcés par un homme (ou une femme), suivis ensuite de 14 stimuli prononcés par une femme (resp. un homme).

2.5.5 Les résultats

Résultats globaux

Comme on peut le voir sur la figure 2.13, les performances moyennes des sujets sont légèrement plus faibles que lors de la première expérience. Toutefois, les sujets présentant les scores les plus faibles, se retrouvent avec un score de corrélation autour de 0.7, ce qui reste suffisamment significatif pour statuer que l'imitation gestuelle est suffisamment bien réussie dans l'ensemble. Notons d'ailleurs qu'un des sujets (NS) obtient un meilleur score pour l'imitation gestuelle que pour l'imitation vocale. En outre, les scores des meilleurs sujets se situent autour de 0.9, en concordance avec les résultats de la première expérience.

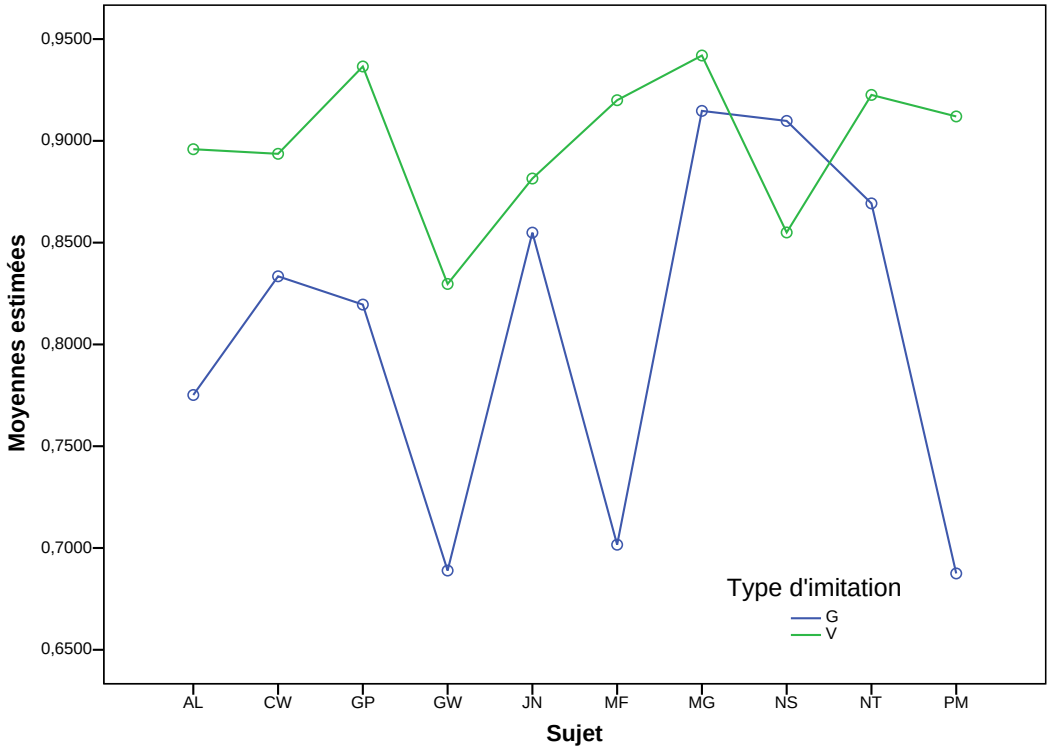


FIGURE 2.13 – *Résultats de corrélation, pour chaque sujet (en abscisses), pour les imitations vocales (en vert) et gestuelles (en bleu)*

Effet du genre

Il n'a pas été noté d'effet du genre sur les performances, que ce soit du côté des sujets ou des stimuli. On pouvait, en effet, s'attendre éventuellement à des performances plus faibles pour le locuteur féminin, à cause des artefacts plus grands introduits par la méthode PSOLA pour une fréquence fondamentale plus élevée, comme mentionné dans la section 2.2. Après l'expérience, plusieurs sujets ont d'ailleurs fait part de leur plus grande difficulté à reproduire l'intonation du locuteur féminin, du fait même de ces artefacts.

Les résultats de corrélation sont en outre légèrement supérieurs pour le locuteur féminin que masculin, quelle que soit la modalité utilisée (interrogatif ou focalisé). On peut en déduire que les artefacts engendrés par l'algorithme PSOLA sur la qualité sonore n'ont pas influencé outre mesure la perception de la mélodie par les sujets. Ces résultats de corrélation sont présentés sur la figure 2.14 suivante.

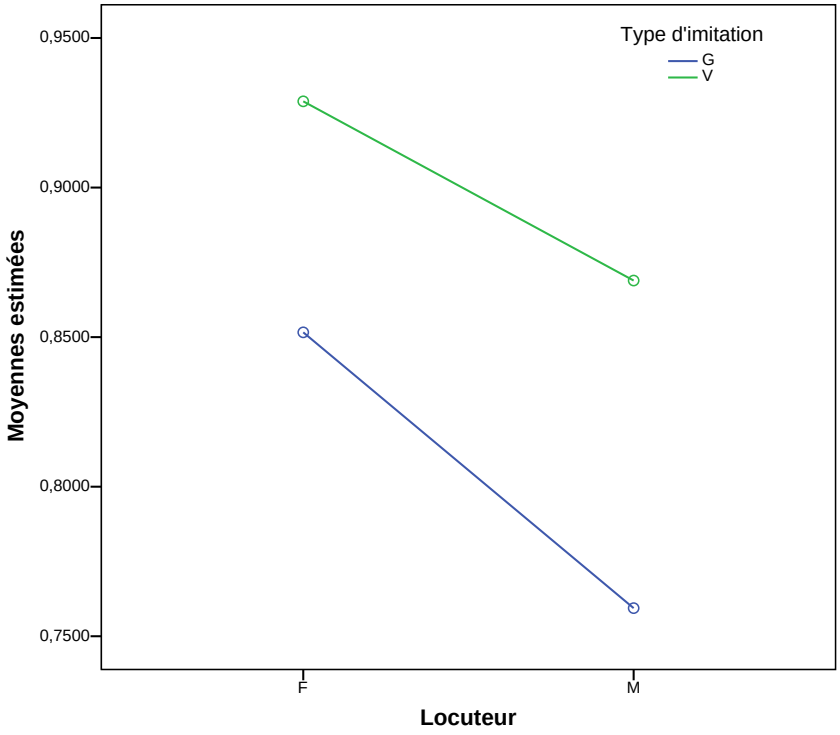


FIGURE 2.14 – *Résultats de corrélation suivant le genre du locuteur (Féminin, Masculin) pour les imitations gestuelles (en bleu) et vocales (en vert).*

Concernant les imitations vocales, lorsqu'un sujet devait reproduire une phrase prononcée par un locuteur du sexe opposé, on observe bien que le sujet essaie de reproduire la courbe intonative et non pas sa valeur absolue, conformément à la consigne demandée. On peut se rendre compte de cet effet sur la figure 2.15. Sur cet exemple, il s'agit de la superposition à la fois du contour intonatif de la phrase originale prononcée par un homme (en gris), du contour intonatif de la reproduction vocale d'un sujet féminin (en vert) et des valeurs du contrôle de la hauteur grâce à la tablette (en rouge).

Mis à part, un décalage en hauteur inévitable, on observe que la microprosodie n'est ici pas reproduite. En ce qui concerne la reproduction avec la tablette, il est assez évident que la dynamique relative aux mouvements microprosodiques est trop élevée pour pouvoir être correctement effectué par des gestes manuels (dont la fréquence maximale se situe entre 30 et 50 Hz). Mais concernant, la reproduction vocale, la plupart de ces mouvements microprosodiques sont également absents, ce qui nous laisse supposer que : soit la hauteur perçue s'abstrait de la microprosodie pour pouvoir être reproduite, soit l'analyse fréquentielle réalisée sur le signal de parole créé des artefacts de mesure sans réel lien avec la nature perceptive de la hauteur. Dans tout les cas, la courbe continue générée par la tablette graphique nous laisse penser qu'elle représente une bonne stylisation de la courbe intonative de parole.

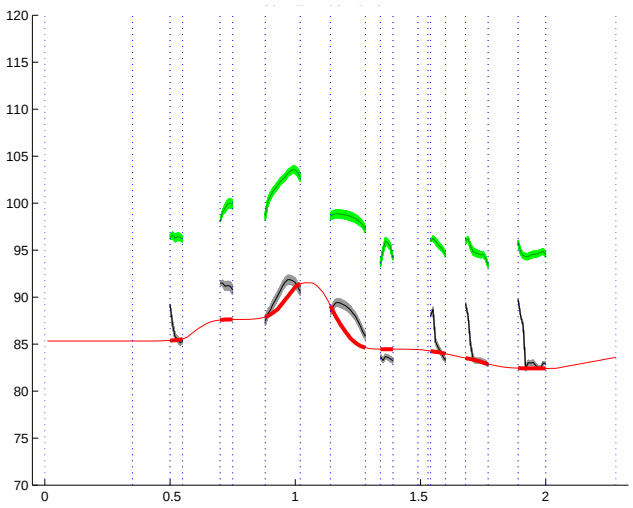


FIGURE 2.15 – *Exemple d'imitation gestuelle (rouge) et vocale (vert), pour une voix cible masculine (gris), par un sujet féminin. Fréquence en demi-tons, temps en secondes.*

Effet de la pratique musicale

Un résultat assez surprenant pour être noté est également ressorti de cette expérience, à savoir que la pratique musicale est fortement corrélée avec les performances des sujets. Si l'on observe les figures 2.16 et 2.17, classées en abscisse par expérience musicale croissante, on s'aperçoit que :

1. Les mesures au sens des moindres carrés diminuent avec la pratique musicale, à la fois pour les imitations vocales et avec la tablette.

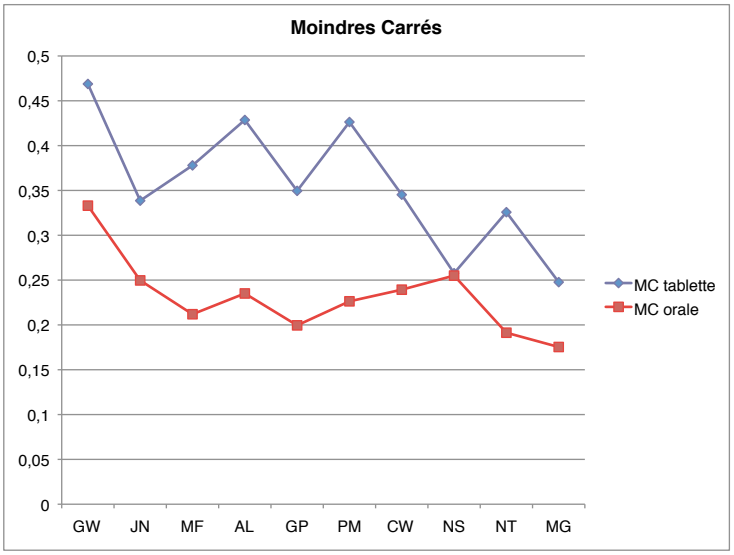


FIGURE 2.16 – *Effet de l'apprentissage musical sur les distances* RMS. *Imitations orales en rouge et avec la tablette en bleu*

2. Tandis que les mesures de corrélations pour les imitations vocales restent sensiblement autour de 0.9 pour tous les sujets, cette corrélation augmente avec la pratique instrumentale.

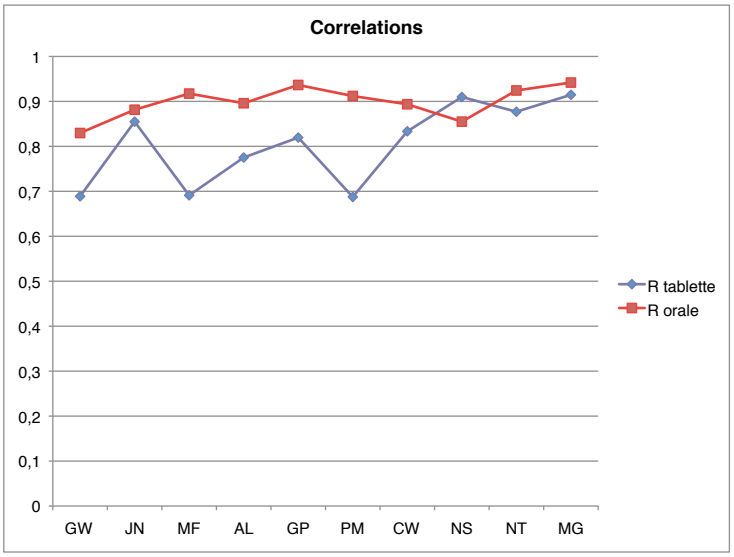


FIGURE 2.17 – *Effet de l'apprentissage musical sur les corrélations. Imitations orales en rouge et avec la tablette en bleu*

On en déduit alors principalement que les performances ne sont pas simplement liées au fait que le sujet possède une "bonne oreille" ou pas, mais plus significativement qu'il avait eu un entraînement instrumental (et donc gestuel) au préalable. Et l'on peut ainsi espérer que les sujets les moins bons en terme de performances pour la tâche considérée pourraient s'améliorer en disposant d'une durée d'entraînement plus longue. Ce phénomène explique également en partie les meilleures performances globales des sujets de la première expérience, qui avait pu disposer de plus de temps pour se familiariser avec l'interface, et qui en outre avaient eu un entraînement instrumental pour trois d'entre eux.

Effet du type de stimuli

Lors de l'expérience précédente, nous avions pu observer un effet de la longueur de la phrase sur les performances (cf. figure 2.12). De prime abord, les résultats de cette nouvelle expérience paraissent en contradiction avec les résultats observés lors de la première expérience. La figure 2.18 suivante, représente les résultats de corrélation, suivant le type d'imitation en fonction de la longueur du stimulus, en nombre de syllabes.

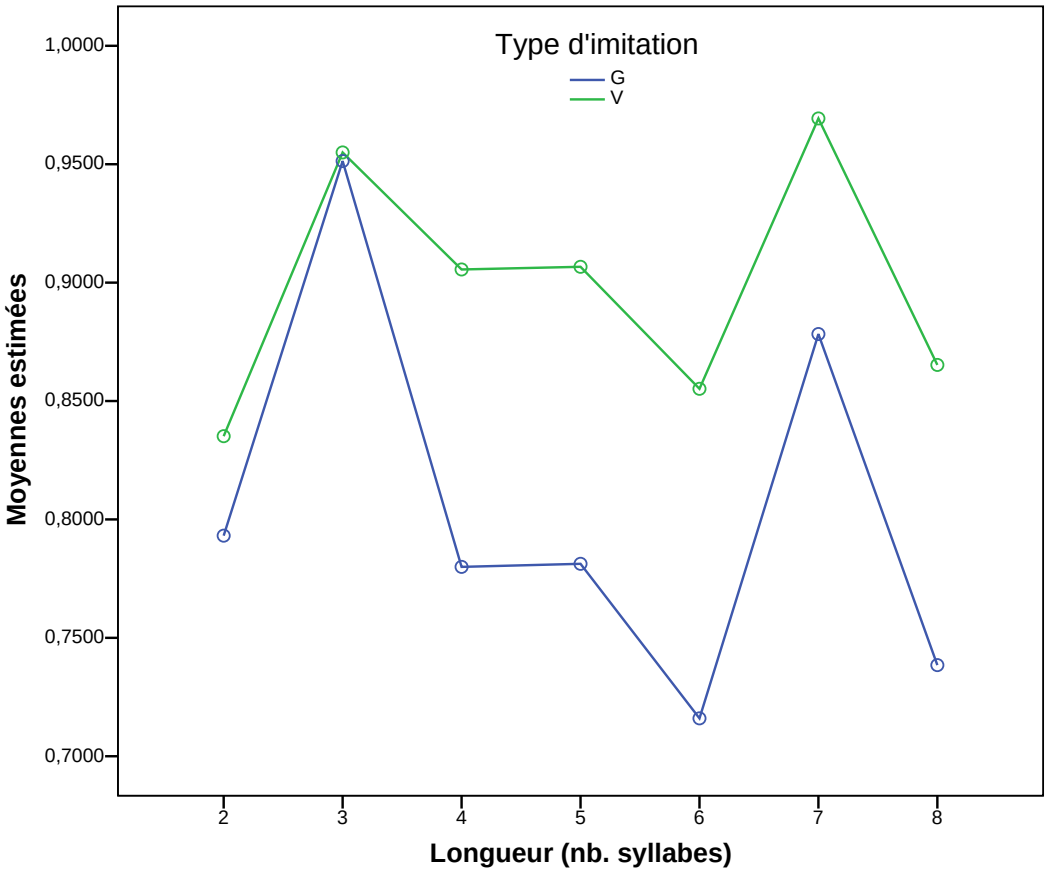


FIGURE 2.18 – *Résultats de corrélation suivant le type de phrase utilisée pour les imitations gestuelles (en bleu) et vocales (en vert).*

La forme des courbes 2.18 et 2.19 présentées ici sont sembables suivant la modalité utilisée pour l'imitation (vocale ou gestuelles). Ces résultats suggèrent donc que la longueur de la phrase n'a pas d'effet significatif sur les performances, puisqu'ici, les meilleurs résultats sont obtenus pour les stimuli de 3 et 7 syllabes. Il paraît plus probable que ce soit la syntaxe de la phrase qui joue un rôle sur les performances. En d'autres termes, la sémantique de la phrase utilisée jouerait un rôle non négligeable sur la capacité de reproduire une mélodie cible.

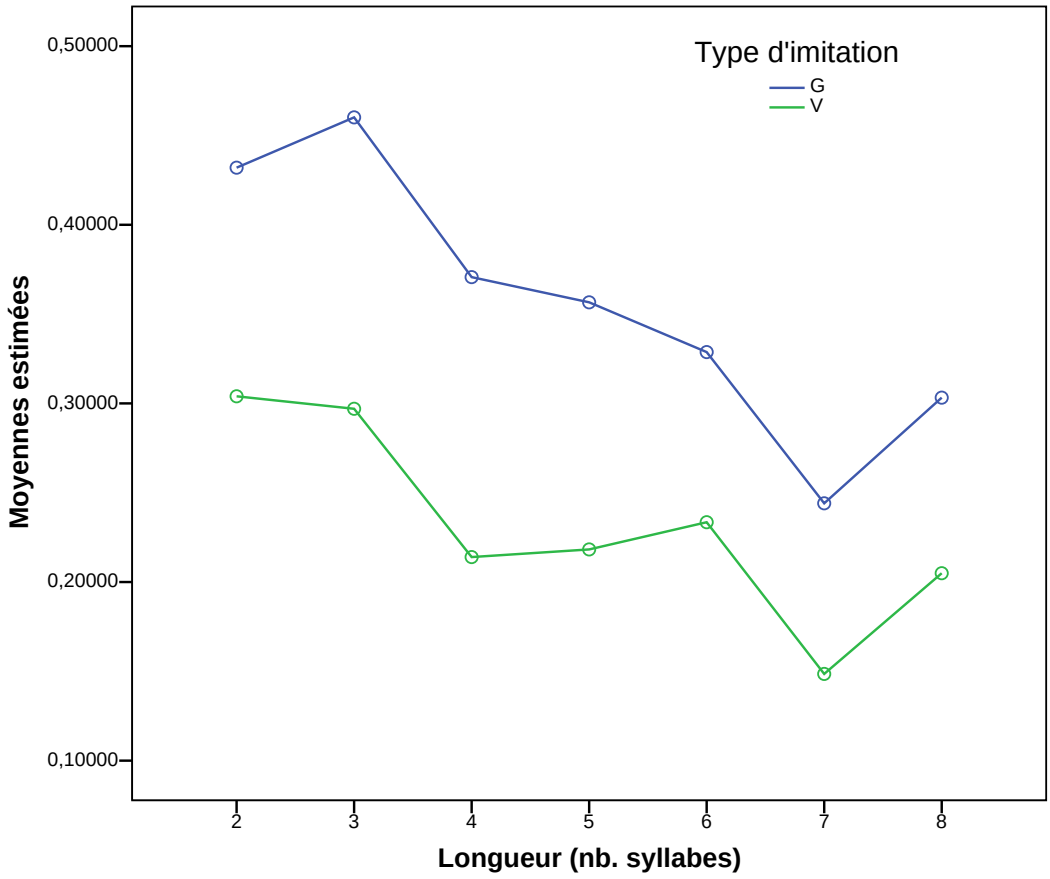


FIGURE 2.19 – *Résultats de moindres carrés suivant le type de phrase utilisée pour les imitations gestuelles (en bleu) et vocales (en vert).*

La figure 2.19 précédente nous montre que les résultats de distance au sens des moindres carrés sont meilleurs plus le nombre de syllabes augmente, que ce soit pour les imitations gestuelles ou vocales. On peut supposer, notamment pour les stimuli de 2 et 3 syllabes, que la durée du stimulus est trop faible pour pouvoir juger correctement de la valeur moyenne de la hauteur.

Concernant la contradiction observée avec la première expérience, il paraît difficile de conclure, du fait que les stimuli utilisés lors de la première expérience ne comprenait que des phrases du locuteur masculin, et que d'autre part, des phrases déclaratives étaient également utilisées.

2.5.6 Analyse gestuelle

Avant de rentrer dans les détails des analyses cinématiques et dynamiques proprement dites, un petit détour théorique s'avère nécessaire afin de bien comprendre les enjeux de ces analyses.

Geste et prosodie

Notre étude sur la modification de l'intonation et du rythme de la parole par le geste nous a amené à nous poser certaines questions concernant l'adéquation avérée des gestes manuels, voire d'écriture, et les gestes vocaux, comme l'intonation, par exemple. En effet, il ne paraît pas évident a priori que les variations de fréquence fondamentale de la parole soient régies par les mêmes lois que celles des mouvements manuels. Concernant la description de ces derniers, il existe une littérature abondante, décrivant notamment des lois d'invariance, permettant de décrire aisément les gestes humains grâce à des lois cinématiques connues. Il existe également en corollaire de ce domaine des études menées sur les gestes musculaires vocaux, tels les mouvements de la mâchoire ou du dos de la langue (Kelso et al., 1985; Tasko and Westbury, 2004; Sanguineti et al., 1998; Shiller et al., 2002).

Mais en ce qui concerne la prosodie, aucune étude, à notre connaissance, n'a été menée afin de caractériser la dynamique intonative en termes de mouvements. De nombreux modèles existent cependant en modélisation prosodique, plus ou moins raffinés, consistant à décrire l'évolution "naturelle" de la fréquence fondamentale pour la parole. Cette présente section a donc modestement pour but de mettre en regard les différentes courbes intonatives issues des imitations avec la tablette graphique avec les lois d'invariance du mouvement gestuel humain. Et, grâce aux expériences que nous avons menées, d'essayer d'aborder la question d'une modélisation prosodique en termes cinématique et dynamique.

Nous allons donc d'abord présenter les différentes lois d'invariance présentes dans la littérature, avant de montrer que nous retrouvons bien ces résultats avec les données gestuelles expérimentales. Parmi les questions que nous nous sommes posées à l'issue des expérimentations, se trouvent : Existe-t-il une méthode permettant de mettre à l'épreuve un modèle prosodique ? Est-ce qu'il est possible de prolonger les courbes d'intonation dans les parties non voisées ? Qu'est-ce que cela implique ?

Les Lois d'Invariance des Gestes Humains

Comme le rapporte S. Gibet (Gibet et al., 2004), toutes les études ayant portées sur les gestes humains peuvent être ramenées à trois ou quatre lois d'invariance de base, permettant de caractériser un mouvement humain, spécialisé ou non. Ce sont ces différentes lois que nous rapportons ici.

Invariance du Profil de Vélocité

Les mouvements avec plusieurs points d'inflexion engendrent des profils de vélocité dont la forme globale est en forme de cloche. En outre, cette forme présente une asymétrie qui dépend de la vitesse du mouvement. Plus la vitesse augmente et plus la courbe devient symétrique jusqu'à ce que l'asymétrie s'inverse.

Le Principe d'isochronie et la loi de Fitts

Le principe d'isochronie est l'un des principes fondamentaux du mouvement humain, et l'on peut le caractériser ainsi : si l'on demande à un sujet de réaliser aussi précisément que possible un mouvement d'aller-retour rectiligne entre deux points, ce sujet mettra autant de temps à réaliser le parcours, quelque soit la distance entre ces deux points. Ceci signifie donc que plus les cibles sont éloignées, et plus la personne se déplacera vite pour atteindre ces cibles. Inversement, plus les cibles sont rapprochées, et plus la personne se déplacera lentement. Evidemment, par induction, dans le premier cas la personne sera moins précise que dans le second cas, par application du principe de compromis entre vitesse et précision.

La loi de Fitts (Fitts, 1954) est une extension du principe d'isochronie décrit précédemment. P. Fitts introduit en effet la notion sous-jacente dans le principe d'isochronie, à savoir celle de précision. Il définit alors une taille de cible, permettant de mesurer si le sujet a effectivement atteint la cible, et avec quelle précision. Cette loi peut alors être exprimée sous la forme de l'équation suivante :

$$T = a + b.I_d = a + b.\log_2(\frac{2A}{W}) \tag{2.12}$$

où I_d représente l'index de difficulté et traduit la précision atteinte pour une taille W donnée de cibles, situées à un distance A l'une de l'autre. T est la durée requise pour exécuter le mouvement.

Des raffinements ont été apporté à cette loi dans le domaine des interfaces hommes-machines pour des mouvements coplanaires à deux dimensions (MacKenzie and Buxton, 1992), de manière à obtenir uniquement des valeurs de mesure positives, selon la formule suivante :

$$T = a + b.\log_2(\frac{A}{W} + c) \quad avec \quad c = 0.5 \text{ ou } c = 1 \tag{2.13}$$

Loi de Puissance Deux-Tiers

Pour les mouvements d'écriture et de dessin dans un plan, Viviani, Terzuolo & Lacquaniti (Viviani and Terzuolo, 1983; Lacquaniti et al., 1983) ont montré qu'il existait une relation entre la cinématique

des mouvements elliptiques et les propriétés géométriques de la trajectoire. La loi appelée loi de puissance deux-tiers établit alors une relation entre la vitesse angulaire ω et la courbure C de la trajectoire, exprimée sous la forme suivante :

$$\omega(t) = k.C(t)^{2/3} \tag{2.14}$$

où de manière équivalente :

$$v(t) = k.R(t)^{\beta} \qquad avec \qquad \beta = 1/3 \tag{2.15}$$

et où $v(t)$ représente la vitesse tangentielle et $R(t)$ le rayon de courbure, calculés de la manière suivante dans un repère cartésien $(O, x(t), y(t))$:

$$v(t) = \sqrt{\dot{x}(t)^2 + \dot{y}(t)^2} \tag{2.16}$$

$$R(t) = \frac{v(t)^3}{|\dot{x}(t).\ddot{y}(t) - \ddot{x}(t).\dot{y}(t)|} \tag{2.17}$$

La valeur empirique de β est de $1/3$ pour un adulte. Viviani (Viviani and Terzuolo, 1983) a alors montré que la valeur de k restait contante pour des mouvements elliptiques et que l'on pouvait ainsi découper le mouvement d'écriture par morceaux, délimités par des changements de valeurs de k et qu'il dénomme par unités d'action motrice. Notons que, bien que cette loi ait pu être vérifiée par un grand nombre d'études, elle reste valide pour des mouvements coplanaires d'amplitude relativement restreinte.

Minimisation de coût

Nelson (Nelson, 1983) a introduit dans son article de 1953, un certain nombre de coûts liés aux mouvements humains (le coût énergétique, le coût impulsionnel, le coût de jerk), ne pouvant être minimisés tous ensemble, mais indépendamment selon la/les contrainte(s) appliquée(s) sur le mouvement. Par exemple, nombre de mouvements sportifs, comme la course de sprint, vont chercher à minimiser le coût énergétique. En revanche, pour des mouvements qu'il appelle spécialisés, comme ceux de l'écriture ou de la pratique instrumentale musicale, l'optimisation recherchée est ici celle de la "douceur" du mouvement et le coût minimisé sera celui de la secousse (jerk).

Ce coût de secousse est exprimé par la moyenne quadratique de la dérivée de l'accélération, selon la formule suivante :

$$C = \frac{1}{2}\int_{t_1}^{t_2} \left[\left(\frac{dx^3}{dt^3} \right)^2 + \left(\frac{dy^3}{dt^3} \right)^2 \right] dt \tag{2.18}$$

avec t_1 et t_2 les instants de début et de fin du mouvement, et $(x(t), y(t))$ les coordonnées de la position. L'hypothèse que l'on souhaitait tester grâce à une analyse des gestes des sujets est la

suivante : est-ce que les performances d'imitation seraient liées à une minimisation de la secousse ? Avant de tester cette hypothèse, nous avons voulu voir si, conformément aux travaux de P. Viviani sur les gestes coplanaires, la loi de puissance deux-tiers était bien vérifiée.

Résultats expérimentaux

Sur la figure 2.20, sont représentées toutes les valeurs du facteur β [10] (en abscisse) selon la valeur de corrélation, pour chacun des sujets. On observe donc bien que pour tous les sujets que la valeur $\beta = \frac{1}{3}$ représente la valeur médiane du nuage de points. En revanche, le fait qu'une imitation soit plus proche ou non de cette valeur médiane ne renseigne en rien sur la performance. En d'autres termes, quelque soit la valeur de corrélation (bonne ou mauvaise), la valeur $\frac{1}{3}$ représente toujours la valeur médiane des points, avec une variance sensiblement égale. Cela dit, ces résultats sont conformes aux résultats de P. Viviani puisque nous avons toujours affaire ici à des mouvements coplanaires, quelque soit la performance.

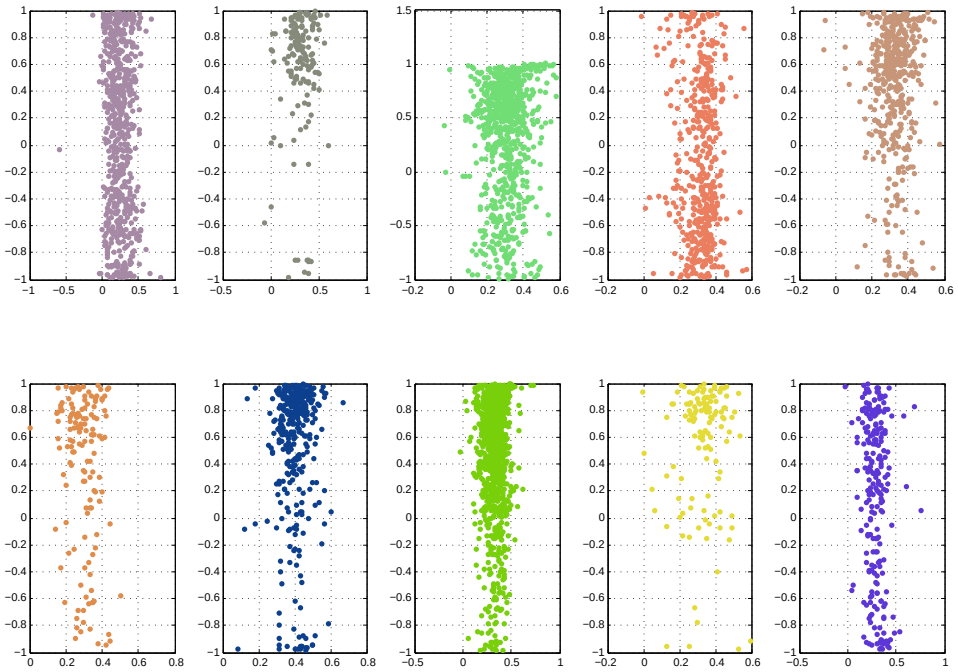


FIGURE 2.20 – *Visualisation, pour chaque sujet, des résultats de corrélation en fonction de la valeur de β obtenue par régression linéaire.*

10. les valeurs de β sont déduites pour chaque imitation par une régression linéaire entre les logarithmes de la vitesse et de celui du rayon de courbure.

Minimisation de la secousse

La figure 2.21 suivante représente les histogrammes, pour chaque sujet, les coûts de secousse, calculés d'après l'équation 2.18, suivant les valeurs de corrélation. Ces valeurs sont représentées à la fois sous la forme d'histogrammes 3D (en haut) et de leur valeurs cumulés sous la forme d'une carte de chaleur (en bas). Les sujets sont numérotés dans l'ordre croissant, en parcourant les graphes de gauche à droite, de haut en bas.

Ces analyses ne démontrent pas de manière claire et nette un lien entre la minimisation de la secousse et les performances. Malgré tout, les valeurs de secousse pour les sujets les meilleurs (7&8) semblent plus compactes et regroupés vers la valeur de secousse nulle, tandis que les sujets les moins bons (6&10) ont eux des valeurs de secousse plus éparses et plus élevées.

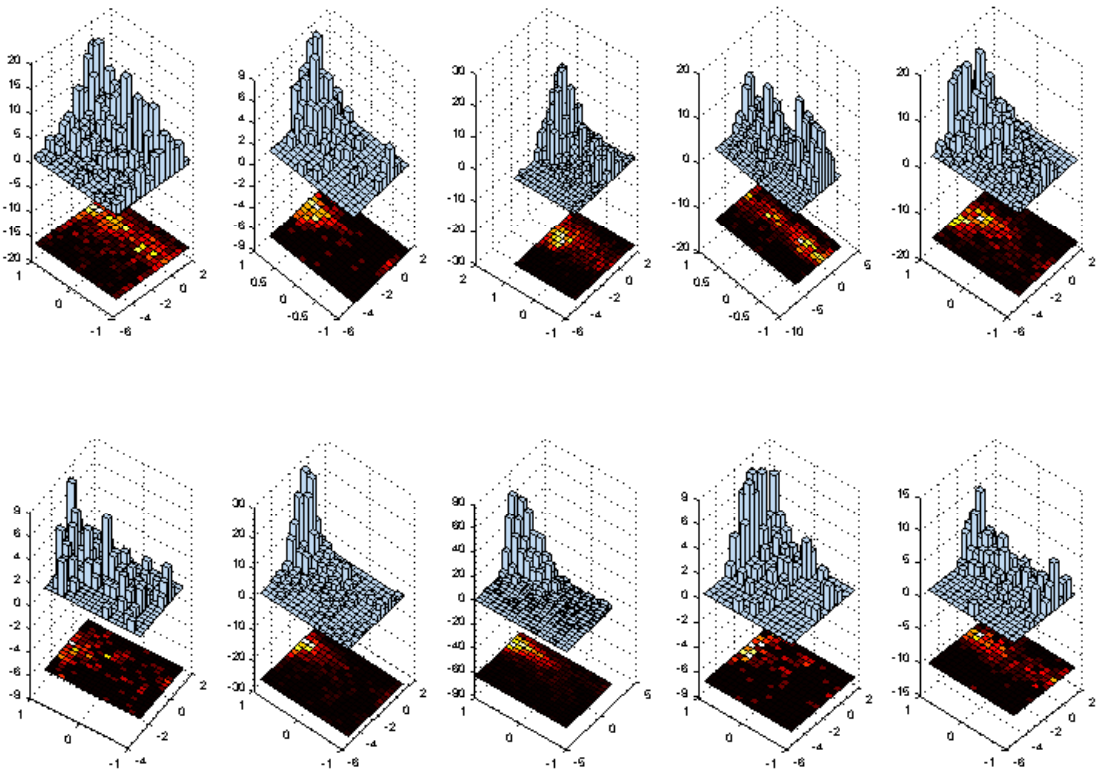


FIGURE 2.21 – *Visualisation, pour chaque sujet, des résultats de corrélation en fonction de la valeur de la secousse.*

Ces analyses cinématiques ne nous permettent pas de valider avec certitude l'hypothèse selon laquelle les performances d'imitations gestuelles des sujets ont un lien quelconque avec la minimi-

sation de la secousse. Cependant, les résultats obtenus confirment à la fois que la loi de puissance deux-tiers, et la minimisation de la secousse sont vérifiées pour tous les sujets. Cependant, ces mesures ne sont pas corrélées avec les performances des sujets. Ce fait permet de déduire que les sujets n'ont pas réalisé des gestes *au hasard*, mais se sont bien reposés sur leurs capacités d'écriture pour la réalisation de l'imitation prosodique. Ces résultats nous laisse donc penser que le système bio-mécanique constitué par le larynx obéirait aux mêmes lois d'invariance que d'autres organes ayant pu être étudiés, en ce qui concerne le contrôle de la fréquence fondamentale. Cette hypothèse doit toutefois être prise avec précaution, d'après certaines études récentes notamment de Maoz, ou Perrier (Maoz et al., 2006; Perrier and Fuchs, 2008). Leurs études montrent en effet que tout système mécanique du second ordre, auquel on ajoute du bruit, répond parfaitement à la loi de puissance deux-tiers.

2.6 Applications

Comme nous avons déjà eu l'occasion de le faire remarquer, le système Calliphonie nous a permis d'étudier sous un nouveau jour la problématique de la stylisation intonative. Outre cette thématique de recherche, il est possible d'envisager d'autres applications au système Calliphonie, comme ne le décrivons ci-dessous.

2.6.1 Enrichissement de base de données

Post-traitement

Une première application du système Calliphonie est directement dérivée du schéma développé pour l'évaluation de notre système : permettre à un utilisateur de changer l'intonation d'une phrase. De telles applications peuvent être utiles dans le domaine des synthétiseurs de parole afin d'augmenter leur expressivité. Le problème principal consiste alors à enregistrer et modéliser de manière adéquate les multiples corpora nécessaires pour faire face à n'importe quel type d'expression pour chaque phrase.

Notre proposition est de fournir la possibilité à l'utilisateur final d'ajouter directement l'expression dont il a besoin en sortie du synthétiseur grâce au système Calliphonie. Ce système est en effet simple d'utilisation et nécessite relativement peu d'entraînement. Par exemple, il sera facile d'ajouter une focalisation sur l'un des mots de la phrase.

Procédure d'évaluation Afin d'évaluer la capacité de notre système pour ajouter une certaine expression à la parole synthétique, une procédure d'évaluation a été mise en place. Elle se fonde exactement sur le même principe que les expériences précédentes pour l'imitation de l'intonation, à la différence près qu'ici la parole naturelle aplatie en entrée du système Calliphonie a été remplacé avec une phrase synthétique produit par le système de synthèse par concaténation SELIMSI (Prudon and d'Alessandro, 2001). L'utilisateur de Calliphonie entend la phrase originale issu de notre corpus, comportant soit une focalisation sur un mot soit une prosodie interrogative. L'utilisateur doit ensuite reproduire le contour intonatif de la phrase originale à partir de la phrase synthétique, de la même façon que précédemment.

La principale différence réside dans la durée segmentale des stimuli modifiés : tandis que pour l'évaluation précédente les durées étaient rigoureusement les mêmes, ici la phrase synthétique possède ses propres durées. Cela implique deux différences majeurs entre les deux protocoles. La première concerne la procédure de modification : il est en effet plus difficile de réaliser l'imitation lorsqu'un allongement trop prononcé est présent au sein de la phrase naturelle. La deuxième concerne la mesure de distance entre les contours intonatifs originaux et modifiés. Comme les valeurs de hauteur sont comparés au niveau des voyelles uniquement, et comme les voyelles

naturelles et synthétiques ne possèdent pas les mêmes durées la plupart du temps, au lieu d'extraire une valeur de hauteur toutes les 10 ms, 10 valeurs sont calculées, régulièrement espacées le long d'une voyelle. Ces 10 valeurs sont ensuite utilisées pour calculer les distances RMS et de corrélation. La table 2.3 fournit les résultats de ces mesures suivant les deux modalités utilisées.

	Corrélation	RMS
Focalisation	0,92	3,18
Interrogation	0,86	4,14
Moyenne	0,89	3,66

TABLE 2.3 – Valeurs de distances moyennes obtenues, pour les phrases focalisées, interrogatives et pour toutes les phrases. (Le Beux et al., 2007)

Résultats et analyses Les résultats obtenus sont sensiblement les mêmes que ceux obtenus avec la voix naturelle. Les distances RMS et les corrélations sont plutôt bonnes (cf. table 2.3), et indiquent une stylisation proche de la courbe intonative pour les stimuli synthétiques, même en présence de différences temporelles. Les scores obtenus pour la focalisation et l'interrogation sont comparables, avec une performance globale légèrement supérieure pour la focalisation. Concernant l'effet de la longueur de la phrase, (cf. figure 2.22), le phénomène est quelque peu plus complexe que pour la parole naturelle : si la corrélation diminue graduellement avec la longueur de la phrase, comme observé précédemment, la distance RMS ne présente pas de tendance particulière, exceptée pour les phrases courtes, présentant des scores plus élevés, contrairement à la parole naturelle.

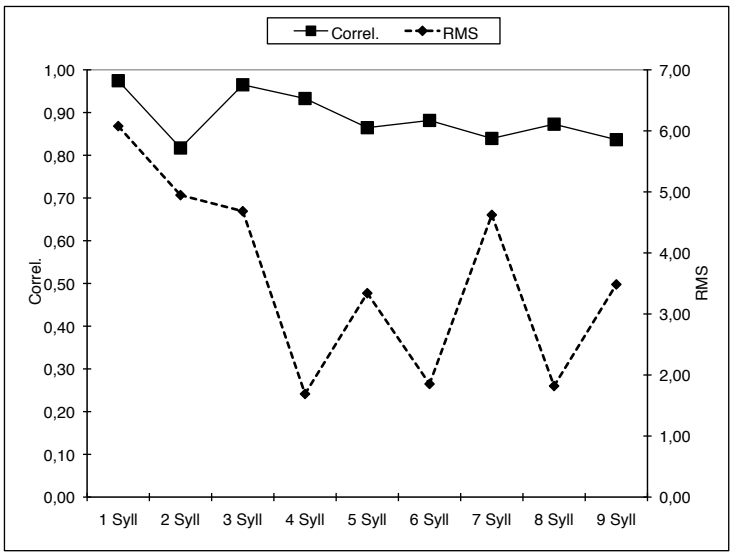


FIGURE 2.22 – *Distances* RMS *(losanges) et corrélations (carrés) suivant la longueur de la phrase. (Le Beux et al., 2007)*

La distance objective entre les paramètres prosodiques modifiés en sortie de Calliphonie et la prosodie naturelle originale, est plutôt faible, donnant une assez bonne idée de la capacité du système pour imiter la prosodie d'une phrase donnée. Le système Calliphonie fournit donc l'opportunité de produire de la parole expressive.

Cependant, il doit être noté que le paramètre de durée n'est pas pris en compte ici. Ce qui n'est pas suffisamment satisfaisant pour une synthèse expressive de qualité, pour laquelle la modification de durée est nécessaire. Dans notre implémentation actuelle, Calliphonie fait appel a deux modifications successives (concaténation et PSOLA), une situation qui n'est pas optimale. Bien qu'il reste encore des améliorations à apporter pour obtenir une meilleure qualité sonore, nous pensons que la capacité des sujets à ajouter de l'expressivité aux synthétiseurs a été démontré de manière convaincante.

Si l'on considère les bases de données n'ayant pas été précédemment annotées, ce système peut toutefois être utilisé d'une manière un peu différente. Lorsque le but est simplement de produire des phrases expressive (pour des expériences perceptives, par exemple), alors il est possible de modifier en ligne les phrases synthétisées et d'enregistrer directement après modification.

Cette procédure fournit la possibilité à une personne pas nécessairement familière avec la synthèse et le traitement de parole de produire des phrases expressives d'une manière simple, sans avoir à acheter un système coûteux ou acquérir une connaissance approfondie en traitement du signal de parole. En outre, il est possible d'utiliser les phrases synthétiques issues de n'importe quel système TTS disponible publiquement sur internet ou d'enregistrer directement quelques phrases soi-même avec un simple microphone, avant de réaliser la modification grâce à Calliphonie.

Enrichissement de base de données

Une autre application du système Calliphonie concerne spécifiquement la synthèse de parole par règles. Les systèmes de synthèse basés sur la sélection/concaténation d'unités acoustiques non uniformes nécessite de larges corpora de parole enregistrée. Notre système peut ainsi être utilisé pour permettre l'enrichissement de la base de données de parole, avant même la synthèse. Dans ce cas précis, la parole naturelle est modifiée, et une même phrase peut être modifiée selon plusieurs variations prosodiques, comme décrit sur la figure 2.23.

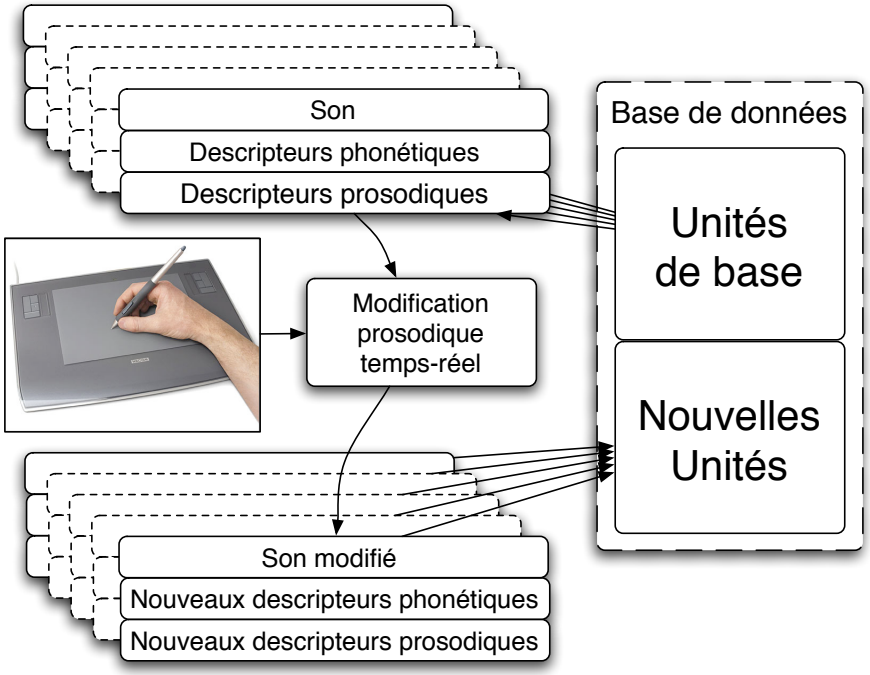


FIGURE 2.23 – *Description du processus d'enrichissement d'une base de données. (Le Beux et al., 2007)*

Plusieurs étapes sont nécessaires pour l'obtention de cet enrichissement et pouvant être appliqué à plusieurs types de bases de données. Il n'existe pas de contrainte sur le contenu de la base de données. Calliphonie peut ainsi être utilisé pour ajouter de nouvelles expressions n'ayant pas été enregistrées, ou pour obtenir plus d'occurrences d'une expression sous-représentée.

Ainsi, le contenu prosodique de la base de données peut être étendu et/ou amélioré sans nécessiter de nouvel enregistrement. Ce traitement est complètement indépendant du type de base de données, puisqu'il s'agit uniquement d'un pré-traitement de cette base.

2.6.2 Voix chantée

Nous n'avons jusqu'alors pas testé la possibilité de modifier la hauteur et la durée d'en enregistrement de voix chantée. Il paraît toutefois parfaitement envisageable de modifier une voix chantée, moyennant quelques adaptations. La principale adaptation concerne la gestion du vibrato, car comme le rappelle M. Castellengo (Castellengo et al., 1989), le vibrato dépendra de la longueur de la note, de sa hauteur et également de la phase. Il paraît néanmoins intéressant d'étudier l'effet de notre système sur de la voix chantée, d'autant plus que notre système contrôlé par le geste permet d'aborder la question du vibrato de manière non supervisée. En effet, les fréquences de variations du vibrato (entre 4 et 10 Hz) sont inférieures aux limites fréquentielles atteignables par le geste humain

(environ $30 - 50$ Hz).

Le succès indéniable de logiciels tels que Auto-Tune de la société Antares (Auto-Tune, 2009) nous montre bien qu'il existe un intérêt pour la correction de hauteur pour des chanteurs inexpérimentés ou l'obtention d'effets vocaux audio-numériques.

2.7 Conclusions du chapitre

D'un point de vue pratique, notre étude sur la modification de la hauteur et de la durée de parole nous a amené à développer un outil de modification en temps réel, reposant sur l'algorithme PSOLA. Ce système nous permet à l'heure actuelle de pouvoir réaliser des changements d'intonation et de rythme de la parole avec une assez grande facilité. Cet outil permet en outre de fournir à des personnes ne possédant pas ou peu de connaissances en traitement du signal de parole, la possibilité d'étudier les modifications prosodiques selon un paradigme simple et interactif. Nos différentes présentations au cours de conférences nous ont à ce sujet confirmé que l'intérêt de disposer d'un tel outil parmi la communauté de recherche en prosodie était vivace.

Des améliorations peuvent encore être apportées du point de vue technique afin de consolider la robustesse de notre approche. L'un des points intéressants à aborder concerne l'organisation rythmique d'une occurrence de parole. Il paraît en effet judicieux de pouvoir disposer d'une modification de la durée de parole, non seulement locale, mais également plus longue dans le temps. En d'autres termes, actuellement notre système offre la possibilité de modifier la durée de la parole en temps réel. Mais, il serait peut-être plus pertinent d'échantillonner la valeur de modification temporelle sur des échelles de temps plus longues telles que les syllabes, les phonèmes ou les p-centres, par exemple. Cette opportunité fournirait sans nul doute une approche novatrice pour l'étude du rythme de la parole.

Sur le plan expérimental, notre approche de modification gestuelle de l'intonation et du rythme de la parole s'est avérée pertinente. Les résultats de nos expériences nous en effet confirmé que la possibilité de modifier l'intonation d'une phrase donnée par le geste manuel est comparable, en termes de performances, à une imitation effectuée avec notre propre voix. Les résultats de notre seconde expérience nous apportent en outre la confirmation, que l'entraînement lié au système n'est pas négligeable, mais avec un temps d'entraînement adéquat, il est possible d'améliorer ses performances avec le système.

Concernant l'aspect théorique, notre approche place sous un nouveau jour l'étude de la modélisation prosodique. En effet, jusqu'à aujourd'hui, l'intonation était étudiée soit sous la forme de segment contrastés (haut, bas, ascendant ...), voire de fonctions plus élaborées comme des splines continues. La modification gestuelle de l'intonation, voire du rythme, offre donc la possibilité de s'abstraire de formes intonatives ou prosodiques établies *a priori*, pour se pencher sur une analyse *a posteriori* des mouvements gestuels les plus efficaces ou pertinents. A ce titre, nos expériences ont confirmé le fait que les gestes effectués pour l'imitation d'une intonation donnée répondait aux mêmes contraintes que les gestes d'écriture, ce qui constitue un résultat intéressant pour approfondir la question de la modélisation prosodique.

Notre expérience avec le système de modification prosodique nous laisse également penser que

la modification rythmique se satisfait également de la dynamique offerte par le geste manuel. Les variations dynamiques relatives de durée des structures de la parole sont en effet du même ordre de grandeur que celle de l'intonation, à ceci près que les variations locales sont sans doute moindres.

Toute notre étude sur la modification prosodique de la hauteur et de la durée de parole laisse présager d'un certain nombre d'applications possibles pour notre outil. En premier lieu pour les systèmes de synthèse à concaténation d'unités acoustiques, pour lesquels il est possible d'enrichir une base de données avec de nouvelles attitudes ou expressions, sans avoir besoin de réenregistrer un nouveau corpus, ou pour accomplir des corrections en post-traitement du synthétiseur, pour des phrases dont la prosodie ne serait pas assez convaincante.

De plus, comme nous l'avons déjà fait remarquer, notre système pourrait s'avérer d'une grande utilité pour la communauté de recherche en prosodie, afin de se rendre compte de manière plus interactive des modifications apportées sur l'intonation et/ou le rythme d'une phrase. Par ailleurs, le fait de pouvoir partager notre expérience et notre système avec un nombre de personnes plus larges contribuerait grandement à l'amélioration de notre système et à l'émulation potentielle des communications dans ce domaine.

Au cours du prochain chapitre, nous allons nous intéresser aux dimensions prosodiques restantes, et plus spécifiquement à la qualité vocale, qui joue également un rôle non négligeable sur les expressions et les attitudes vocales.

Chapitre 3

Synthèse de source glottique

Sommaire

3.1 Synthèse de Source Glottique et Qualité Vocale **93**
3.1.1 Modèle Linéaire Source/Filtre 94
3.1.2 Les Principaux Modèles Signal de Source Glottique 96
3.1.3 Le Modèle Linéaire Causal/Anticausal (CALM) 106
3.2 Phonétique de la qualité vocale **111**
3.2.1 La notion de registre vocal 112
3.2.2 La dimension de bruit 120
3.2.3 L'effort vocal 121
3.2.4 La dimension tendue/relâchée 121
3.3 Le modèle CALM en temps réel ou RTCALM **123**
3.3.1 Les contraintes 123
3.3.2 Les solutions 123
3.3.3 Composantes non périodiques 127
3.3.4 Description des fonctions de mapping 131
3.4 Les différents instruments basés sur RTCALM **140**
3.4.1 Premier instrument 140
3.4.2 Deuxième instrument 142
3.4.3 Méta-instrument 145
3.4.4 Exploration haptique du phonétogramme 150
3.4.5 Réflexions sur l'adéquation interface/synthétiseur 155
3.5 Applications **156**
3.5.1 Voix chantée 156
3.5.2 Synthèse de qualité vocale et génération de stimuli 157
3.5.3 Apprentissage phonétique 157

3.1 Synthèse de Source Glottique et Qualité Vocale

Le but principal de ce chapitre est de s'intéresser à la question de la qualité vocale, déjà évoquée dans les précédents chapitres, et comme nous allons le voir, faisant également partie intégrante de la prosodie, dans son acception élargie (Campbell and Mokhtari, 2003; Pfitzinger, 2006). Nous avons choisi, afin d'étudier les phénomènes liés à la source glottique, d'utiliser le cadre de travail de l'analyse par la synthèse. Cette méthodologie largement utilisée dans le domaine de recherche en qualité vocale, a sans doute été popularisée grâce à Kenneth Stevens (Stevens, 1960) en reconnaissance de parole, puis largement utilisée ensuite en synthèse par Dennis & Laura Klatt (Klatt and Klatt, 1990). Le principe de base de cette méthode consiste à dire qu'en synthétisant une occurrence vocale, en réalisant son analyse objective ou subjective [1], puis en faisant une comparaison avec le son que l'on cherche à synthétiser, cela nous permet de connaître le chemin parcouru jusqu'à l'occurrence *idéale*.

En outre, plus récemment, J. Kreiman rapporte dans un article (Gerratt and Kreiman, 2001) que cette méthode permet de mener avec plus de précision et de consistance des tests perceptifs relatifs aux voix pathologiques, qu'avec des tâches plus classiques d'évaluation par échelle graduées. Plus précisément, elle montre que le fait de pouvoir manipuler chacun des paramètres de synthèse (*presqu'en temps-réel* [2]) permet d'obtenir des corrélations inter et intra individuelles plus fortes qu'avec les méthodologies traditionnelles.

Il est toujours difficile de dissocier ces deux notions que sont la qualité vocale et la modélisation de source glottique. En effet, lorsque l'on évoque le sujet de la qualité vocale, cela revient quasi systématiquement à étudier le fonctionnement de la source glottique, car même s'il a pu être montré qu'un couplage de la source glottique et du conduit vocal existait (Flanagan, 1965; Rothenberg, 1986), celui-ci n'est observable et modélisable que dans le cadre d'une étude aéro-acoustique de la production de parole. En revanche, dans le cadre du modèle linéaire source/filtre, ce couplage est par définition occulté puisque la séparation de la contribution de la source glottique et du conduit vocal représente le principe de base de ce modèle.

Nous nous plaçons dans ce chapitre dans le cadre de la modélisation source/filtre de la production de parole, et cela pour plusieurs raisons :

1. D'un point de vue calculatoire, ce modèle est bien plus simple à manipuler.
2. En première approximation, le couplage entre la source glottique et le conduit vocal peut être considéré comme négligeable.
3. Il existe encore aujourd'hui de nombreuses incertitudes quant à la définition même de la qualité vocale et de ses dimensions perceptives.

1. via des méthodes d'analyse du traitement du signal ou des tests d'écoute perceptifs
2. *in near-real time* selon les termes de l'article

4. Un modèle "signal" de source glottique nous permet déjà de modifier les paramètres de la source et d'observer les effets produits en termes de qualité vocale. L'effet de la modification de ces paramètres engendre par la suite un raffinement de la modélisation.

Notons ici également que notre étude a pour but principal d'approfondir la notion de synthèse vocale expressive. Aussi, nous n'avons délibérément pas essayé de construire un synthétiseur permettant de réaliser tous les sons de parole, pour pouvoir nous concentrer sur la synthèse de voyelles, et surtout le contrôle des paramètres de source glottique au cours de la production de ces voyelles.

Avant de décrire les différentes expérimentations et implémentations de synthétiseurs de source glottique que nous avons pu réalisé, nous allons réaliser une brève introduction du modèle source/filtre linéaire, puis faire une revue des principaux modèles de source glottique historiques. Dans un second temps, nous nous intéresserons à la définition de la qualité vocale en phonétique, en parallèle avec des méthodes de traitement du signal pour réaliser les différentes dimensions perceptives de la source. Enfin, nous présenterons un certain nombre d'instruments vocaux que nous avons pu réaliser au cours de notre étude.

3.1.1 Modèle Linéaire Source/Filtre

L'organe vocal peut être décrit simplement par un système source/filtre (Fant, 1960). La source glottique est un générateur débit d'air non linéaire où le son est produit par des mouvements complexes des plis vocaux (situés dans le larynx) sous la pression des poumons. Une étude approfondie des caractéristiques temporelles et spectrales de la source glottique peut être trouvée dans (Henrich, 2001). Les sons produits par le larynx se propagent ensuite dans les cavités orales et nasales qui peuvent être vues comme un filtrage variable dans le temps, suivant les mouvements et les configurations des différents articulateurs de ces cavités supra-glottiques. Enfin, le flux de débit est converti en une onde de pression rayonnée par les lèvres et les narines. On retrouve les différentes étapes de la production vocale selon le modèle source-filtre sur la figure 3.1.

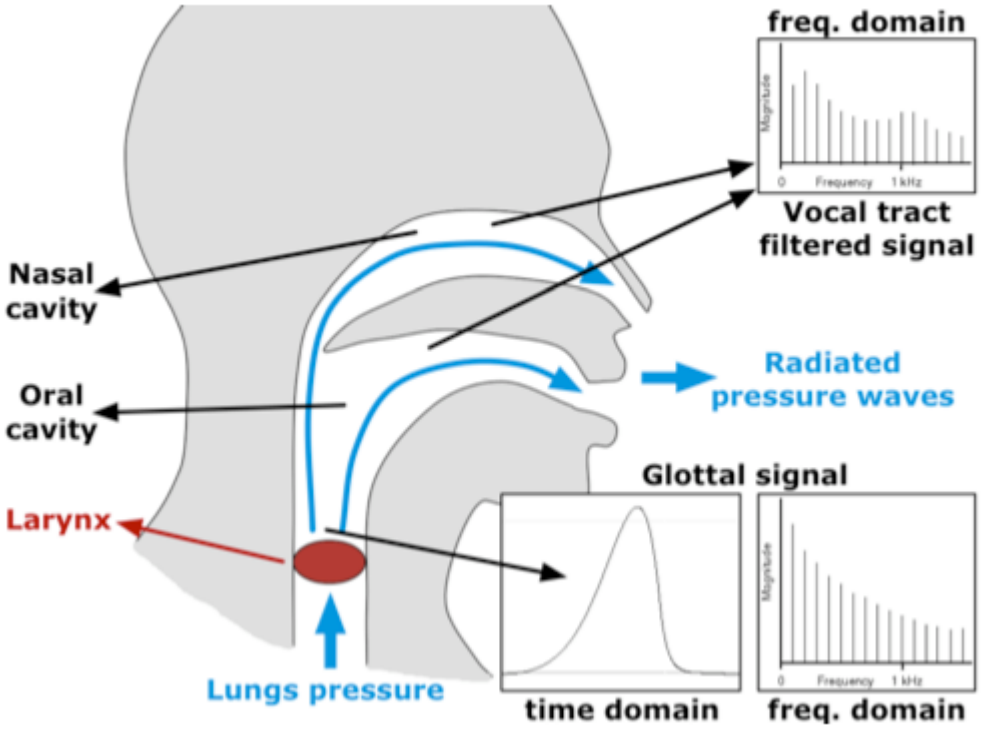


FIGURE 3.1 – *Représentation schématique du fonctionnement du modèle source-filtre (D'Alessandro et al., 2006)*

Le modèle source/filtre de production de la parole stipule que :

$$s(t) = e(t) * v(t) * l(t) \tag{3.1}$$

où l'opérateur $*$ représente l'opération de convolution, $s(t)$ est le signal de parole rayonné, $v(t)$ est la réponse impulsionnelle du conduit vocal, $e(t)$ est la source d'excitation vocale et $l(t)$ est la réponse impulsionnelle de la composante de rayonnement sonore au niveau des lèvres.

Si l'on traduit cette formule dans le domaine spectral, on obtient alors :

$$S(\omega) = E(\omega) \times V(\omega) \times L(\omega) \tag{3.2}$$

où $S(\omega), E(\omega), V(\omega)$ et $L(\omega)$ sont respectivement les transformées de Fourier des fonctions $s(t)$,$e(t)$, $v(t)$ et $l(t)$. Ceci implique donc qu'il est théoriquement possible de séparer simplement les différentes contributions des différents organes vocaux (i.e. larynx, conduit vocal et lèvres) dans le domaine de Fourier.

En outre, il est communément admis que la composante de rayonnement des lèvres consiste en une simple dérivation, et il est alors possible de simplifier cette formule en ramenant la dérivation au

niveau de la source d'excitation, pour obtenir alors la formule simplifiée suivante :

$$S(\omega) = E(\omega)' \times V(\omega) \tag{3.3}$$

L'effet des lèvres et des narines est généralement réalisé par un filtre linéaire passe-haut du premier ordre invariable dans le temps (Fant, 1960). L'effet du conduit vocal peut, quant à lui, être modélisé par un filtrage du signal glottique par plusieurs filtres linéaires résonants du second ordre (formants) [3].

La contribution de la source $e(t)$ est en réalité un signal composé, somme d'une composante quasi-périodique $p(t)$ et d'une composante de bruit $r(t)$. En effet, le passage de l'air à travers la glotte engendre des turbulences qui se propagent le long du conduit vocal. Cette composante de bruit peut s'exprimer dans le domaine temporel sous la forme :

$$s(t) = [p(t) + r(t)] * v(t) * l(t) = \Big[\sum_{i=-\infty}^{+\infty} \delta(t - it_0) * u_g(t) + r(t) \Big] * v(t) * l(t) \tag{3.4}$$

Ce qui revient, dans le domaine spectral, à :

$$S(w) = [P(w) + R(w)] \times V(w) \times L(w) \tag{3.5}$$

$$S(w) = \Big[\Big(\frac{1}{2\pi} \sum_{i=-\infty}^{+\infty} \delta(w - iw_0) \Big) \big|U_g(w)\big| e^{j\theta_{U_g(w)}} + \big|R(w)\big| e^{j\theta_{r(w)}} \Big] \times \big|V(w)\big| e^{j\theta_{v(w)}} \times \big|L(w)\big| e^{j\theta_{l(w)}} \tag{3.6}$$

où $p(t)$ est la composante quasi-périodique de l'excitation, $u_g(t)$ est le signal de débit glottique, t_0 est la période fondamentale, $r(t)$ est la composante de bruit de l'excitation glottique, δ la distribution de Dirac, $P(w), R(w), U_g(w)$ sont respectivement les transformées de Fourier de $p(t), r(t), u_g(t)$ et où $f_0 = \frac{1}{t_0}$ est la fréquence fondamentale du voisement.

Dans la prochaine section, nous allons nous intéresser aux différents modèles de source glottique, et plus particulièrement, dans la continuité du présent paragraphe, à l'onde de débit glottique $u_g(t)$ (*resp.* $U_g(\omega)$) et à sa dérivée. Nous détaillerons également plus tard dans ce chapitre, la contribution liée à la composante apériodique de la source.

3.1.2 Les Principaux Modèles Signal de Source Glottique

Lorsque l'on traite de l'expressivité vocale, l'intonation et le rythme occupent bien souvent une place primordiale, car ils constituent les paramètres acoustiques les plus facilement accessibles, surtout en analyse. Cependant, la synthèse vocale par formants nécessite de pouvoir générer

3. 4 ou 5 filtres formantiques suffisent pour l'obtention d'un signal vocal réaliste

un train impulsionnel qui sera ensuite filtré par le conduit vocal. Ce train de pulsation peut être plus ou moins raffiné, et les premières tentatives de modélisation reposaient simplement sur des trains d'impulsions (selon Dudley, un bourdonnement - *buzz* - (Dudley, 1939)). Bien qu'en première approximation, cette modélisation soit acceptable en termes d'intelligibilité, on se retrouve alors bien loin de pouvoir traiter le problème de l'expressivité grâce à de tels modèles.

Si la modification prosodique permet de réaliser certaines expressions (telle que l'interrogation, le doute ...), elle ne fournit cependant pas les outils nécessaires pour traiter de la notion de qualité vocale. La qualité vocale constitue le cadre relatif aux évolutions dynamiques de la source glottique permettant de catégoriser, par exemple, des voix stressées, relâchées, soufflées, voire pathologiques. Aussi, l'amélioration des modèles de source glottique s'est poursuivie tout au long du $\mathrm{XX}^{\mathrm{ème}}$ siècle afin d'adresser justement le problème de l'expressivité, ou tout du moins au départ, du naturel de la voix synthétique. Il existe aujourd'hui plusieurs modèles de source glottique plus ou moins équivalents, parmi lesquels les populaires KLGLOTT (Klatt and Klatt, 1990; Klatt, 1980) et LF (Fant, 1995; Fant et al., 1985). Nous allons donc ici, dans un premier temps décrire les particularités de ces différents modèles avant de montrer l'influence de leurs différents paramètres sur l'expressivité en synthèse.

Dans leur récent article, B. Doval (Doval et al., 2006) décrivent les caractéristiques temporelles et spectrales des principaux modèles historiques de source glottique. Nous ne reportons ici que les définitions temporelles, qui sont finalement leurs descriptions originelles par leurs auteurs respectifs, des différentes fonctions permettant de réaliser les différents synthétiseurs associés. Par la suite nous décrirons plus en détails les caractéristiques spectrales d'un modèle générique dérivé de tous ces différents modèles temporels, mais permettant une manipulation plus efficace et pertinente des dimensions de qualité vocale.

Il a été montré par B. Doval (Doval et al., 2003) que certains des principaux modèles de source glottique, nommément les modèle KLGLOTT88, R++, Rosenberg-C et LF, présentent tous des caractéristiques temporelles communes concernant l'onde de débit glottique (ODG).

On peut retrouver les formes typiques de l'ODG et de sa dérivée (ODGD) sur la figure 3.2 suivante.

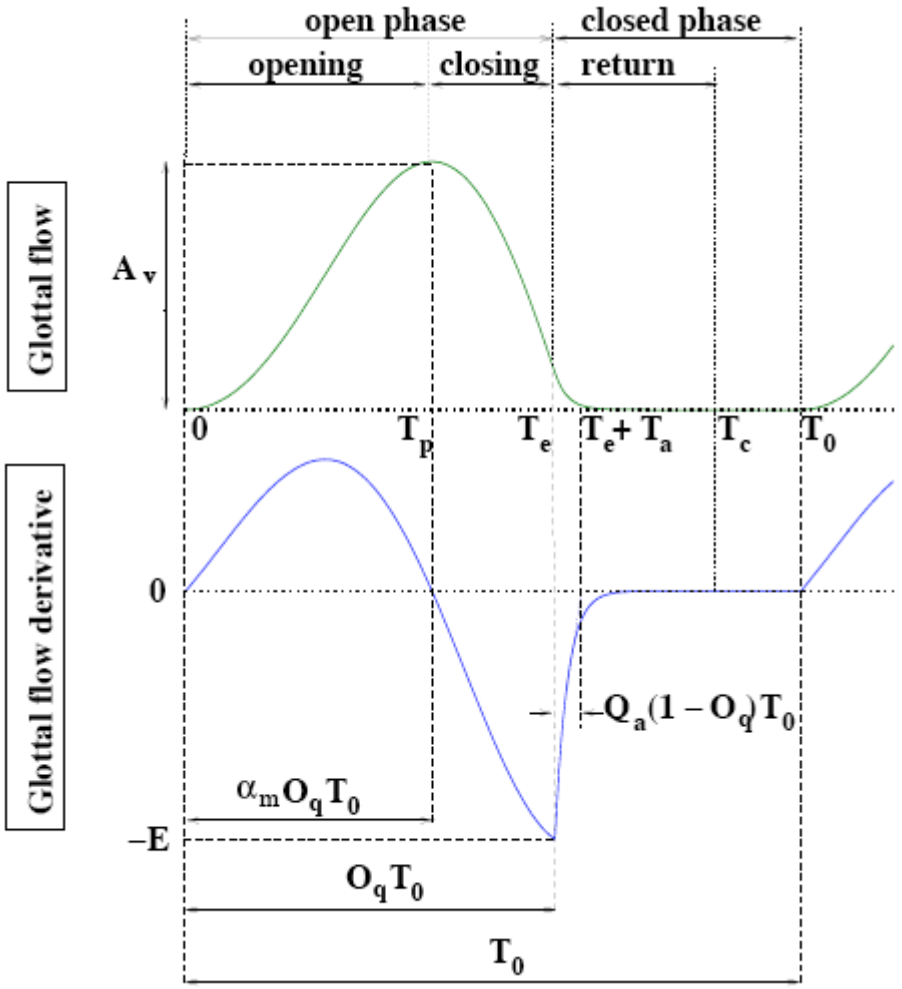


FIGURE 3.2 – *Formes typiques de l'onde de débit glottique (en haut) et de sa dérivée (en bas), et leurs paramètres temporels respectifs (d'après (Doval et al., 2003))*

Concernant les caractéristiques de celles-ci, il est bon de se souvenir de certaines descriptions qualitatives :

- Le débit glottique est toujours positif ou nul
- Le débit glottique et sa dérivée sont quasi-périodiques
- Pendant une période fondamentale, le débit glottique a une forme de cloche : il augmente, décroît, puis devient nul.
- Pendant une période fondamentale, la dérivée de l'onde de débit glottique est positive, négative, puis nulle
- Le débit glottique et sa dérivée sont continues et dérivables, sauf, pour la dérivée, lors de l'instant de fermeture de la glotte (GCI [4]).

4. Glottal Closure Instant

Dans la suite de ce document, nous dénommerons l'onde de débit glottique et sa dérivée respectivement par les termes ODG et ODGD pour plus de commodité. Ces différents modèles ont été regroupés, selon un modèle générique commun par B. Doval (Doval et al., 2006), mais avant de définir ce modèle générique, nous allons faire une incursion sur les définitions des différents modèles de source glottique précités.

Modèles de Rosenberg

Dans son article de 1970, A. E. Rosenberg (Rosenberg, 1971) a testé l'influence de la forme de l'onde de débit glottique sur la qualité vocale. Pour cela, il a utilisé au total 6 formes d'ondes différentes, dénommés par des lettres allant de A à F. Le modèle A est un modèle triangulaire, le B polynomial, C,D et E sont trigonométriques et enfin F est trapézoïdal. Les différentes formes d'ondes de débit glottique sont reportées sur la figure 3.3.

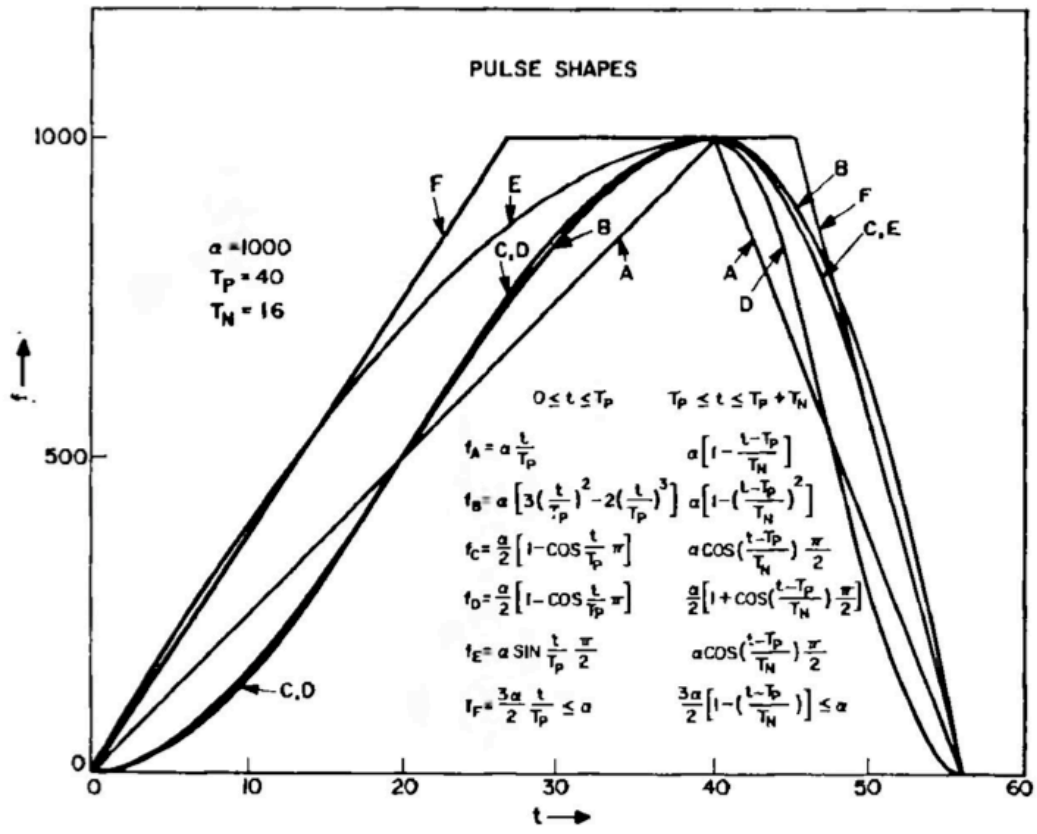


FIGURE 3.3 – *Formes et expressions des différents modèles de Rosenberg (d'après Rosenberg (Rosenberg, 1971)).*

Tous les modèles de Rosenberg sont exprimés à l'aide de 4 paramètres :

- A : l'amplitude
- T_0 : la période fondamentale
- T_p : l'instant du maximum de l'ODG.
- T_n : l'intervalle de temps entre le maximum de l'ODG et le GCI

Le modèle Rosenberg-C est constitué de deux parties sinusoïdales :

$$U_g(t) = \begin{cases} \frac{A}{2}(1 - \cos(\pi \frac{t}{T_p})) & 0 \leq t \leq T_p \\ A\cos(\frac{\pi}{2} \frac{t - T_p}{T_n}) & T_p \leq t \leq T_p + T_n \\ 0 & T_p + T_n \leq t \leq T_0 \end{cases} \tag{3.7}$$

et sa dérivée est alors exprimée par la fonction continue par morceaux :

$$U_g^{'}(t) = \begin{cases} \frac{\pi A}{2T_p} \sin(\pi \frac{t}{T_p}) & 0 \leq t \leq T_p \\ -\frac{\pi A}{2T_n} \sin(\frac{\pi}{2} \frac{t - T_p}{T_n}) & T_p \leq t \leq T_p + T_n \\ 0 & T_p + T_n \leq t \leq T_0 \end{cases} \tag{3.8}$$

Nous détaillons ici également le modèle Rosenberg-B, car il est à la base des modèles KLGLOTT et $R++$. Ce modèle est, quant à lui, constitué de deux fonctions polynomiales :

$$U_g(t) = \begin{cases} A(3(\frac{t}{T_p})^2 - 2(\frac{t}{T_p})^3) & 0 \leq t \leq T_p \\ A(1 - (\frac{t - T_p}{T_n})^2) & T_p \leq t \leq T_p + T_n \\ 0 & T_p + T_n \leq t \leq T_0 \end{cases} \tag{3.9}$$

et sa dérivée est alors exprimée par la fonction continue par morceaux :

$$U_g^{'}(t) = \begin{cases} \frac{6A}{T_p^2}(t - T_p) & 0 \leq t \leq T_p \\ \frac{2A}{T_n^2}(t - T_p) & T_p \leq t \leq T_p + T_n \\ 0 & T_p + T_n \leq t \leq T_0 \end{cases} \tag{3.10}$$

Après vérification, je me suis rendu compte d'une erreur dans l'expression de la version D, qui ne revient pas à 0. La correction supposée (suppression du facteur $1/2$ dans le cosinus) de cette expression est la suivante :

$$U_g(t) = \begin{cases} \frac{A}{2}(1 - \cos(\pi \frac{t}{T_p})) & 0 \leq t \leq T_p \\ \frac{A}{2}(1 + \cos(\pi \frac{t - T_p}{T_n})) & T_p \leq t \leq T_p + T_n \\ 0 & T_p + T_n \leq t \leq T_0 \end{cases} \tag{3.11}$$

Modèle KLGLOTT88

Ce modèle est issu du modèle Rosenberg-B (Eqs. 3.9 et 3.10) et a été utilisé pour le synthétiseur KLSYN88 de Klatt & Klatt (Klatt and Klatt, 1990).

La forme d'onde est caractérisée par 4 paramètres également :
- $F_0 = 1/T_0$: la fréquance fondamentale
- A_V : l'amplitude de voisement
- O_q : le quotient ouvert
- T_L : l'atténuation du filtre en hautes fréquences (*spectral tilt*)

Rappelons, à toutes fins utiles, que le paramètre O_q est égal à T_e/T_0, comme illustrée précédemment sur la figure 3.2.

Dans le cas où le paramètre T_L est nul, le modèle KLGLOTT88 est alors identique au modèle Rosenberg-B, et son équation est la suivante :

$$U_{g|T_L=0}(t) = \begin{cases} at^2 - bt^3 & 0 \leq t \leq O_q T_0 \\ 0 & O_q T_0 \leq t \leq T_0 \end{cases} \tag{3.12}$$

$$U'_{g|T_L=0}(t) = \begin{cases} 2at - 3bt^2 & 0 \leq t \leq O_q T_0 \\ 0 & O_q T_0 \leq t \leq T_0 \end{cases} \tag{3.13}$$

$$avec \begin{cases} a = \frac{27 A_V}{4 O_q^2 T_0} \\ b = \frac{27 A_V}{4 O_q^3 T_0^2} \end{cases} \tag{3.14}$$

La condition $T_L = 0$ correspond pour l'ODGD à un passage d'une amplitude négative maximale (i.e. $-E$, figure 3.2) directement à 0 à l'instant de fermeture glottique. Physiologiquement, cela correspond à une fermeture abrupte des plis vocaux. Dans le cas où $T_L \neq 0$, alors $U_{g|T_L=0}(t)$ est filtré par un filtre passe-bas du premier ordre, tel que l'atténuation à 3000 Hz soit égale à T_L dB.

Modèle R++

R. Veldhuis a proposé une autre amélioration du modèle Rosenberg-B (Veldhuis, 1996), appelé R++, en ajoutant un coefficient d'asymétrie et un paramètre de phase retour.

Ce modèle possède ainsi cinq paramètres :
- K : coefficient d'amplitude
- T_0 : période fondamentale
- T_e : minimum de la dérivée de l'onde de débit glottique (instant d'excitation)
- T_p : maximum de l'onde de débit glottique
- T_a : constante de temps de la phase retour

Contrairement aux précédents modèles, ce n'est plus ici l'ODG qui est formalisée mais l'ODGD. Cette ODGD contient une partie polynômiale du $3^{ème}$ ordre jusqu'à l'instant T_e, puis une phase de

retour exponentielle de T_e à T_0.

$$U_g'(t) = \begin{cases} 4Kt(T_p - t)(T_x - t) & 0 \leq t \leq T_e \\ U_g'(T_e) \times \frac{e^{-(t-T_e)/T_a} - e^{-(T_0-T_e)/T_a}}{1-e^{-(T_0-T-e)/T_a}} & T_e \leq t \leq T_0 \end{cases} \tag{3.15}$$

où :

$$T_x = T_e\left(1 - \frac{\frac{1}{2}T_e^2 - T_eT_p}{2T_e^2 - 3T_eT_p + 6T_a(T_e - T_p)D(T_0, T_e, T_a)}\right) \tag{3.16}$$

et

$$D(T_0, T_e, T_a) = 1 - \frac{(T_0 - T_e)/T_a}{e^{(T_0-T_e)/T_a} - 1} \tag{3.17}$$

Alors, pour ce modèle, l'expression de l'ODG $U_g(t)$ est obtenue par intégration, sous la forme :

$$U_g(t) = \begin{cases} Kt^2(t^2 - \frac{4}{3}t(T - p + T_x) + 2T_pT - x) & 0 \leq t \leq T_e \\ U_g(T_e) + T_aU_g'(T_e) \times \frac{1-e^{-(t-T_e)/T_a} - ((t-T_e)/T_a)e^{-(T_0-T_e)/T_a}}{1-e^{-(T_0-T-e)/T_a}} & T_e \leq t \leq T_0 \end{cases} \tag{3.18}$$

Modèle LF (Liljencrants-Fant)

Le modèle LF (Fant et al., 1985) résulte en réalité de la combinaison de deux modèles différents : le modèle "L" de J. Liljencrants et le modèle "F" de G. Fant. Le but principal du modèle LF, selon ses auteurs, était de pouvoir s'adapter globalement à n'importe quelle forme d'onde usuelle grâce à un jeu de paramètres restreint et de permettre de synthétiser les types de phonations extrêmes. En outre, son implémentation numérique était également facilitée. Pour une comparaison plus détaillée des deux modèles L et F, le lecteur intéressé pourra se reporter à (Fant et al., 1985). Ces deux modèles, dans leurs versions originales, traduisaient une fermeture abrupte des plis vocaux (i.e. avec une phase de retour nulle) comme pour le modèle KLGLOTT88, précédemment décrit dans la section 3.1.2. Le modèle LF visait justement à améliorer cette modélisation, pour des cas où la fermeture de la glotte ne serait pas complète ou plus graduelle. Pareillement au modèle R++, le modèle LF représente l'onde de débit glottique dérivée (ODGD) et non l'ODG[5]. Cela signifie que le modèle LF fait implicitement référence au modèle source-filtre de la production vocale et n'a pas vocation, en premier lieu, à décrire les interactions de la source avec le conduit vocal. L'ODGD du modèle LF est représentée sur la Figure 3.4

5. N.B. : le modèle LF est évidemment antérieur au modèle R++

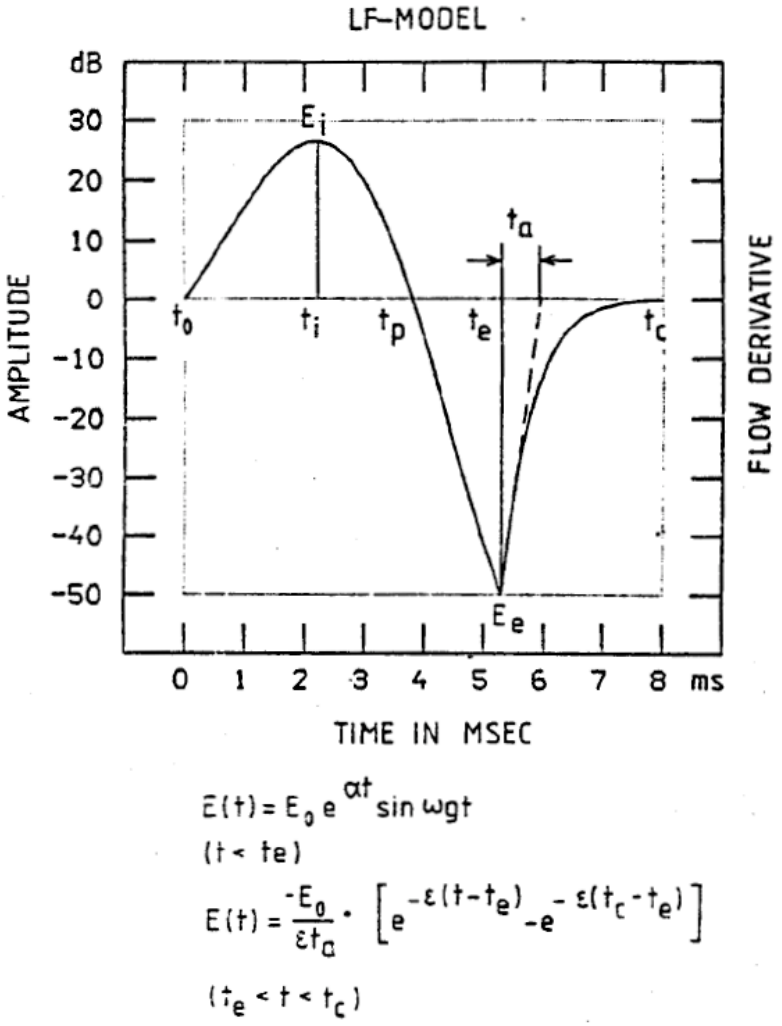


FIGURE 3.4 – *Forme et expression de l'onde de débit glottique dérivée du modèle LF (d'après Fant (Fant et al., 1985))*

Ce modèle comporte quatre paramètres :

- E_e : l'amplitude du minimum de l'onde de débit glottique dérivée (i.e. maximum négatif)
- T_e : l'instant d'excitation maximale
- T_p : l'instant du maximum de l'onde de débit glottique
- T_a : la constante de temps de la phase de retour

La période fondamentale T_0 est ensuite déterminée de manière unique par ces quatre paramètres. Par rapport à la Figure 3.4, l'équation suivante 3.20 tient en outre compte d'une part de la normalisation de la forme d'onde par l'amplitude E_i et la période fondamentale T_c (ou T_0). De façon à se conformer à la réalité physique de la production de l'ODGD, une contrainte est apportée au gain sur

une période. Celle-ci s'exprime sous la forme :

$$\int_0^{T_0} U_g^{'}(t) = 0 \tag{3.19}$$

Cette contrainte nous assure que le signal revient bien à zéro à la fin de chaque période. En d'autres termes, on s'assure que l'aire positive de la courbe est égale à l'aire négative, évitant ainsi qu'un gain ne se cumule à chaque nouvelle période fondamentale.

Ce modèle est constitué d'une partie sinusoïdale modulée par une exponentielle croissante jusqu'à T_e puis décroissante pour la phase de retour, entre T_e et T_0 :

$$U_g^{'}(t) = \begin{cases} -E_e e^{a(t-T_e)} \frac{\sin(\pi t/T_p)}{\sin(\pi T_e/T_p)} & 0 \leq t \leq T_e \\ -\frac{E_e}{\epsilon T_a}(e^{-\epsilon(t-T_e)} - e^{-\epsilon(T_0-T_e)}) & T_e \leq t \leq T_0 \end{cases} \tag{3.20}$$

Le paramètre ϵ étant défini par l'équation implicite suivante :

$$\epsilon T_a = 1 - e^{T_0 - T_e} \tag{3.21}$$

Le paramètre a est quant à lui défini par une autre équation implicite, sous la forme :

$$\frac{1}{a^2 + (\frac{\pi}{T_p})^2}\left(e^{-aT_e}\frac{\pi/T_p}{\sin(\pi T_e/T_p)} + a - \frac{\pi}{T_p}\text{cotan}(\pi T_e/T_p)\right) = \frac{T_0 - T_e}{e^{\epsilon(T_0-T_e)} - 1} - \frac{1}{\epsilon} \tag{3.22}$$

Enfin, pour obtenir l'ODG, il suffit d'intégrer l'expression précédente, ce qui nous donne :

$$U_g(t) = \begin{cases} -\frac{E_e e^{-aT_e)}}{\sin(\pi T_e/T_p)} \frac{1}{a^2+(\frac{\pi}{T_p})^2}(\frac{\pi}{T_p} + ae^{at}\sin(\pi t/T_p) - \frac{\pi}{T_p}e^{at}\cos(\pi t/T_p)) & 0 \leq t \leq T_e \\ -E_e(\frac{1}{\epsilon T_a} - 1)(T_0 - t + \frac{1}{\epsilon}(1 - e^{-\epsilon(T_0-t)}) & T_e \leq t \leq T_0 \end{cases} \tag{3.23}$$

Modèle générique

Selon Doval et al. (Doval et al., 2006), l'ensemble de ces modèles peuvent être exprimés grâce à un jeu réduit de cinq paramètres, à savoir :

- E : le maximum d'excitation
- T_0 : la période fondamentale
- O_q : le quotient ouvert
- α_m : le coefficient d'asymétrie
- Q_a : le quotient de phase retour

Ils ont ainsi pu montrer que dans le cas d'une fermeture abrupte, l'ODG $U_g(t)$ et sa dérivée peuvent toujours être formulées de la manière suivante, quelque soit le modèle :

$$U_g(t;P) = EO_q T_0 n_g(\frac{t}{O_Q T_0};\alpha_m) \tag{3.24}$$

$$U_g^{'}(t;P) = En_g^{'}(\frac{t}{O_Q T_0};\alpha_m) \tag{3.25}$$

pour $0 \leq t \leq T_0$ et avec $P = \{E, T_0, O_q, \alpha_m, Q_a\}$.

Cette formalisation globale des différents modèles historiques d'ODG et d'ODGD leur ont permis d'aboutir à un nouveau modèle de source glottique, le modèle CALM, dont la manipulation se trouve simplifié, comme nous allons le voir dans la section 3.1.3 suivante.

En épilogue de cette première section 3.1.2 sur les différentes modèles, on peut observer sur la figure 3.5 la comparaison dans le domaine temporel de l'ODG (en haut) et de l'ODGD pour les 4 différents modèles présentés. Sur cette figure, la fermeture est abrupte ($T_L = 0$) et E est conservé constant pour tous les modèles. On observe sur cette figure notamment, que le modèle LF possède un comportement médian, en se situant au centre des autres modèles. Notons également, qu'en gardant une amplitude négative maximale constante pour l'ODGD, l'amplitude maximale de l'ODG est quant à elle modifiée.

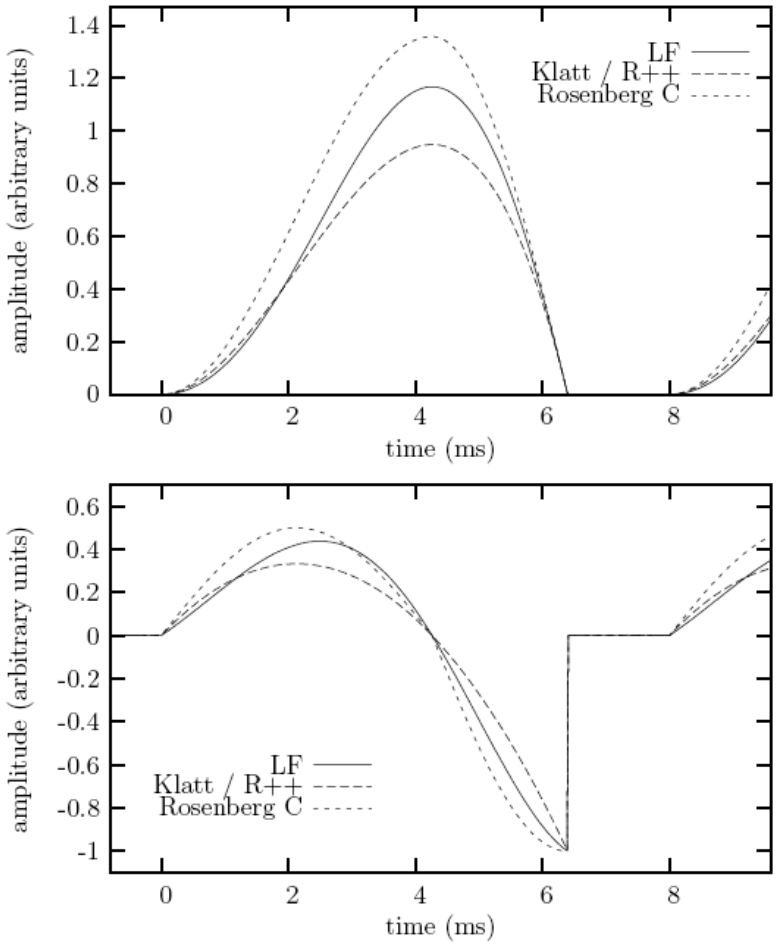


FIGURE 3.5 – *Comparaison des quatre modèles de source glottique, pour un même jeu de paramètres, lors d'une fermeture abrupte et pour* E *constant (d'après (Doval et al., 2003))*

3.1.3 Le Modèle Linéaire Causal/Anticausal (CALM)

Le modèle linéaire causal/anticausal a été développé par Doval, d'Alessandro et Henrich (Doval et al., 2003), en se basant sur une modélisation spectrale de l'onde de débit glottique, plutôt que temporelle. La modélisation du conduit vocal dans le domaine spectral (avec les fréquences centrales des filtres résonants, leurs amplitudes et bandes passantes) est très efficace en terme de manipulation car la description spectrale des sons est plus adaptée du point de vue de la perception auditive. Traditionnellement, et comme le montre les différentes définitions de la section précédente, l'onde de débit glottique est exprimée dans le domaine temporel. Une approche spectrale peut être considérée comme équivalente à condition de prendre en compte à la fois le spectre en amplitude et en phase.

Pour le spectre d'amplitude, deux effets différents peuvent être isolés (cf. figure 3.6). D'une part, une certaine quantité d'énergie est concentrée en basses fréquences (i.e. au dessous de 3 kHz). Ce pic est généralement appelé *formant glottique*. Lors de variations de qualité vocale on peut observer que la bande passante, l'amplitude et la position du formant glottique sont modifiées. D'autre part, la variation de la pente spectrale en hautes fréquences (appelé *tilt spectral* ou *inclinaison spectrale* et représenté sur la figure 3.6 par une modification de la valeur de la pente pour les fréquences supérieures à $\mathrm{F_c}$) est également liée aux modifications de qualité vocale, principalement en ce qui concerne l'effort vocal, comme nous le verrons par la suite.

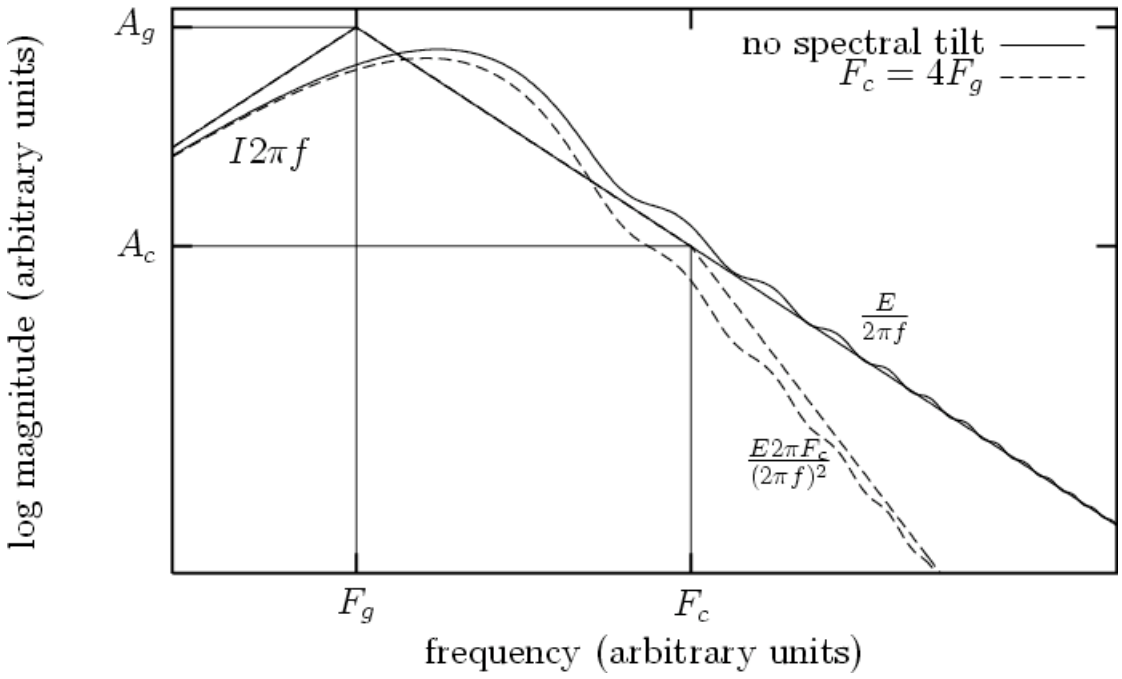


FIGURE 3.6 – *Spectre en amplitude de l'onde de débit glottique dérivée : illustration du formant glottique ($\mathrm{F_g}$, $\mathrm{A_g}$) et du tilt spectral ($\mathrm{F_\alpha}$, $\mathrm{A_\alpha}$) (d'après (Doval et al., 2003))*

En considérant à la fois les effets du formant glottique et du tilt spectral, l'ODG peut être implémentée numériquement de manière simplifiée grâce à deux filtres en cascade. Un filtre passe-bas résonant du second ordre ($\mathrm{H_1}$) pour le formant glottique, et un filtre passe-bas du premier ordre ($\mathrm{H_2}$) pour le tilt spectral.

Cependant, l'information de phase nous indique que ce système n'est pas totalement causal. En réalité, comme il est montré sur la figure 3.7 pour l'ODGD, l'ODG est la combinaison d'une partie montante (ou active) et d'une partie descendante (ou passive). La partie descendante, ou phase retour, influence principalement le tilt spectral et constitue par conséquent une partie causale. Or l'instant auquel débute cette phase retour n'est autre que le GCI, l'instant de fermeture glottique, à partir duquel est calculé le modèle. Et il est alors possible de montrer que le filtre passe-bas du

second ordre doit être anticausal de manière à fournir une bonne représentation de phase. Cette information est parfois appelée représentation de phase mixte de la production vocale (Bozkurt, 2004).

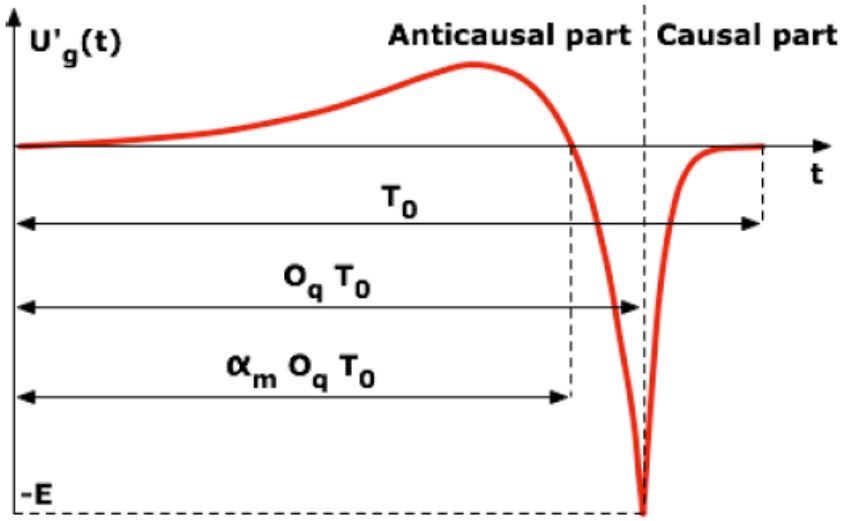


FIGURE 3.7 – *Représentation dans le domaine temporel de l'onde de débit glottique dérivée : partie anticausale et partie causale. (d'après (D'Alessandro et al., 2006))*

Une étude complète des caractéristiques spectrales de l'onde de débit glottique, détaillée dans (Doval et al., 2003), nous fournit les équations liant les paramètres pertinents de la pulsation glottique (F_0 : fréquence fondamentale, O_q : quotient ouvert, α_m : coefficient d'asymétrie et T_l : tilt spectral, en dB à 3000 Hz) aux coefficients de H_1 et H_2. Notons que l'expression de b_1 (corrigé par rapport à celle présente dans (Doval et al., 2003) et (d'Alessandro et al., 2005b)) contient également des équations liant les paramètres temporels aux paramètres spectraux, exprimant les implémentations des deux filtres.

La transformée en z du filtre résonant anti-causal du second ordre, s'exprime alors sous la forme :

$$H_1(z) = \frac{b_1.z}{1 + a_1.z + a_2.z^2} \tag{3.26}$$

où :

$$a_1 = -2e^{a_p T_e} \cos(b_p T_e),\ a_2 = e^{2a_p T_e} \tag{3.27}$$

$$b_1 = ET_e \tag{3.28}$$

$$a_p = -\frac{\pi}{O_q T_o \tan(\pi\alpha_m)},\ b_p = \frac{\pi}{O_q T_0} \tag{3.29}$$

Et la transformée en z du filtre causal du premier ordre est telle que :

$$H_2(z) = \frac{b_{T_l}}{1 - a_{t_l} z^{-1}} \tag{3.30}$$

où :

$$a_{T_l} = \nu - \sqrt{\nu^2 - 1},\ b_{T_l} = 1 - a_{T_l} \tag{3.31}$$

$$\nu = 1 - \frac{1}{\mu}, \mu = \frac{\frac{1}{e^{-T_l/10 \ln 10}} - 1}{\cos(2\pi \frac{3000}{F_e}) - 1} \tag{3.32}$$

La source vocale dans le domaine spectral peut être ainsi considérée simplement comme un système passe-bas. Cela signifie que l'énergie de la source vocale est principalement concentrée dans les basses fréquences et décroît rapidement lorsque la fréquence augmente.

L'inclinaison spectrale, dans le spectre de parole rayonnée (qui est principalement visible sur l'ODGD) est au plus de $-6\ \mathrm{dB/octave}$ en hautes fréquences. Comme cette pente est de $+6\ \mathrm{dB/octave}$ à la fréquence nulle, la forme globale du spectre est celle d'un large pic spectral. A ce pic, figure un maximum, assez similaire dans sa forme aux pics de résonance du conduit vocal (mais différent par sa nature). Ce *formant* se remarque souvent sur les spectrogrammes, où il est également appelé *barre de voisement*, en-dessous du premier formant du conduit vocal.

Les propriétés spectrales de la source peuvent être étudiées en termes de propriétés de ce formant glottique. Ces caractéristiques sont :

1. Sa position ou *fréquence centrale*
2. Sa largeur ou *bande passante*
3. Sa pente en hautes fréquences ou *tilt spectral*
4. Sa hauteur ou *amplitude*

On peut montrer que la fréquence du formant glottique est inversement proportionnelle au quotient ouvert O_q (Doval et al., 2006). Cela signifie que le formant glottique est bas pour une voix relâchée, correspondant à un quotient ouvert élevé. Inversement, une voix tendue possède un formant glottique élevé, et le quotient ouvert est alors faible.

L'amplitude du formant glottique est directement proportionnelle à l'amplitude de voisement. La largeur du formant glottique est liée à l'asymétrie de la forme d'onde glottique. La relation n'est toutefois pas triviale, mais l'on peut considérer qu'une forme d'onde symétrique (S_q faible) engendre un formant glottique plus étroit et légèrement plus bas. Inversement, une asymétrie plus prononcée engendre un formant glottique plus large et plus élevé.

Autour de la valeur standard du coefficient d'asymétrie ($\alpha_m = 2/3$) et pour des valeurs normales de O_q (entre 0.5 et 1), le formant glottique est situé légèrement au-dessous ou proche du premier harmonique ($F_g \approx H_1$). En revanche, en prenant des valeurs extrêmes telles que $Oq = 0.4$ et $\alpha_m = 0.9$, F_g peut alors atteindre le quatrième harmonique.

Jusqu'à maintenant, nous avons considéré une fermeture abrupte des cordes vocales. Une fermeture adoucie des cordes vocales est obtenue grâce à un Q_a positif dans le domaine temporel. Dans le domaine spectral, l'effet d'une fermeture adoucie correspond à une augmentation de l'atténuation spectrale. La position fréquentielle où cette atténuation additionnelle commence est alors inversement proportionnelle à Q_a. Pour un Q_a faible, l'atténuation affecte uniquement les hautes fréquences, car le point correspondant dans le spectre est élevé. Pour un Q_a élevé, cette atténuation modifie les fréquences débutant en un point plus bas du spectre.

En définitive, l'enveloppe spectrale des modèles de débit glottique peut être considérés comme un gain sur un filtre passe-bas. L'enveloppe spectrale de la dérivée peut être vue comme le gain d'un filtre passe-bande. Le spectre de la source peut être stylisé par trois segments linéaires ayant des pentes respectives de $+6\ \mathrm{dB/octave}$, $-6\ \mathrm{dB/octave}$, $-12\ \mathrm{dB/octave}$ (voire parfois $-18\ \mathrm{dB/octave}$) comme représentée sur la figure 3.6. Les deux points de cassure dans le spectre correspondent respectivement au pic du formant glottique et à la fréquence de coupure du tilt spectral.

3.2 Phonétique de la qualité vocale

Arrivé à ce point de notre argumentation, le but de cette section est de démontrer la pertinence de la modélisation de la source vocale pour l'étude de la synthèse de voix expressive. Et cela parce que la *qualité vocale*, notion que l'on peut réduire, dans une première approche, aux effets de la configuration laryngée sur la production vocale, réside au coeur même de l'expressivité, bien qu'elle soit souvent laissée pour compte, voire occultée dans bon nombre d'études sur la prosodie (tendance qui tend cependant à se réduire ces dernières années).

La notion de qualité vocale est liée aux différentes configurations laryngées permettant ou donnant naissance à des types de voix telles que des voix soufflées, tendues, rauques ... Il paraît ainsi assez clair que, de la même façon que l'on avait précédemment soutenu l'idée que la prosodie représentait le support, ou le vecteur, au sein duquel l'expressivité était exprimée, en sa qualité de dimension prosodique (Campbell and Mokhtari, 2003; Pfitzinger, 2006), la qualité vocale joue un rôle à part entière dans l'expressivité de la production vocale.

Mais avant de décrire les possibilités offertes par la modélisation de source glottique pour la synthèse de voix expressive, nous allons réaliser un petit détour par la phonétique de la qualité vocale, afin de mieux comprendre le fonctionnement et les enjeux relatifs à la production expressive. Souvenons-nous que la notion de qualité vocale est issue principalement de trois domaines de recherche différents : la phoniatrie (l'étude et le traitement des voix pathologiques), la voix chantée (qu'est-ce que *bien chanter* ?) et plus récemment l'analyse et la synthèse vocale (afin de mieux comprendre les mécanismes extrêmes ou standards).

C'est sans doute la phoniatrie qui introduisit en premier le terme de qualité vocale, en faisant ainsi référence à une qualité de voix *normale* par opposition à une qualité de voix *pathologique*, comme la diplophonie (i.e. présence de deux fréquences fondamentales distinctes) pour ne citer qu'un exemple. Cela dit, au fur et à mesure de l'élaboration d'une catégorisation des voix pathologies, l'on s'est rendu compte que certaines caractéristiques habituellement prêtées aux voix pathologiques étaient également présentes, dans une moindre mesure, pour les voix dites "normales". Pour reprendre l'exemple précédent, la diplophonie peut être considérée comme un cas extrême de jitter.

Avec l'évolution de la puissance de calcul des ordinateurs s'est accompagnée la possibilité de synthétiser des occurences vocales, permettant une comparaison avec des échantillons de voix naturelle, pathologique ou non. Ces comparaisons sont principalement effectuées grâce à des tests d'écoute perceptifs, soit pour catégoriser grâce à des écoutes expertes les différents types de pathologies (Grade of hoarseness, Rough, Breathy, Asthenic, Strained (De Bodt et al., 1997)), soit en fournissant la possibilité au sujet de modifier certains paramètres du synthétiseur afin de *coller* au plus proche d'un stimulus précédemment entendu.

En tout cas, il est nécessaire de retenir deux faits importants concernant ce point, corroborés par les éminents travaux de J. Kreiman et B. Gerratt : (i) tout d'abord, la possibilité fournie aux utilisateurs de pouvoir modifier de façon continue, et non plus sur une échelle discrète, les paramètres du synthétiseur (Gerratt and Kreiman, 2001; Kreiman and Gerratt, 1998) offre de bien meilleurs résultats en termes de fiabilité (intra-individuelle et inter-individuelle) ; (ii) ensuite, pour étendre ce terme de fiabilité, il faut ici bien comprendre, justement, que la qualité vocale est par essence relative à la perception que nous avons de telle ou telle voix (Kreiman et al., 1993). Il n'existe donc pas de bijection entre une certaine configuration phonétique et la qualité vocale perçue de cette configuration. Autrement dit, même s'il existe certaines constantes phonétiques, une qualité vocale donnée ne sera pas *nécessairement* produite de la même manière pour deux locuteurs différents voire même pour un même locuteur dans deux situations (contextes) différents.

Cela dit, l'utilisation de la synthèse de source glottique nous permet, avec un jeu de paramètres réduit donné, de produire de manière constante une même occurence en sortie du synthétiseur (moyennant les effets du contexte : taille de la pièce d'écoute, réverbération ...). Ainsi, grâce aux tests d'écoute précités, il est alors possible de catégoriser, avec une plus grande fiabilité, différentes qualités vocales données. On ne cherche pas ici à reproduire *à l'identique* la voix naturelle, puisque l'on utilise une modélisation, mais à obtenir une production la plus *transparente* possible (i.e. dont le seuil différentiel [6] soit suffisamment faible).

Concernant l'étude de la prosodie et de l'expressivité de la voix, les paramètres des modèles de source glottique peuvent être regroupés selon quatre dimensions perceptives principales relativement indépendantes (d'Alessandro, 2006), à savoir :

1. La notion de registre ou mécanisme laryngé
2. L'effort vocal
3. Les apériodicités : souffle et âpreté
4. La dimension tendue/relâchée

3.2.1 La notion de registre vocal

C'est donc selon ces quatre dimensions, que nous allons décrire leurs caractéristiques phonétiques afférentes, pour ensuite donner une description des relations existant entre ces dimensions et les paramètres de modélisation de la source glottique. C. d'Alessandro (d'Alessandro, 2006) nous donne quelques exemple de l'importance de ces dimensions relativement à l'expressivité de la voix.

6. Le seuil différentiel, ou Difference Limen en anglais, également appelé Just Noticeable Difference (JND), caractérise la différence la plus faible perceptible par l'oreille humaine pour la variation d'un paramètre de synthèse pris isolément. Pour comprendre ce dont il est question ici, on peut s'amuser à synthétiser deux fréquences pures très proches, dont on fait varier progressivement l'une d'elles. On perçoit alors *alternativement* un son composé (fusion) ou deux fréquences distinctes. Le seuil différentiel représente alors la différence fréquentielle à partir de laquelle la fusion disparaît, et qui peut être différente pour chaque auditeur.

Il nous dit notamment que les différents registres peuvent être utilisés dans un but stylistique lors de tâches expressives. Le chuchotement, caractéristique de la proximité ou de l'intimité, peut être considéré comme un marqueur de séduction. La dimensions tendue/ralâchée sert également au sein des langues à tons, comme un marqueur distinctif tonal. Enfin, l'effort vocal est une dimension importante pour signaler l'accentuation, et sera à ce titre souvent corrélé avec la hauteur de la voix.

Dans le domaine de la voix chantée, la notion de voix de poitrine et de voix de tête, souvent utilisée dans l'étude et la pratique du chant, bien que liée indubitablement à la qualité vocale, ne constitue pas à proprement parler une configuration phonétique pertinente pour l'étude de la qualité vocale. En effet, ces termes possèdent presque autant d'acceptions que d'auteurs ayant étudié ces registres. Et aucune ne permet d'associer clairement tel ou tel registre à un mode particulier de phonation du larynx.

En définitive, la qualité vocale se réfère plus justement dans le domaine de la phonétique aux modes de phonation adoptés par les cordes vocales permettant de produire un certain timbre de voix. Ainsi, Ladefoged (Ladefoged, 1971) dénombre pas moins de neuf modes de vibration des cordes vocales, eux-mêmes non exhaustifs et se recoupant parfois. Une notion importante relative aux modes de phonation est celle de "voix neutre", car afin de définir des timbres de voix particuliers, il convient de s'accorder sur la nature d'une voix supposée normale ou standard. Les phonéticiens s'entendent ici pour accepter le terme de voix modale, définie par Hollien (Hollien, 1974) comme étant le type de phonation neutre. Mais ici encore il s'agit plus d'une description que d'une définition au sens strict du terme. Laver (Laver, 1980) parle ainsi de la voix modale comme d'un type de vibration des cordes vocales dont la théorie phonétique assume qu'elle a lieu lors d'un voisement ordinaire, quand aucune caractéristique particulière n'est changée ou ajoutée.

Classification des modes de phonation

Dans le but de désambiguiser les confusions liées au modes phonatoires et aux registres utilisés dans la terminologie du domaine de la voix chantée, notamment, nous reportons ici, en premier lieu la classification réalisée récemment par B. Roubeau et al. (Roubeau et al., 2009). Cette classification fait référence à quatre mécanismes laryngées ($\mathrm{M0}, \mathrm{M1}, \mathrm{M2}, \mathrm{M3}$) liés aux différentes configurations *mécaniques* du larynx. Cette classification repose sur le fait que lors d'une accélération de la vibration des cordes vocales, du plus grave au plus aiguë, le système biomécanique constitué par les muscles, les tissus et les os du larynx, adopte différentes configurations non linéaires, modifiant de manière *quantique* les formes possibles de l'onde de débit glottique. L'étude de B. Roubeau (Roubeau et al., 2009) se fonde essentiellement sur des mesures électroglottographiques, et le tableau 3.1 suivant répertorie de manière exhaustive les différents termes présents dans la littérature pour qualifier, certains modes phonatoires ou registres.

Mécanisme $M0$	**Mécanisme** $M1$	**Mécanisme** $M2$	**Mécanisme** $M3$
Friture (fry)	Modale	Voix de fausset	Voix de sifflet (whistle)
Impulsion (pulse)	Normale	Void de tête (F)	Flageolet
Strohbass	Voix de poitrine	Voix de loft (loft voice)	Flûte
Voix de contrebasse	Lourde	Légère	
	Epaisse	Etroite	
	Voix mixte (H - mixed)	Voix mixte (F - mixed)	
	Voce finta (H)		
	Voix de tête d'opéra (H)		

TABLE 3.1 – Classification des modes phonatoires, selon les mécanismes laryngées, pour les hommes (H) et les femmes (F) (d'après (Roubeau et al., 2009))

Voix modale ($M1$**)**

La phonation modale se réfère à la phonation typique ou basique, et inclut l'étendue de fréquences fondamentales utilisées pour la voix parlée (Hollien, 1974). Cette phonation est également associée à une vibration périodique des cordes vocales, une fermeture glottique complète avec un spectre glottique riche. La voix modale, ne constitue pas le mode de phonation le plus bas (i.e. le plus grave), mais nous l'introduisons ici en premier, car c'est le mode de phonation à partir duquel sont définis les autres modes de phonation, ainsi que les apériodicités.

H. Hollien (Hollien, 1974) commente ainsi son choix à propos du terme "modal" pour ce type de phonation, tout en montrant son scepticisme quant à la taxonomie utilisée en musique vocale :

> Le registre *modal* est un terme que j'utilise depuis quelques années. Originellement, je préférais le terme "normal" pour identifier ce registre. Cependant, comme le faisait remarquer van den Berg (lors d'un échange personnel) l'utilisation du label "normal" impliquerait que les autres registres soient anormaux et, évidemment, sa logique est correcte. En conséquence, le registre modal s'appelle ainsi car il comprend l'intervalle des fréquences fondamentales utilisées couramment en voix parlée et chantée (i.e. le mode). C'est un terme plutôt inclusif et de nombreuse personnes - en particulier celles travaillant dans la musique vocale - contesterait le fait que cette entité constitue un jeu de registres et de sous-registres incluant deux (poitrine et tête), voire trois (bas, moyen, haut) entités différentes. Je comprend la tradition d'une telle approche , mais ... je dois encore trouver une preuve raisonnablement convaincante que de tels sous-registres existent véritablement.

Dans un souci de clarté, nous nommerons donc indifféremment, dans la suite de ce manuscrit, par voix modale ou $M1$ ce que nous appelions par commodité voix normale ou neutre précédemment. Il convient également d'ajouter à la description pertinente de Hollien, le fait qu'habituellement la voix modale est obtenue lors d'une phonation sans effort particulier, à une hauteur moyenne pour le locuteur en question. La voyelle [e] (ou le schwa) est généralement privilégiée pour cette

description, car celle-ci correspond à une configuration neutre du conduit vocal, c'est à dire sans ajout de cavités résonantes supplémentaires que celles présentes pour un tube de cette longueur (entre 15 et 17 cm approximativement). On observe alors traditionnellement un premier formant autour de 500 Hz et des formants d'ordres supérieurs espacés de 1000 Hz environ. Cependant, certaines études sur la qualité vocale, privilégient l'utilisation de stimuli avec la voyelle [a]. Du point de vue aéro-acoustique, il a été montré que le conduit vocal engendrait une interaction avec le larynx, notamment par l'apparition d'ondes réfléchies revenant vers la source vocale. Ces interactions bien qu'importante d'un point de vue théorique, n'ont qu'une influence négligeable sur l'ODG en première approche [7]. Il est toujours possible ensuite d'ajouter sous la forme de règles issues d'analyses acoustiques une interaction entre les harmoniques de la source et les formants du conduit vocal. L'étude menée dans le présent manuscrit ne traitera cependant la notion de qualité vocale que selon les différentes aspects de phonation de la source glottique.

Sur la figure 3.8, est représentée la configuration classique des cordes vocales, et de l'onde de débit glottique correspondante pour la voix modale. On peut observer sur la figure du haut, que les cordes vocales sont accolées en leurs extrémités, et que la glotte s'ouvre lorsque la pression sub-glottique devient suffisamment importante, pour venir se refermer une fois que les pressions sub-glottique et supra-glottique se sont équilibrées. La figure du bas représente ainsi l'évolution temporelle du débit glottique typique pour la voix modale.

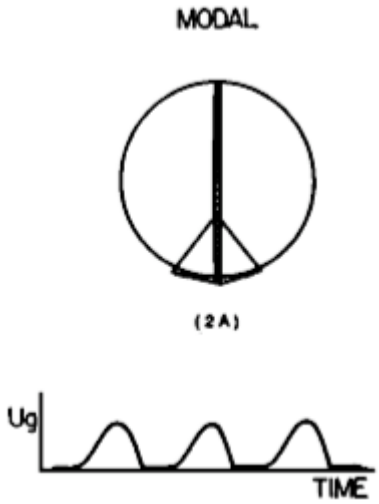


FIGURE 3.8 – *Configuration typique pour la voix modale, de la glotte (en haut) et, de l'ODG (en bas), d'après (Klatt and Klatt, 1990)*

Le type de phonation modale trouve également sa justification dans la description des plis vocaux en tant que système physique. Prenons, pour illustrer notre propos, l'exemple du modèle à

7. et sont d'ailleurs occultées selon l'approche du modèle linéaire source-filtre

deux masses des cordes vocales (Ishizaka and Flanagan, 1972), constituant le premier modèle biomécanique des cordes vocales. Notons que des modèles physiques plus évolués existent (Titze and Strong, 1975), mais nous ne rentrerons pas ici dans les détails sachant que la description du modèle à deux masses suffit ici à notre propos. Ce modèle est décrit sur la figure 3.9 suivante.

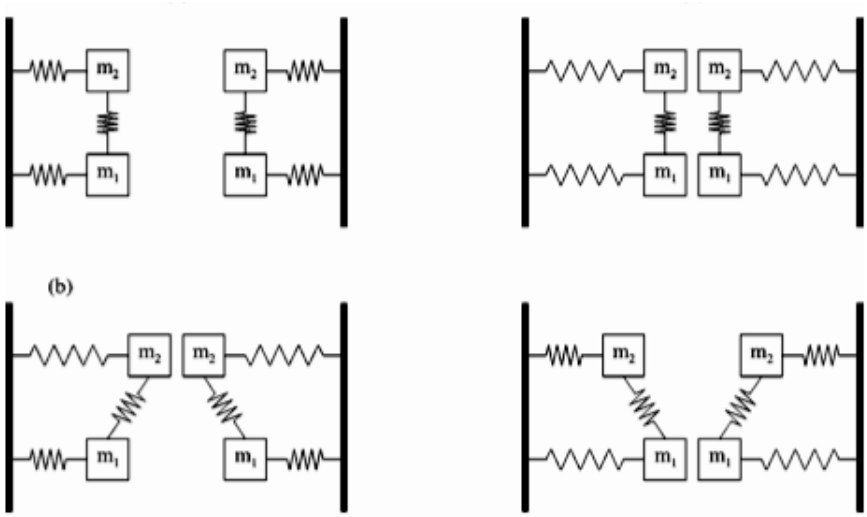


FIGURE 3.9 – *Les deux modes propres possibles du modèle à deux masses, avec en haut un mode propre où les deux masses vibrent en phase et, en bas, un mode où les deux masses sont en quadrature de phase (d'après (Titze and Strong, 1975)).*

En fait, l'un des postulats de base de la théorie vibratoire linéaire est que les modes vibratoires d'un système, tel que celui constitué par les cordes vocales, peuvent être réalisés à partir des mêmes modes propres sous-jacents. Pour le modèle à deux masses, il existe deux modes propres possibles, l'un où les masses vibrent en phase et l'autre où elles sont en quadrature de phase (180˚). Alors, toute vibration des cordes vocales est une combinaison de ces deux modes de vibration. Concernant les cordes humaines, elles contiennent une infinité de degrés de liberté et comportent donc théoriquement une infinité de modes propres. En pratique, seuls un nombre restreint de ces modes sont excités, et il est possible de décrire bon nombre de vibrations complexes des cordes vocales seulement par un quelques modes propres sous-jacents (Berry et al., 1994).

La voix de fausset ($M2$)

Selon Hollien (Hollien, 1974), la voix de fausset et la voix modale sont des "opérations laryngées complètement différentes". Selon Laver (Laver, 1980), il existe un certain consensus concernant la production de la voix de fausset : les cartilages aryténoïdes rapprochent les cordes vocales, les muscles vocaux le long de chacune des cordes vocales restent relâchés, bien que le reste des cordes soit ferme et immobile. La glotte reste légèrement ouverte (i.e. la fermeture ne s'opère pas complètement), avec une pression sub-glottique inférieure à la voix modale. Cette situation conduit

à l'apparition de bruits de friction, pouvant être comparés à ceux d'une voix murmurée (plutôt que soufflée). Laver (Laver, 1980) rapporte les propos de Chiba et Kajiyama à ce sujet :

> Il est apparu que les bords des cordes vocales restent recouverts, de part en part, par de petits morceaux de mucus, ce qui signifie que l'air n'est pas expiré de manière abrupte. (Pour la voix de poitrine, en particulier pour la "voix forte", ces résidus de mucus aux bords des cordes vocales sont éjectés aussitôt que la voix commence).

La voix de fausset est également caractérisée par une échelle moyenne de hauteur assez élevée. Pour un falsetto masculin, Hollien et Michel (Hollien and Michel, 1968) (voir tableau 3.2) reportent un intervalle de 275 à 634 Hz, par comparaison avec 94 à 287 Hz pour une voix modale.

	Fry	**Modal**	**Falsetto**
Hommes			
Etendue (en Hz)	$7-78$	$71-561$	$156-795$
Etendue moyenne (en Hz)	$24-52$	$94-287$	$275-634$
Etendue moyenne (en tons)	6.7	9.7	7.2
Femmes			
Etendue (en Hz)	$2-78$	$122-798$	$210-1729$
Etendue moyenne (en Hz)	$18-46$	$144-538$	$495-1131$
Etendue moyenne (en tons)	8.1	11.4	7.2

TABLE 3.2 – Etendues globales et moyennes en Hz et étendues moyennes en tons pour les modes de phonations modale, fry et falsetto pour un groupe de 12 hommes et 10 femmes (d'après (Hollien and Michel, 1968))

La troisième caractéristique acoustique concerne l'inclinaison spectrale de la forme d'onde glottique qui est bien plus élevée que pour la voix modale, tombant environ à -20 dB/octave (Monsen and Engebretson, 1977). De plus, inversement à l'ODG de la voix modale, la phase la plus abrupte est la phase d'ouverture.

La Raucité (M0)

La raucité est également appelé friture vocale (vocal fry), friture glottique (glottal fry), voix craquée (creaky voice). Catford (Catford, 1977) donne les détails suivants :

> Une fréquence basse (autour de 40 Hz) de vibration périodique d'une faible section des cordes vocales [...] Le mécanisme physiologique précis de la raucité est inconnu, mais seule une très petite partie du ligament glottique, près du bord thyroïde, est impliqué. L'effet acoustique est une série rapide de coups, tel un bâton frotté le long d'un grillage.

Ce qui différencie principalement la voix rauque de la voix rocailleuse est la fréquence fondamentale moyenne : de 34.6 Hz avec un ambitus allant de 24 à 52 Hz pour la voix rauque et de 122.1 Hz avec un ambitus allant de 94 à 287 Hz pour la voix rocailleuse (pour une voix masculine - (Hollien and Michel, 1968), voir tableau 3.2). Ces résultats conduisent à décrire la voix rauque "comme une

registre phonatoire apparaissant à des fréquences inférieures à celles du registre modal". Notons, par ailleurs, que, contrairement aux autres modes phonatoires, la voix rauque posséderait des caractéristiques de hauteur sensiblement identiques, quel que soit le genre considéré, conformément au tableau 3.2. Contrairement aux mécanismes M_1 et M_2 qui possèdent une zone de fréquences communes, pour lesquelles il est possible d'avoir une production vocale selon l'un ou l'autre des mécanismes, les mécanismes M_0 et M_1 sont le plus souvent séparés sur le plan des fréquences, à plus forte raison pour les femmes.

En outre, Hollien et al. (Hollien et al., 1966) décrivent certaines caractéristiques acoustiques, comme suit : (i) les cordes vocales, une fois collées, sont relativement épaisses et apparemment compressées (ii) les cordes ventriculaires (i.e. aussi appelées fausses cordes vocales) sont quelque peu collées, et (iii) les surfaces inférieures des fausses cordes entrent effectivement en contact avec les surfaces supérieures des cordes vocales. Ainsi, une structure inhabituellement épaisse, compacte (mais pas nécessairement tendue) est créée avant le début de la phonation.

Sous ces conditions on pourrait s'attendre à ce que les fausses cordes vibrent en synchronie avec les vraies cordes. Des études récentes suggèrent à ce sujet que les fausses cordes vocales, entrent en vibration lors du chant de gorge (qui serait le pendant de la voix rauque pour le chant), avec une fréquence de vibration deux fois plus faible (Bailly et al., 2007).

Dans des travaux postérieurs, (Hollien et al., 1969) soutiennent l'hypothèse que le contrôle de la fréquence fondamentale lors de la phonation rauque est différent de celui de la voix modale. Là où, pour la voix modale, la longueur des cordes vocales augmente la fréquence fondamentale et leur épaisseur est inversement proportionnelle à la hauteur, pour la voix rauque aucun de ces changements physiologiques (ou myoélastiques) n'a d'effet sur la hauteur, faisant ainsi pencher pour un contrôle aérodynamique (i.e. dépendant de la variation de la pression subglottique).

Du point de vue spectral, on remarque que la voix rauque possède une pente spectrale bien plus raide que les autres modes de phonation. De plus, la voix rauque est hautement apériodique, deux périodes consécutives pouvant passer du simple au double (Monsen and Engebretson, 1977). Hollien & Wendahl (Hollien and Wendahl, 1968) corroborent ce phénomène en décrivant le fry vocal comme "un train de pulsations ou d'excitations discrètes produites par le larynx" séparées par des périodes de non excitation pendant lesquelles la contribution du conduit vocal est très étouffée (de plus de 40 dB (Coleman, 1963)). Des études postérieures ont même pu mettre à jour d'autres variations du fry où les excitations ne sont plus simples, mais doubles, voire triples (Moore and von Leden, 1958; Monsen and Engebretson, 1977; Hollien and Wendahl, 1968).

Un synonyme parfois utilisée pour la voix rauque est voix "laryngée"[8] (Klatt and Klatt, 1990). (Ladefoged, 1971) écrit à ce sujet :

> Un autre mode de vibration des cordes vocales intervient pour les sons laryngés. Lors de ce type de phonation, les cartilages aryténoïdes sont pressés vers l'intérieur de telle façon que les portions postérieures des cordes vocales sont conservées accolées et seules les portions antérieures (ligamentaires) sont capables de vibrer. Le résultat est souvent un son rugueux avec un pitch comparativement bas. Ceci est aussi appelé friture vocale ou voix rauque.

Le tableau 3.2, et la figure 3.10, résument les caractéristiques fréquentielles des trois modes phonatoires principaux $\mathrm{M}0$, $\mathrm{M}1$ et $\mathrm{M}2$. Sur la figure 3.10 de droite, est représenté non pas le spectre de voix en $\mathrm{M}2$, mais le spectre d'une voix dite soufflée. Cependant, le mécanisme $\mathrm{M}2$, comme on a pu le voir précédemment, constitue une configuration laryngée pour laquelle les cordes vocales ne se ferment pas complètement, ce qui engendre la particularité que le flux d'air provenant des poumons génère plus de turbulence, et provoque ainsi un bruit supplémentaire additif, également appelé bruit d'aspiration, caractéristique de la qualité de voix soufflée, et qui se traduit sur le spectre vocal par une structure hautes fréquences inharmonique (i.e. bruitée). Des études montrent, à ce sujet, que les voix de femmes ont tendance à avoir une qualité de voix plus soufflée, liée au fait que les femmes auront plus facilement tendance à utiliser le mécanisme $\mathrm{M}2$.

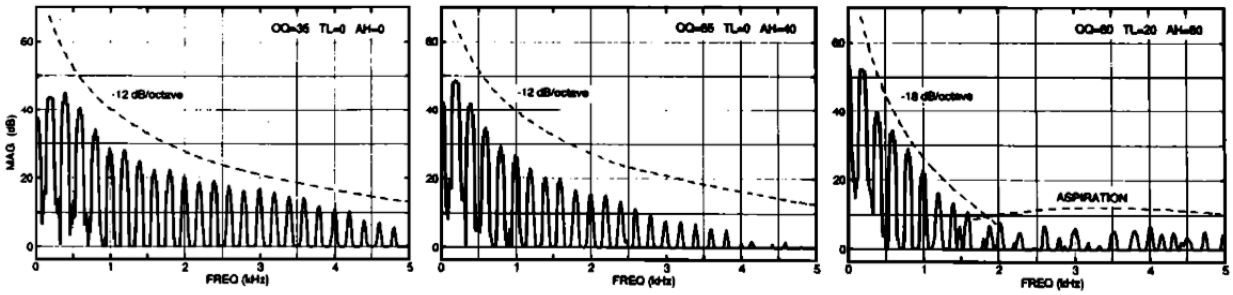


FIGURE 3.10 – *Spectres en amplitude typiques des mécanismes* $\mathrm{M}1, \mathrm{M}2$ *et de la voix soufflée, de gauche à droite (d'après (Klatt and Klatt, 1990))*

On peut observer sur cette dernière figure, pour le passage de M_0 à M_2 en passant par M_1, d'une part une augmentation du quotient ouvert, conjointement à une diminution de l'asymétrie. Ce fait est ici confirmé par l'augmentation du formant glottique, présent dans les basses fréquences. L'évolution de l'inclinaison spectrale de M_0 vers M_2 est celle d'une augmentation, traduit par une diminution progressive des composantes en hautes fréquences, pour être remplacées par le bruit d'aspiration lors de la production de voix soufflée.

8. laryngealization

3.2.2 La dimension de bruit

La présence de bruit dans la production vocale est essentiellement de deux natures : le bruit structurel et le bruit additif. Le bruit structurel est relatif au fait que le système biomécanique formé par le larynx n'est pas une source sonore parfaite. Les différents organes constituant le larynx interagissent de manière naturelle entre eux et sont à l'origine de deux phénomènes apériodiques principaux le jitter et le shimmer, respectivement des variations d'une période à l'autre de la fréquence fondamentale et de l'amplitude. Une voix avec un jitter et un shimmer prononcé sera qualifié de âpre. Le bruit additif est lui simplement lié à la singularité créée au niveau des cordes vocales lors du passage de l'air provenant des poumons, et créant ainsi des turbulences propagées le long du conduit vocal. Ce bruit additif est essentiellement impliqué lors de deux types de qualité vocale différentes : la voix soufflée, lorsque les cordes vocales sont au repos (pas de vibration) et la voix chuchotée, pour laquelle les cordes vocales n'entrent pas en vibration, mais sont accolées, sauf en un lieu restreint laissant passer l'air des poumons de façon réduite.

L'âpreté

L'âpreté ne doit pas être considérée comme un mode de phonation à proprement parler, mais plutôt comme un effet appliqué à la phonation modale, un renforcement de certains paramètres de ce mode de phonation. La caractéristique prédominante de l'âpreté est celle de l'apériodicité de la fréquence fondamentale, autrement appelé jitter, entendue comme une composante de qualité auditive plutôt que de hauteur perçue. Comme le rapporte une étude de Wendahl, pour une fréquence de 100 Hz, une variation période à période aussi faible que 1 Hz (soit 1%) est perçue comme rugueuse. Cependant, cette apériodicité est une notion relative et non absolue par rapport à la hauteur, confirmant le fait que l'impression d'âpreté est plus présente chez les voix masculines que les voix féminines (Hess, 1959). En outre, la durée du stimuli peut jouer un rôle non négligeable et ainsi, un signal bref avec jitter important pourra être jugé moins âpre qu'un signal de longue durée avec un jitter faible. (Coleman and Wendahl, 1967) (p. 128).

Le Chuchotement

La physiologie du mécanisme phonatoire du chuchotement ne prête pas à controverse. Elle peut être décrite schématiquement par une ouverture triangulaire des cartilages, formant ainsi un Y inversé. La taille de ce triangle glottique est inversement proportionnel au volume du chuchotement. Catford (Catford, 1977) rapporte que le spectre acoustique du chuchotement est similaire à celui de la respiration mais avec une concentration bien plus importante de l'énergie acoustique dans les bandes de fréquences formantiques, résultant en un son silencieux relativement *riche*. En voix parlée, le chuchotement peut apparaître dans le processus de dévoisement final de certaines langues.

Le souffle

D. et L. Klatt (Klatt and Klatt, 1990) rapportent un certain nombre de faits concernant la qualité de voix soufflée. Du point de vue physiologique, la qualité de voix soufflée est caractérisée par une augmentation de l'ouverture glottique postérieure. Ce phénomène se traduit par un flux glottique moyen plus important, une forme sinusoïdale plus prononcée, et une phase d'ouverture plus longue. La transition d'une consonne non voisée vers une voyelle contient souvent une courte période de voisement soufflé pendant lequel l'amplitude du premier harmonique est augmentée. La voix soufflée peut donc ainsi être caractérisée par (i) une augmentation du quotient ouvert et (ii) une tendance à remplacer les hautes harmoniques par un bruit d'aspiration. Ces caractéristiques peuvent s'accompagner d'une augmentation de la bande passante du premier formant et/ou l'apparition de pôles et de zéros supplémentaires dans la fonction de transfert du conduit vocal à cause d'une ouverture glottique plus large.

D. Hermes (Hermes, 1991) décrit quant à lui un fait important concernant la synthèse de voyelles soufflées. Afin de pouvoir être correctement fusionné, le bruit additif nécessaire à la production du bruit d'aspiration doit être modulé en amplitude de façon synchrone à la pulsation glottique. Un décalage temporel du bruit additif sera plutôt perçu comme une qualité de voix rauque que soufflée. Si ces conditions ne sont pas remplies, l'auditeur ne parvient généralement pas à fusionner les deux sources quasi-périodique et bruitée, ce qui rend la voix de synthèse trop artificielle.

3.2.3 L'effort vocal

L'effort vocal peut être réduit en première approche au volume sonore, qui sert notamment comme indice prosodique pour l'accentuation. Cependant, si lors d'un effort vocal élevé, la pression sub-glottique augmente, ainsi que l'énergie globale du signal de parole, il ne s'agit pas ici d'une simple *amplification* du signal. Avec l'augmentation de l'effort vocal s'accompagne notamment une énergie en hautes fréquences plus importante. L'origine de cette augmentation est encore mal connue, même si l'on peut observer une tension et une rigidité accrue des cordes vocales.

Nous verrons dans la prochaine section 3.3 que l'effort vocal dépend principalement de deux facteurs non indépendants : le mécanisme laryngé et la fréquence fondamentale. Ces différentes grandeurs sont reliées de manière simple par le phonétogramme.

3.2.4 La dimension tendue/relâchée

Indépendamment de l'effort vocal, la pression exercée par les cordes vocales l'une sur l'autre sur leurs extrémités postérieures, vont définir l'aire de la glotte. Plus la pression de la partie postérieure des cordes vocales sera élevée et plus la voix sera qualifiée de tendue ou pressée.

Inversement, lorsque les cordes vocales sont parfaitement relâchées, correspondant à une vibration non contrainte des cordes vocales, c'est-à-dire vibrant principalement par la pression sub-glottique exercée, alors la qualité de voix est dite relâchée, ou détendue. Pour une voix totalement relâchée, la forme de l'ODG est quasiment sinusoïdale. Pour en revenir à l'expressivité vocale, la voix d'une personne stressée, aura tendance à avoir une qualité de voix plus pressée que d'ordinaire.

On comprend aisément ici que les dimensions de tension et de souffle possèdent des caractéristiques antagonistes. Plus les plis vocaux seront tendus et donc pressées l'un contre l'autre, plus l'aire de la glotte aura tendance à diminuer et donc les turbulences créées lors du passage de l'air également. Inversement, plus la voix sera relâchée, plus la possibilité que des turbulences se créent est important. On peut ainsi, en première approche placer sur un continuum dimensionnel la tension d'un côté et la voix relâchée et/ou soufflée de l'autre.

Cela dit, pour certaines langues particulières, comme le japonais, cette classification n'est pas aussi simple et l'utilisation concurrente d'une voix tendue et soufflée peut apparaître sous certaines conditions. Il convient donc de garder une certaine indépendance de ces deux dimensions en synthèse, et pour les relier éventuellement suivant le phénomène que l'on cherche à observer.

Dans la prochaine section 3.3, nous allons décrire plus précisément l'implémentation de notre synthétiseur qui tient compte des caractéristiques glottiques que nous venons de décrire, à la fois concernant la modélisation de la source glottique, et dans la mesure du possible des différentes configurations phonétiques explicitées afin d'obtenir une synthèse permettant d'adresser la problématique de la qualité vocale.

3.3 Le modèle CALM en temps réel ou RTCALM

Comme nous l'avons décrit dans la section 3.1, le modèle de source glottique CALM est défini par deux filtres simples : un filtre passe-bas du second ordre anti-causal (H_1) et un filtre passe-bas du premier ordre (H_2). Le but de cette section est donc de décrire l'implémentation en temps-réel de ce modèle et plus généralement le logiciel développé pour générer et contrôler un synthétiseur de voix expressive.

3.3.1 Les contraintes

La principale contrainte liée à l'implémentation en temps réel du modèle CALM réside dans le fait que l'un des filtres utilisé est anti-causal. Plusieurs solutions ont été adoptées pour pallier ce problème au nombre desquels nous en retiendrons deux principales.

La première solution consiste à renverser l'axe temporel et ainsi considérer ce filtre comme causal. Il convient alors de sauvegarder le calcul de cette forme d'onde dans un tampon, puis de transmettre un à un les échantillons une fois la période courante calculée. L'inconvénient principal de cette méthode est que l'on ajoute une certaine latence, de l'ordre de la période fondamentale instantanée, mais qui ne se révèle pas rédhibitoire pour notre application.

La seconde solution consiste à considérer le filtre anti-causal stable comme un filtre causal instable et de calculer véritablement la fonction analytique du filtre causal pour générer directement les échantillons à synthétiser. Le point clé de cette méthode est qu'en procédant ainsi le filtrage utilisé est instable. Après le GCI, le filtre diverge et le risque est alors de faire exploser le calcul et saturer la sortie audio. Toutefois, une solution simple à ce problème existe et consiste à stopper le calcul de la partie anti-causale au moment adéquat, c'est-à-dire lorsque l'on est arrivé au GCI.

3.3.2 Les solutions

Solution par renversement temporel

Ce modèle est calculé par filtrage d'un peigne de Dirac par un système causal du second ordre, calculé selon la fréquence et la bande passante du formant glottique, et dont la réponse est inversée temporellement pour obtenir une réponse anti-causale. L'inclinaison spectrale est ensuite introduite par filtrage de cette réponse anti-causale par la composante d'inclinaison spectrale du modèle. La forme d'onde est ensuite normalisée afin de contrôler l'amplitude.

Cette implémentation se situe dans la continuité des tâches de développement réalisées lors de l'atelier eNTERFACE'05 (d'Alessandro et al., 2005b) et du travail présenté lors de NIME'06

(D'Alessandro et al., 2006). Pour cet algorithme, la réponse impulsionnelle est générée par un traitement anti-causal synchrone à la fréquence fondamentale. Ceci signifie qu'afin de construire la forme d'onde souhaitée, la réponse impulsionnelle d'une version causale de H_1 (formant glottique) est calculée, mais stockée à l'envers dans un tampon audio. Cette forme d'onde est ensuite tronquée à une durée correspondant à la fréquence fondamentale instantanée ($\mathrm{F}_0 + \mathrm{Jitter}$).

Cet algorithme est désormais intégré à la fois dans Max/MSP (Max/MSP, 2008) et Pure Data (Puckette, 1996) sous la forme d'un objet externe (pour Max OS X, Windows et Linux) : `almPulse~`. Ensuite la forme d'onde résultante est filtrée par H_2 (inclinaison spectrale). Ce second filtre est également intégré à la fois dans Max/MSP et Pure Data sous la forme d'un objet externe : `stFilter~`. Les coefficients de H_1 et H_2 sont calculés à partir des équations décrites dans la section 3.1.3 d'après (Doval et al., 2003). Ainsi, simultanément, les paramètres dans les domaines temporel et spectral peuvent être récupérés. D'une part, les pulsations glottiques sont dérivées pour produire l'onde de débit glottique dérivée, d'autre part, l'ODG est utilisée pour moduler la quantité de bruit additif. L'algorithme RT-CALM par renversement temporel complet est illustré sur la Figure 3.11.

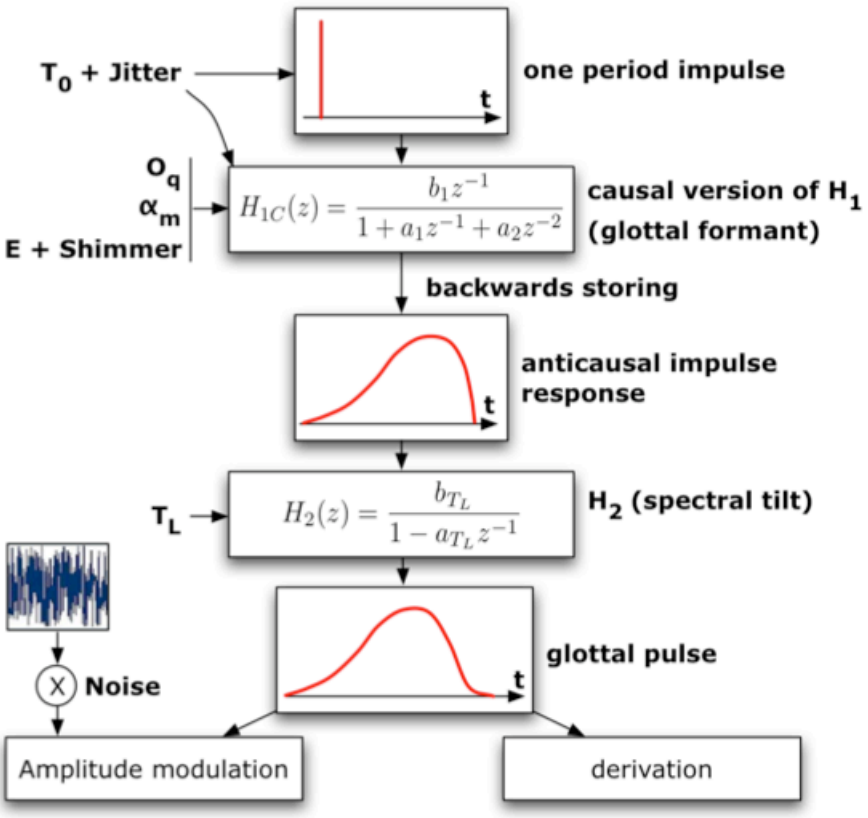


FIGURE 3.11 – *Description schématique du fonctionnement de l'algorithme RT-CALM (d'après (D'Alessandro et al., 2006))*

En réalité, nous tirons parti des propriétés physiques de la glotte pour proposer cet algorithme en temps réel. En effet, la pulsation glottique correspond aux phases d'ouverture et de fermeture des plis vocaux. Cela signifie que les réponses impulsionnelles générées par les filtres H_1 et H_2 ne peuvent pas se superposer. Ainsi, si l'excursion des paramètres est proprement limitée, les réponses impulsionnelles peuvent être stockées à l'envers et tronquées de façon synchrones à la période fondamentale sans modifier outre mesure leurs propriétés spectrales.

La troncature de la forme d'onde du modèle CALM à chaque période fournit de bons résultats de synthèse. Néanmoins, certaines combinaisons de paramètres (par exemple, une valeur élevé de α_m avec une faible valeur de O_q) provoque une oscillation de la réponse impulsionnelle au sein même d'une période, engendrant un signal ne modélisant plus correctement les phénomènes de source glottique et modifiant la perception de qualité vocale. Pour pallier ce problème, des points de troncature anticipée et des options de fenêtrage ont été testés. (par exemple, le premier passage à 0 de l'OGD ou de l'ODGD). Cette étude nous a montré qu'il n'était pas possible d'obtenir simultanément des résultats de synthèse corrects à la fois pour l'ODG et l'ODGD (même avec avec un demie fenêtre de Hann synchrone [9]). Ce problème de modélisation et les limitations dues à l'utilisation d'un tampon audio périodique nous ont amené à réviser l'architecture de ce module de synthèse. Les discontinuités de l'ODGD causées par la troncature de l'ODG sont illustrées sur la Figure 3.12.

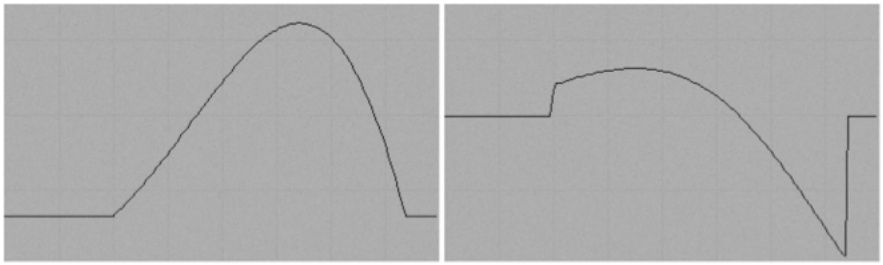


FIGURE 3.12 – *Discontinuité de l'ODGD (à droite) due à la troncature de l'*ODG *au premier passage à* 0 *de la pulsation* CALM *(à gauche). (D'Alessandro et al., 2007)*

Solution par calcul direct

Cette partie décrit une autre version de l'implémentation de la réponse impulsionnelle du filtre anticausal, visant à résoudre le problème précité. En outre, cette dernière solution évite l'utilisation d'un tampon d'une période, comme décrit précédemment, réduisant ainsi la latence globale du système. L'idée principale derrière cette solution était de réduire la charge en allocation mémoire, et son avantage est de pouvoir générer simultanément l'ODG et l'ODGD, avec leurs propres points de

9. Cette méthode de fenêtrage multiplie la partie croissante de la pulsation glottique (ou de sa dérivée) - c'est-à-dire la partie entre le passage à 0 et le maximum positif - par la partie croissante de la fenêtre de Hannning

troncature et de fenêtrage indépendamment.

Au lieu de calculer une version causale de la réponse impulsionnelle hors ligne puis de copier celle-ci à l'envers dans un tampon fixe, le calcul se fait ici directement. L'équation itérative correspond en réalité à un filtre causal instable. La divergence exponentielle du filtre est alors évitée en stoppant le calcul exactement à l'instant de fermeture glottique (GCI). On peut noter que l'ODG et l'ODGD peuvent toutes deux être accomplies grâce à la même équation itérative, par simple modification des valeurs des deux premiers échantillons utilisés comme conditions initiales du processus d'itération.

Le problème majeur de cette implémentation repose alors sur le fait que la génération directe de la forme d'onde doit être synchronisée avec la taille de tampon standard de Pure Data (Puckette, 1996) ou Max/MSP (Max/MSP, 2008). Cette taille de tampon standard est de 64 échantillons, ce qui, à un taux d'échantillonnage audio de 44.1 kHZ, correspond approximativement à une fréquence de 690 Hz. Dans la majorité des cas, la fréquence fondamentale de l'onde glottique est inférieure à 690 Hz, ce qui signifie que plusieurs tampons audios sont nécessaires pour accomplir le calcul complet d'une période donnée. Toutefois, à chaque fois que le tampon est vide (i.e. arrive à terme), la routine de calcul principale est appelée et les valeurs de a_1 et a_2 doivent être sauvegardées jusqu'à la fin de la période de calcul courante. Un drapeau indicateur lié à l'ouverture de la glotte a donc été introduit, et fixé à la valeur de la durée de la période courante (en nombre d'échantillons), et les valeurs de a_1 et a_2 ne sont pas modifiées tant que cet index n'a pas atteint 0. Une fois les valeurs de T_0, T_e, γ, a_p et b_p calculées à l'instant d'ouverture glottique, seuls a_1 et a_2 doivent être conservés, car ce sont les seules variables à prendre en compte dans les équations de l'ODGD.

Des tests poussés menés sur cette implémentation nous ont permis de révéler que cette version est plus robuste que la précédente. En particulier, cette implémentation ne se bloque pas lorsque des valeurs exotiques sont envoyées à l'algorithme (même si ces configurations ne possèdent pas de signification physiologique). Finalement, il convient de noter que cette amélioration ne concerne que le module de génération glottique (`almPulse~`). Le filtre d'inclinaison spectrale (`stFilter~`) n'a quant à lui pas été modifié.

La figure 3.13 représente le patch Pure Data de cette version du synthétiseur RTCALM, permettant de calculer et d'afficher l'ODG et l'ODGD. Ce patch contient également le filtre d'inclinaison spectrale, ainsi qu'un filtre vocalique [10] permettant ainsi de produire la synthèse de quelques voyelles prédéfinies.

10. filtre `vowel~` disponible dans la librairie `sigpAck` développée par Y. Degoyon.

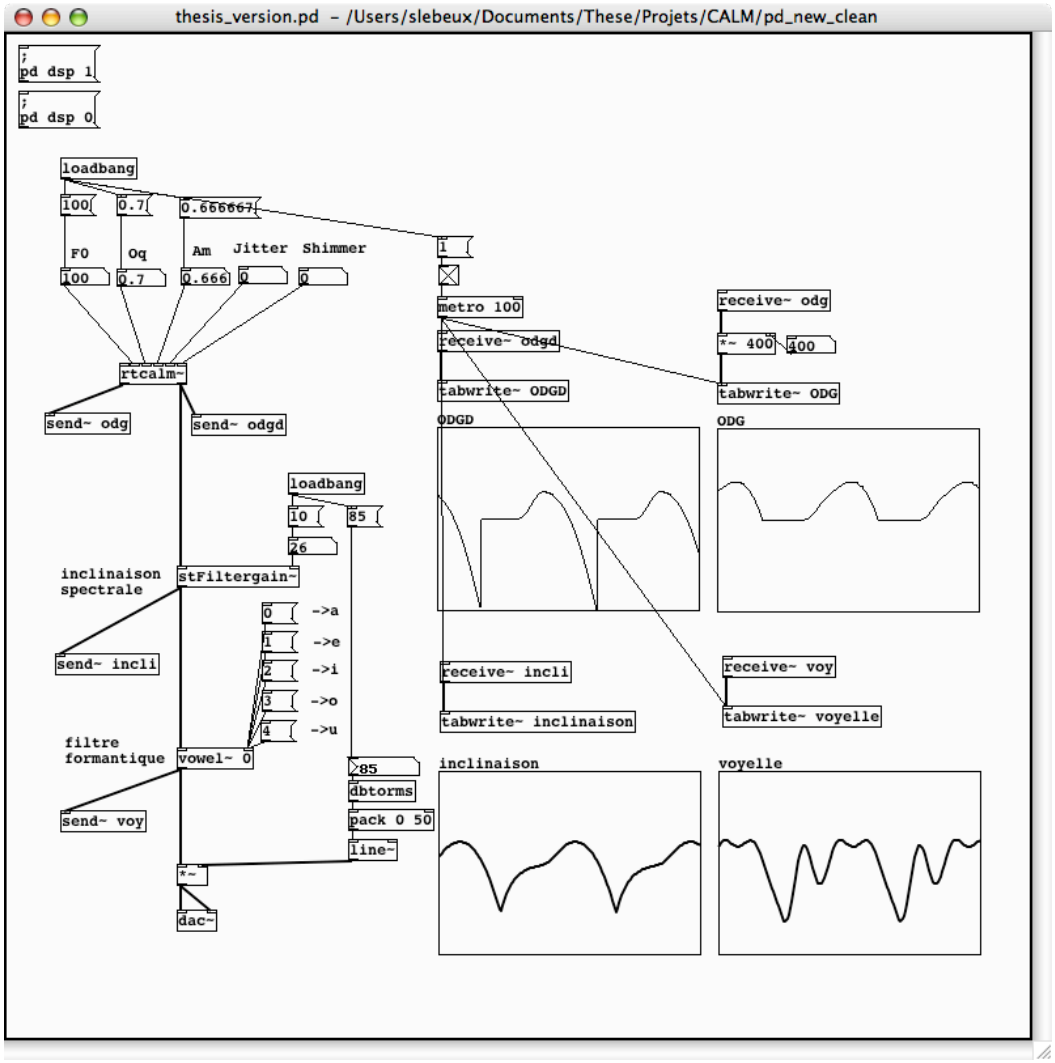


FIGURE 3.13 – *Capture d'écran du patch Pure Data, pour la génération de l'*ODGD *et de son filtrage.*

Sur cette illustration, on peut voir sur les différents graphiques représentés, de haut en bas et de gauche à droite, l'ODGD et l'ODG sans filtrage d'inclinaison spectrale, puis l'ODGD après le filtrage d'inclinaison spectrale et enfin le signal résultant après le filtre formantique.

3.3.3 Composantes non périodiques

Comme décrit théoriquement dans (Doval et al., 2003), le flux glottique est un signal déterministe, complètement contrôlé par un jeu de paramètres. Ajouter du naturel implique l'utilisation de certainres composantes aléatoires que nous allons décrire maintenant.

Jitter

Le jitter est une instabilité naturelle dans la valeur de la fréquence fondamentale, et est défini comme la fluctuation en fréquence d'une période fondamentale à la suivante. Le jitter est une grandeur sans unité, définit en pourcentage de la fréquence fondamentale instantanée. Ainsi, pour un son voisé de fréquence $F_0 = 125\ Hz$, correspondant à une période fondamentale de $T_0 = 8\ ms$, un jitter de 1%, signifie que la fréquence fondamentale estimée du signal de parole pourra varier théoriquement entre $123,75\ Hz$ et $126,25\ Hz$, soit une période fondamentale comprise entre $7.92\ ms$ et $8.08\ ms$.

Il peut être modélisé par une valeur aléatoire (distribution gaussienne autour de zéro avec une variance dépendant de la quantité de jitter introduite), rafraîchie à chaque période et ajoutée à la valeur stable de la fréquence fondamentale.

Le jitter est parfois défini sur plusieurs périodes fondamentales consécutives, de la manière suivante :

$$G = \frac{1}{N-1} \sum_{1}^{N-1} |f(n+1) - f(n)| \tag{3.33}$$

où $f(i)$ représente la fréquence fondamentale instantanée de la période i.

On trouve également dans la littérature des modèles statistiques plus sophistiqués (Schoentgen and de Guchteneere, 1995), utilisés principalement en analyse de la parole. Toutefois, l'ajout d'une valeur aléatoire à la fréquence fondamentale paraît suffisante pour une synthèse plus naturelle. Dans notre objet externe, le jitter est calculé de la manière suivante :

$$rJ = (\frac{Rand()\%1000}{500} - 1) \times Jitter \quad avec \quad 0 \leq Jitter \leq 2 \tag{3.34}$$

La valeur aléatoire renvoyée par la fonction $Rand()$ est comprise entre 0 et 1000 de façon à augmenter la dynamique du jitter. La valeur de rJ entre 0 et 1 (correspondant alors à un pourcentage) est ensuite ajoutée à la valeur de F_0 comme suit :

$$F_0 = F_0 \times (1 + rJ) \tag{3.35}$$

Shimmer

Le shimmer est une instabilité naturelle dans la valeur d'amplitude. Elle peut être modélisée par une valeur aléatoire (distribution gaussienne autour de zéro avec une variance dépendant de la quantité de shimmer introduite), rafraîchie toutes les périodes et ajoutée à la valeur stable de l'amplitude.

De la même manière, il est possible de définir le shimmer pour l'amplitude, défini pareillement

comme un pourcentage de déviation de l'amplitude d'une période fondamentale à l'autre, et qu'il est possible d'exprimer sur plusieurs périodes de la manière suivante :

$$S = \frac{1}{N-1} \sum_{1}^{N-1} |a(n+1) - a(n)| \tag{3.36}$$

où $a(i)$ représente l'amplitude instantanée de la période i.

Dans notre objet externe, le Shimmer est calculé de la manière suivante :

$$rS = 1 - \frac{Rand()\%1000}{1000} \times \frac{Shimmer}{100} \quad avec \quad 0 \leq Shimmer \leq 2 \tag{3.37}$$

Notons que la valeur d'amplitude du signal audio est comprise entre -1 et 1, aussi la valeur de rS est également comprise entre 0 et 1, et égale à 1 pour un shimmer nul. Ensuite cette valeur est multipliée simplement à la valeur d'amplitude de la façon suivante :

$$U'(n) = U'(n) \times rS \tag{3.38}$$

où, $U'(n)$ représente la valeur de l'$ODGD$ à un instant quelconque.

Notons, par ailleurs, que le jitter et le shimmer sont des notions conjointes, car de nombreuses études sur de la parole naturelle ont montrées (Ferrand, 1995; Fuller and Horii, 1986; Pause-wang Gelfer and Fendel, 1995; Sorensen and Horii, 1983) que l'on observe jamais de jitter sans shimmer. Enfin, pour un signal de parole réel, il est impossible d'observer un jitter (et/ou un shimmer) nul.

Turbulence

Pour ce qui est du bruit additif, le problème est un peu plus complexe. Pour ce qui est de la modulation, la réalisation est assez directe. Il suffit d'utiliser l'$ODGD$ en sortie de l'objet externe `almPulse~` pour effectuer une multiplication avec le bruit et ajouter ensuite cette quantité à l'$ODGD$ avant le filtre d'inclinaison spectrale. Selon Dik Hermes (Hermes, 1991) le fait est que afin d'obtenir une fusion entre la composante quasi-périodique et la composante de bruit, il est nécessaire que celles-ci soient en synchronie. Au cas échéant, les deux composantes sonores paraissent disjointes, conduisant à une perception moins naturelle de la voix. Par ailleurs, Hermes rapporte le fait que lors d'une désynchronisation des deux composantes, la voix peut être perçue comme rugueuse. Il paraît donc intéressant malgré tout de conserver un contrôle sur le délai entre l'$ODGD$ et la source de bruit, suivant que l'on souhaite une voix plus soufflée, ou avec plus de rugosité.

Dans un premier temps, nous avons utilisé un bruit rose, qui semblait nous donner un résultat satisfaisant. Cela dit, les études montrent (Hillman et al., 1983) que la répartition spectrale du bruit

correspondrait plutôt à une forme en cloche en échelle logarithmique, comme l'indique la figure 3.14 suivante.

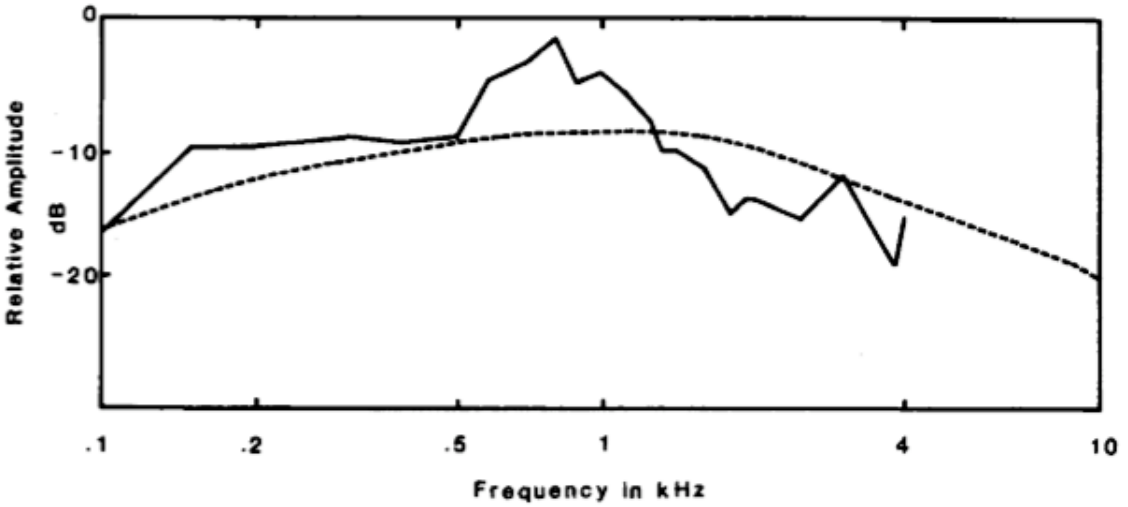


FIGURE 3.14 – *Comparaison du spectre idéal et mesuré de la pression glottique de la source(d'après (Hillman et al., 1983))*

Toutefois, il faut noter ici que cette courbe est obtenue par une mesure acoustique après filtrage par le conduit vocal. Comme le relève donc Hillman, un simple bruit blanc suffit à produire une voix soufflée ou chuchotée. Aussi, dans un second temps, nous avons cherché à reproduire ce type de bruit par filtrage d'un bruit blanc par un filtre dont la réponse en fréquence corresponde mieux à ce type de bruit. D. Hermes mentionne par ailleurs, que le bruit d'aspiration caractéristique de la voix soufflée, se situe plutôt dans les moyennes et les hautes fréquences. Ainsi, l'utilisation d'un bruit blanc filtré par un filtre passe-haut semble une bonne méthode pour modéliser les turbulences. Il est ensuite possible de régler a posteriori la fréquence de coupure et le gain du filtre passe-haut pour obtenir une voix plus ou moins soufflée, ou chuchotée. Ces résultats sont confirmés par Stevens & Hanson (Stevens and Hanson, 1995), qui suggèrent un bruit blanc filtré passe-haut, avec une fréquence de coupure atour de $2-3$ kHz. Ils indiquent en outre, rapport signal sur bruit d'environ 10 dB jusqu'à 5 kHz pour une voix modale, passant à 0 dB à 4 kHz lors d'une fermeture incomplète de la glotte (voix soufflée).

Notons ici qu'une méthode efficace pour la génération des apériodicités est celle de l'utilisation de formes d'ondes formantiques, générées de manière aléatoire par un processus de Poisson, comme indiqué par G. Richard (Richard and d'Alessandro, 1996), permettant de produire les bruits structurels et additifs précités. En outre, cette méthode fournit également un moyen de produire des bruits de types plosifs ou impulsionnels, moyennant l'utilisation d'une densité de probabilité non gaussienne (Grau et al., 1993). Ces méthodes dépassent toutefois le cadre strict de notre étude, mais constituent un point de comparaison intéressant pour une future évaluation de notre

synthétiseur.

3.3.4 Description des fonctions de mapping

On peut trouver un inventaire approfondi d'un grand nombre de systèmes de contrôle de la synthèse sonore dans l'article de M. Wanderley et P. Depalle (Wanderley and Depalle, 2004), et nous allons essayer de situer le cadre de nos explorations au sein de cette typologie des différentes interfaces pour le contrôle de la synthèse sonore. Concernant l'acquisition des données, dans les différents systèmes que nous avons développés, celle-ci a toujours été effectuée de manière directe. L'autre possibilité d'acquisition des données, également relevée par W. Goebl (Goebl et al., 2008) est celle de l'acquisition indirecte des données par analyse et rétro-action de certaines caractéristiques du signal audio généré en premier lieu. Compte tenu des problèmes liés à l'analyse de la source glottique évoqué précédemment, une acquisition indirecte n'était en effet pas souhaitable, malgré les interdépendances présentes parmi les dimensions vocales. Enfin, Wanderley et Depalle différencient un troisième moyen d'acquisition qui est celui de l'analyse de signaux physiologiques, telles que l'activité musculaire ou cérébrale. Bien que nous ayons eu l'opportunité de tester de tels capteurs, nous n'avons pas pu approfondir ce types de solutions. Notons toutefois que l'évolution temporelle de ce type de variations physiologiques s'opère la plupart du temps sur des échelles temporelles trop grandes pour pouvoir fournir un moyen de contrôle efficace de la synthèse vocale, si ce n'est pour le contrôle d'un chant de type diphonique, pour lequel l'évolution des formants et des caractéristiques de la source est relativement lente.

Concernant la typologie de contrôleur, notre approche se situe parmi les contrôleurs alternatifs. Nous n'avons pas essayé en effet, à l'instar de la machine de von Kempelen, ou plus proche de nous, des implémentations de P. Cook (Cook and Leider, 2000; Cook, 2005) de chercher à simuler le rôle des organes sub-glottiques. L'utilisation de tels dispositifs pour le contrôle de la synthèse vocal paraît en effet plus approprié dans le cadre d'une modélisation physique du conduit vocal. Nous avons plutôt tenté d'explorer les possibilités offertes par des contrôleurs ad-hoc, tels que le Méta-instrument, voire d'étudier l'effet de contrôleurs haptiques permettant d'adresser la problématique de l'effort à fournir pour pouvoir produire le signal vocal.

Wanderley et Depalle notent également que l'on peut classer les instruments de musique numériques selon deux catégories suivant que ces derniers cherchent à reproduire le plus fidèlement possible un phénomène acoustique donné, issu de l'analyse d'un instrument standard, ou au contraire qu'ils cherchent à s'abstraire de toute correspondance avec un quelconque instrument réel, et par là-même explorer de nouvelles sonorités. Notre approche a évidemment été de coller le plus possible à la production du signal vocal, notre idée de départ étant que l'on peut toujours étendre les plages de valeurs des paramètres du synthétiseur pour obtenir des sons qui ne possèdent aucune réalité physique. Par ailleurs, dans le cadre de la synthèse sonore, et à plus forte raison de la

synthèse vocale, c'est bien souvent la dynamique temporelle des paramètres qui fournit le naturel souhaité, là où une version trop "stable" est généralement directement perçue comme synthétique voire robotique.

Enfin, concernant le mapping, le modèle prédominant et qui fait consensus à l'heure actuelle est celui d'une approche en multi-couches du contrôle de la synthèse sonore. Les fonctions de mapping ou fonctions de correspondance définissent les connexions effectuées entre les différents capteurs utilisés et les paramètres bas niveau du synthétiseur. Comme ont pu le décrire J.B. Rovan etal. (Rovan et al., 1997), Hunt et Kirk (Hunt et al., 2000; Hunt and Kirk, 2000), le mapping fait partie intégrante de l'instrument, et ainsi les possibilités offertes par l'instrument, ou en d'autres termes son expressivité, dépendent en grande partie des correspondances réalisées avec les paramètres de synthèse.

Ces correspondances peuvent être réalisées de différentes manières, comme, par exemple, avec des réseaux de neurones (Fels and Hinton, 1998), ou des stratégies de type un-vers-un, plusieurs-vers-un, un-vers-plusieurs (Rovan et al., 1997). Ces dernières fournissent, selon Hunt et Kirk (Hunt and Kirk, 2000) les stratégies de correspondance fournissant la plus grande expressivité. Il a été montré par la suite, que l'utilisation d'une couche supplémentaire facilitait grandement la manipulation des paramètres, pour peu que l'espace dans lequel est effectué la manipulation soit un espace perceptif (Arfib et al., 2002). Cet espace perceptif, permet de manipuler des dimensions plus pertinentes du point de vue auditif, et d'éviter ainsi de manipuler un trop grand nombre de paramètres pour se concentrer sur un nombre réduit de paramètres de plus haut niveau, mais possédant une pertinence du point de vue de l'auditeur et donc de l'interprète.

Dans le cadre de la synthèse vocale, comme nous venons de le dire, le but est d'essayer de reproduire une voix qui soit la plus convaincante possible. Il existe une littérature abondante en analyse de parole, permettant de décrire, au moins qualitativement les domaines de variation des paramètres de source glottique selon les différentes dimensions vocales. Les dimensions vocales sont, comme nous l'avons vu précédemment, intrinsèquement perceptives. Aussi, l'espace perceptif de manipulation de notre synthétiseur sera directement défini par ces dimensions vocales.

Conformément aux descriptions données dans la section 3.2, afin de pouvoir contrôler les différents aspects de qualité vocale, il convient de s"intéresser aux différentes dimensions de qualité vocale. D'après ces définitions, deux tâches principales peuvent être traitées. Tout d'abord, l'implémentation des fonctions de mapping entre ces dimensions et les paramètres bas niveau. Puis, l'identification et l'implémentation des phénomènes de dépendances inter-dimensionnelles. Dans ce domaine, de nombreuses théories différentes ont été proposées reliant plusieurs aspects intra et inter-dimensionnels de la production vocale (Klatt and Klatt, 1990; Hanson and Chuang, 1999; Alku and Vilkman, 1996; Traunmüller and Eriksson, 2000; Henrich et al., 2004; Henrich et al., 2005).

Nous avons décidé de nous focaliser uniquement sur quelques uns d'entre eux, à savoir l'implémentation directe de la tension et de l'effort vocal, la réalisation du phonétogramme, et de concevoir notre plate-forme de synthèse de façon à pouvoir être entendue facilement (par exemple, la correction des relations existantes, l'ajout de nouvelles fonctions de mapping, ...).

Relations entre les dimensions et les paramètres de synthèse

Au cours de l'atelier eNTERFACE'06, nous nous sommes focalisés sur plusieurs aspects du processus dimensionnel. Tout d'abord, nous avons considéré les relations entre un jeu restreint de dimensions (F_0,V,T et M_i) et de paramètres de synthèse (O_q, α_m, T_l). Puis, nous avons décidé de réaliser notre schéma de correspondance en considérant deux processus orthogonaux du contrôle dimensionnel. D'une part, l'effort vocal (V) (également lié aux variations de F_0 par le phonétogramme, voir prochain paragraphe) et les mécanismes (M_i) contrôlent les valeurs standards des paramètres (O_{q_0}, α_{m_0}, T_{l_0}). D'autre part, la tension (T) contrôle l'excursion des valeurs de O_q et α_m autour de leurs valeurs standards (ΔO_q, $\Delta\alpha_m$). Selon cette approche, les valeurs effectives des paramètres de synthèse sont décrites par les relations suivantes :

$$O_q = O_{q_0} + \Delta O_q \tag{3.39}$$

$$\alpha_m = \alpha_{m_0} + \Delta\alpha_m \tag{3.40}$$

$$T_l = T_{l_0} \tag{3.41}$$

Les équations suivantes considèrent les paramètres V et T normalisés entre 0 et 1, tandis que M_i représente l'$i^{ème}$ mécanisme de phonation (M_0 est traité indépendamment).

- $O_{q_0} = f(V|M_i)$

$$O_{q_0} = 1 - 0.5 \times V \mid M_1 \tag{3.42}$$

$$O_{q_0} = 0.8 - 0.4 \times V \mid M_2 \tag{3.43}$$

- $\alpha_{m_0} = f(M_i)$

$$\alpha_{m_0} = 0.6 \mid M_1 \tag{3.44}$$

$$\alpha_{m_0} = 0.8 \mid M_2 \tag{3.45}$$

- $T_{l_0} = f(V)$

$$T_{l_0} = 55 - 49 \times V \tag{3.46}$$

- $\Delta O_q = f(T)$

$$\Delta O_q = (1 - 2T) \times O_q + 0.8T - 0.4 \mid T \leq 0.5 \tag{3.47}$$

$$\Delta O_q = (2T - 1) \times O_q + 2T + 1 \mid T > 0.5 \tag{3.48}$$

– $\Delta\alpha_m = f(T)$

$$\Delta\alpha_m = (0.5T - 1) \times \alpha_m - 1.2T + 0.6 \mid T \leq 0.5 \tag{3.49}$$

$$\Delta\alpha_m = (0.25 - 0.5T) \times \alpha_m + 0.4T - 0.2 \mid T > 0.5 \tag{3.50}$$

La dernière adaptation de paramètres concerne la distortion perceptive de O_q (distortion quadratique) et α_m (distortion en racine carrée) au sein de leurs étendues de variations respectives (O_q : 0.4 à 1 ; α_m : 0.6 à 0.8).

Relations inter-dimensionnelles : le phonétogramme

Une caractéristique importante de la production vocale humaine concerne le fait que nous ne soyons pas capables de produire n'importe quelle hauteur (F_0) pour n'importe quel effort vocal (V) donné. Une relation stricte existe entre ces propriétés acoustiques particulières. Par exemple, il n'est pas possible de produire une hauteur très basse (autour de 80 Hz) avec un niveau de pression sonore supérieur à 80 dB (pour un homme) ou, inversement, de produire une hauteur élevée avec un effort vocal faible. Une tentative d'atteinte de telles extrémités conduit à un arrêt brusque de la production vocale. Cette relation existant entre la hauteur et l'effort vocal est appelée un *phonétogramme*, et l'évolution de cette dépendance varie ostensiblement d'un locuteur à un autre, selon par exemple que la personne est un chanteur professionnel ou non, un homme ou une femme, possède une voix pathologique ou non, ... Dans une première approche, nous avons décidé d'implémenter un phonétogramme "moyen", d'après le travail réalisé par N. Henrich (Henrich et al., 2005) et B. Roubeau (Roubeau et al., 2009). La figures 3.15 représente des phonétogrammes moyens pour un homme et une femme.

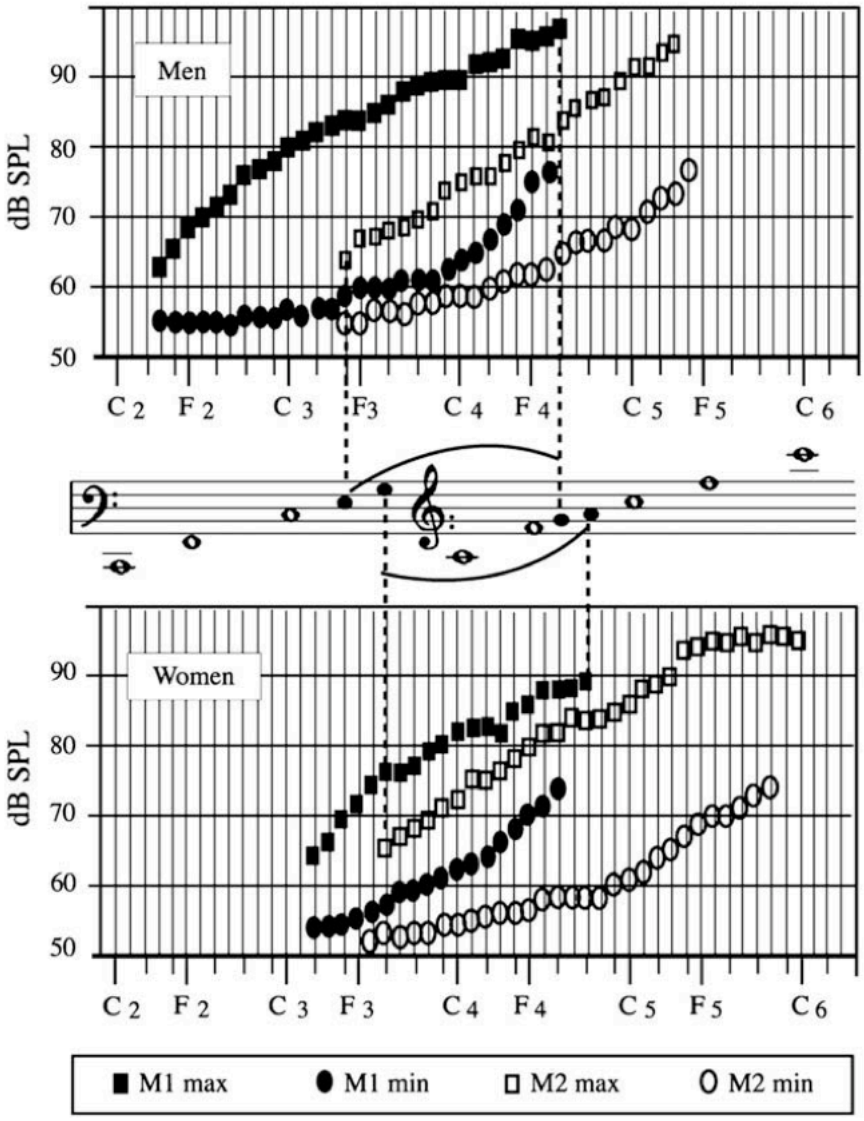


FIGURE 3.15 – *Phonétogramme moyen (*M1, M2*) pour une voix masculine (en haut) et féminine (en bas) (d'après (Roubeau et al., 2009))*

En outre, ce phénomène implique différents types de configurations laryngées. Nous nous intéressons ici principalement à deux configurations, les premier et second mécanismes des plis vocaux (M_1 et M_2). Ces deux mécanismes laryngés sont, selon la typologie classique du chant, appelés voix de poitrine et voix de fausset (ou voix de tête). Alors, comme le montre la figure 3.15, il n 'est pas possible de produire n'importe quelle fréquence selon les deux mécanismes, néanmoins les deux mécanismes possèdent une région commune vers le milieu du phonétogramme. Cette région permet le passage d'un mécanisme à l'autre. Suivant le travail présenté dans (Bloothooft et al., 2001), l'étendue fréquentielle où ce passage s'effectue est d'environ une octave (soit 12 demi-tons).

La caractéristique principale de ce passage est de provoquer un saut en fréquence (F_0). Ainsi, lors de la production d'un glissando croissant depuis M_1 vers M_2, il se produit un saut d'environ 8 demi-tons, tandis que ce saut est d'environ 12 demi-tons lors d'un glissando descendant. Les probabilités des intervalles de saut sont illustrés sur les figures 3.16 et 3.17. Sur le premier, on peut noter que le saut en fréquence dépend également de la fréquence fondamentale à laquelle il s'opère. Ainsi que l'on peut le noter, plus le passage s'effectue à une hauteur élevée, plus la probabilité pour que ce saut soit faible est grande (lors d'un passage M_1 vers M_2).

On déduit de ces observations que ce phénomène introduit un hysteresis, c'est à dire que le saut ne se fait pas symétriquement suivant le sens de ce dernier. Pour simplifier, on comprend aisément qu'afin de ne pas perturber la phonation, on aura tendance à conserver le plus longtemps possible le mécanisme phonatoire dans lequel on se trouve avant le saut. Pour la plupart des locuteurs ou chanteurs non entraînés, ce saut n'est pas maîtrisé tandis que les chanteurs entraînés arrivent à cacher plus ou moins doucement ce saut, bien qu'ils ne puissent pas empêcher le changement de mécanisme.

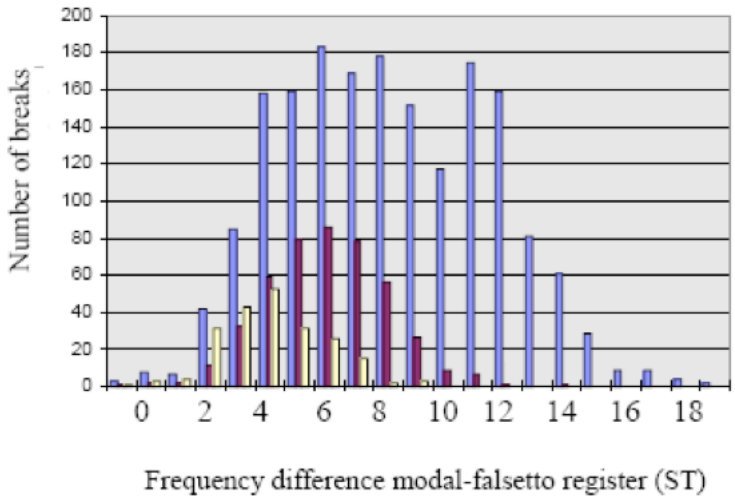


FIGURE 3.16 – *Sauts en fréquence exprimés en demi-tons depuis le registre de poitrine (modal) vers le falsetto. En bleu, pour un saut à* 200 Hz*, en rouge pour un saut à* 300 Hz*, et à* 400 Hz *en jaune. (d'après (Bloothooft et al., 2001))*

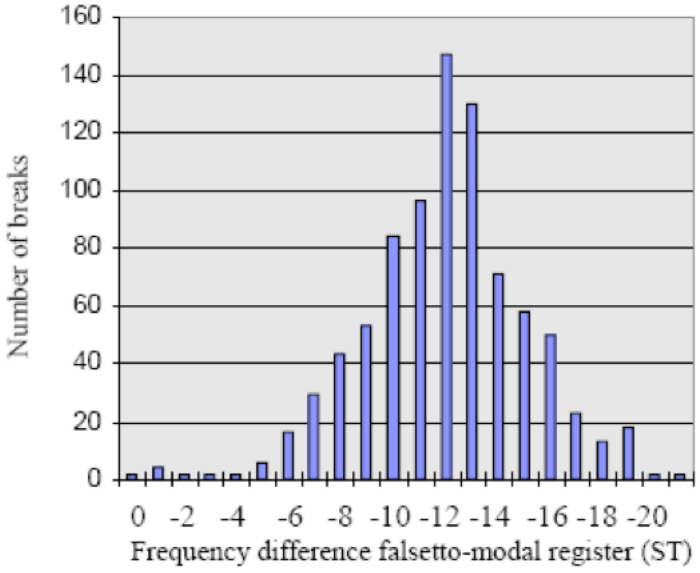


FIGURE 3.17 – *Sauts en fréquence en demi-tons depuis le falsetto vers le registre de poitrine (modal). (d'après (Bloothooft et al., 2001))*

Cependant, le fait d'utiliser un phonétogramme générique empêche, dans une certaine mesure, de pouvoir donner une véritable identité à la voix de synthèse, même si cela ajoute néanmoins du naturel à la voix de synthèse, par rapport à un simple augmentation du volume, indépendante des paramètres de source glottique. Nous verrons dans la section 3.4.4, une solution adoptée permettant de charger dans le synthétiseur n'importe quel phonétogramme à partir d'un simple fichier texte.

Le formant du chanteur

Une des particularités spectrales de la voix chantée, est ce que l'on appelle le *formant du chanteur* (Sundberg, 2001). Ce phénomène se traduit par une augmentation de l'énergie spectrale autour de 3000 Hz. Cet enrichissement spectral est appelé formant du chanteur, tout simplement parce que c'est un peu comme si un nouveau pic apparaissait dans le spectre autour de cette fréquence. En réalité, il s'agit plutôt d'un mouvement attractif des $3^{\text{ème}}$ et $5^{\text{ème}}$ formants, la fréquence de résonance du $3^{\text{ème}}$ formant augmentant, et celle du $5^{\text{ème}}$ descendant. Ainsi, on se retrouve avec une augmentation significative de la bande passante du $4^{\text{ème}}$ formant. On peut observer l'effet de ce phénomène sur la figure 3.18 suivante.

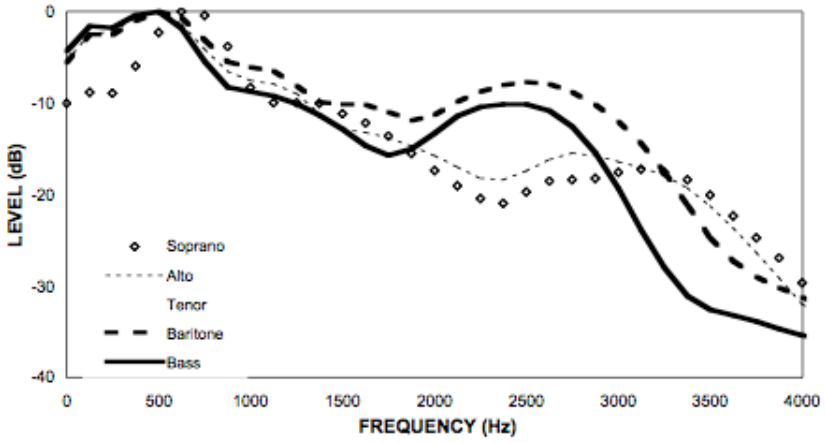


FIGURE 3.18 – *Enveloppes spectrales à long terme pour différents types de chanteurs (d'après (Sundberg, 2001))*

Du point de vue de la synthèse, on peut assez aisément simuler le formant du chanteur en augmentant sensiblement la fréquence centrale du $3^{\text{ème}}$ filtre formantique et en abaissant celle du $5^{\text{ème}}$. Comme ce phénomène peut être plus ou moins prononcé, un contrôle continu est effectué permettant de rapprocher progressivement les $3^{\text{ème}}$ et $5^{\text{ème}}$ formants du $4^{\text{ème}}$. La figure 3.18 précédente nous informe toutefois que la position et l'amplitude de ce formant sera différente suivant le type de chanteur. Ainsi, pour les basses et les barytons, le formant sera plus bas (autour de 2500 Hz) avec une amplitude relative plus prononcé, tandis que pour les sopranos et les altos le formant se situera autour de 3000 Hz avec une amplitude plus faible. Malheureusement, sur cette figure n'apparaît pas la courbe pour les ténors. Ces derniers ont des caractéristiques entre les deux classes pré-cités, c'est-à-dire un formant à la fois élevé en fréquence et en amplitude.

Toujours est-il que si l'on exclut les ténors, on peut dire que le formant du chanteur va dépendre du mécanisme utilisé puisque les basses et les barytons utiliseront quasi exclusivement le mécanisme M_1, tandis que les sopranos et les altos utiliseront principalement le mécanisme M_2. Les analyses effectués par Sundberg reposent sur des analyses spectrales à long terme ($LTAS$), par conséquent on peut penser que les résultats observés pour les ténors sont la traduction du fait que le mécanisme utilisé était parfois M_1, parfois M_2, conduisant en moyenne à une augmentation de la position et de l'amplitude du formant. En résumé, on peut considérer que la position et l'amplitude du formant du chanteur sera dépendante du mécanisme utilisé.

Vibrato

Même, en voix chantée, où la note à jouer est fixe et imposée, apparaît un autre phénomène appelé *vibrato*. Le vibrato, quant à lui, est une modulation de la fréquence fondamentale, apparaissant le plus souvent pour des notes tenues suffisamment longues. Le vibrato peut être efficacement introduit

en synthèse de la manière suivante :

$$F(t) = F_0 + A_0 \times \sin(\omega_v t) \tag{3.51}$$

où, $F(t)$ est la fréquence instantanéé, F_0 la fréquence fondamentale, $\omega_v = 2\pi F_v$ la pulsation du vibrato et A_0 son amplitude, également appelée profondeur.

Le vibrato ne représente pas à proprement parler une apériodicité, vis-à-vis de la fréquence fondamentale, puisque c'est une modulation de fréquence. Cependant, il paraît judicieux d'introduire cette notion ici puisque il constitue tout de même une variation autour de la fréquence fondamentale. Enfin, l'effet d'une modulation d'amplitude en voix chantée est appelée *tremolo*, et sera défini de la même manière que le vibrato pour F_0.

3.4 Les différents instruments basés sur RTCALM

Cette section décrit certaines des différentes configurations que nous avons réalisées. Le but premier de ce travail était de réaliser des tests intensifs en temps réel du modèle de synthèse RTCALM ainsi que des mappings des dimensions de qualité vocale. Aucun contrôleur dédié n'a cependant été conçu pour cela. Seuls des interfaces existantes comme des tablettes, des joysticks ou des claviers ont été utilisés.

3.4.1 Premier instrument

Ce premier instrument utilise à la fois des mouvements manuels libres dans l'espace ainsi que des gestes d'ouverture/fermeture de la main. Cette application est développée dans l'environnement Max/MSP (Max/MSP, 2008). Les deux types de gestes semblent adéquats pour un contrôle précis des dimensions de qualité vocale comme la mélodie, l'effort, la pression vocale, la rugosité et le souffle.

Dans cette première implémentation, nous avons utilisé un clavier pour jouer des notes MIDI de façon à déclencher des voyelles à différents hauteurs. Ainsi, en utilisant le clavier, nous pouvons garder le gant de données libre pour un contrôle fin de F_0 pour réaliser un vibrato, un portamento ou d'autres types d'ornements mélodiques. Le contrôle précis de F_0 par la position du gant seule s'est révélée ardue du fait de l'absence de références justes pour les notes de la gamme, dû au caractère approximatif des gestes manuels.

Les notes tempérées (ou toute autre convention) délivrées par le clavier peuvent être modifiées dans une certaine mesure grâce à la localisation du gant selon un certain axe (l'axe transversal donnant une meilleure ergonomie car l'on n'a pas besoin de plier le coude pour accomplir le vibrato). Le gain général est connecté à l'axe longitudinal du gant. Puis, le vibrato et l'enveloppe d'amplitude du son peuvent être produits par des mouvements circulaires de la main.

Les dimensions de qualité vocale sont contrôlées par flexion des doigts. Le pouce contrôle l'effort vocal (tilt spectral), l'index contrôle le souffle (lié au bruit additif), le majeur contrôle la dimension tendu/relâché (liée au formant glottique), l'annulaire contrôle la ruguosité (liée au jitter et au shimmer). Les modifications de qualité vocale sont accomplies par ouverture et fermeture de la main entière ou de certains doigts. Les voyelles paramétrées sont associées aux touches du clavier de l'ordinateur. Les formants des voyelles peuvent être modifiés par des appareils additionnels, comme le pédalier ou les joysticks.

En résumé, pour cet instrument nous avons :

1. La main gauche contrôle la hauteur grâce au clavier MIDI (notes tempérées)

2. Les mouvements de la main droite contrôlent à la fois les modulations fines, et le phrasé de note.
3. Les doigts de la main droite contrôlent la tension, l'effort, la rugosité et le souffle.

Dans cette implémentation, le phrasé de note consiste en de relativement larges mouvements de la main. Une solution alternative consiste à coupler l'effort et le phrasé de note par les mouvements des doigts, et de garder une dimension pour le mouvement de la main pour contrôler une autre dimension vocale (par exemple, le souffle). Alors, le phrasé est contrôlé par des mouvements de doigts plus petits et plus rapides.

Cependant, cet instrument ne paraît pas très aisé pour le chant de style "classique". Les transitions mélodiques permises par les claviers n'ont pas de qualité chantée particulière. Les gestes d'ouverture/fermeture de la main sont plus ou moins comparables aux gestes d'ouverture/fermeture dans le conduit vocal et aux abductions/adductions des cordes vocales. Cette analogie est potentiellement utile pour la synthèse de voix chantée.

CALM, clavier et gant

Pour ce synthétiseur, seul le gant de données P5 est utilisé. L'interface d'entrée autorise 8 paramètres variables continus en parallèle : 3 positions spatiales $\mathrm{x}, \mathrm{y}, z$, associées avec le mouvement du gant relativement à une borne réceptrice fixe placée sur la table et les 5 paramètres associés à la flexion des doigts de la main. Plusieurs touches du clavier de l'ordinateur contrôlent les configurations de voyelles. Le gant dirige le CALM. Seules deux dimensions spatiales (x, z) sont utilisées comme suit : la variable x est liée à l'intensité E et la variable z à la fréquence fondamentale. Tous les doigts sauf l'auriculaire sont utilisés pour contrôler respectivement (à partir du pouce) le ratio de bruit, le quotient ouvert, le tilt spectral et l'asymétrie. Ce mapping est plus fiable et efficace (comparé au clavier utilisé dans le premier instrument). Seule une courte phase d'entraînement est nécessaire pour obtenir des variations de source vocale très naturelles. Le clavier de l'ordinateur est utilisé pour modifier les valeurs des filtres formantiques pour synthétiser les différentes voyelles, et les articulations basiques du conduit vocal. Cependant, il n'est pas très facile de plier un doigt de la main sans en entraîner un autre. Par exemple, il est assez difficile de plier le majeur sans que l'annulaire bouge aussi. Une description schématique générale de ces deux instruments est donné sur la figure 3.19 suivante.

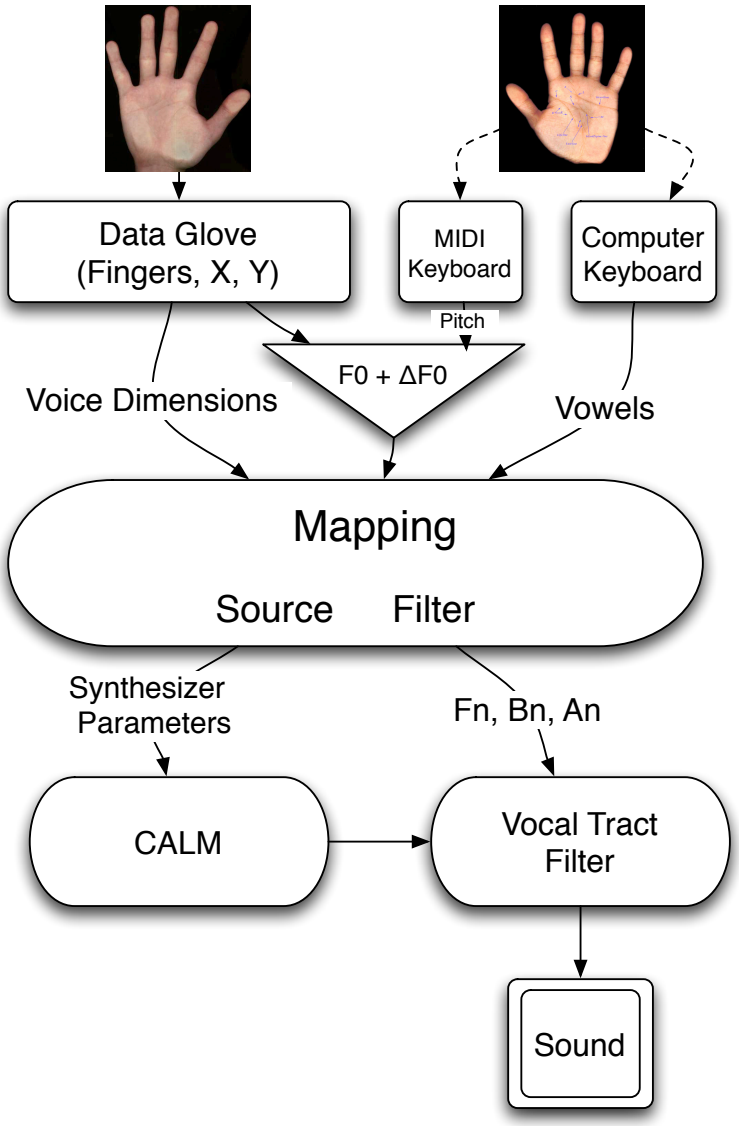


FIGURE 3.19 – *Mapping des dimensions vocales de contrôle aux paramètres du modèle de source glottique. (d'Alessandro et al., 2006)*

3.4.2 Deuxième instrument

Le point central de ce deuxième instrument réside dans sa simplicité d'utilisation et d'apprentissage. Les différents choix ont été faits pour obtenir ce résultat. En premier lieu, nous avons décidé de nous focaliser sur la qualité vocale. Le contrôle du conduit vocal serait alors limité aux changements de configuration de voyelles. Nous avons donc tiré avantage de notre habileté naturelle d'écriture pour connecter les caractéristiques de flux glottique à seulement trois dimensions de la tablette

graphique (axes x et y et la pression).

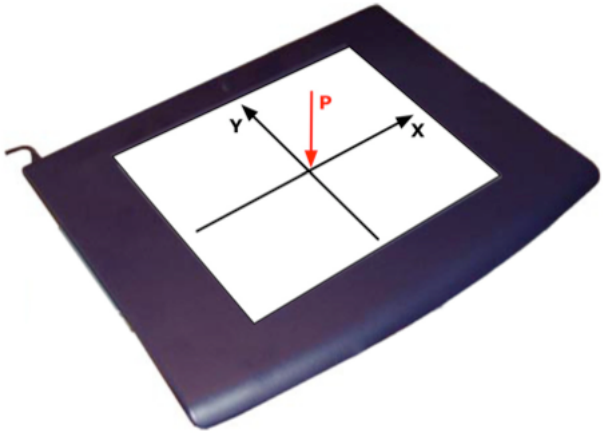


FIGURE 3.20 – *Les trois degrés de liberté de la tablette graphique (D'Alessandro et al., 2006)*

Comme décrit sur la 3.20, l'axe horizontal est mappé à la fréquence fondamentale. Les tests ont démontré qu'après un court entraînement, deux ou trois octaves peuvent être gérées sur la tablette graphique. D'autre part, des caractéristiques de mise à l'échelle de la transposition ont été implémentées. L'axe vertical, quant à lui, contrôle à la fois la dimension tendu/relâché et l'effort vocal. Le mapping est réalisé en utilisant la valeur de y comme un facteur d'interpolation entre les deux configurations différentes des paramètres O_q, α_m et T_l depuis une voix douce jusqu'à une voix tendue (cf. figure 3.21). Enfin, le paramètre de pression est mappé au gain E.

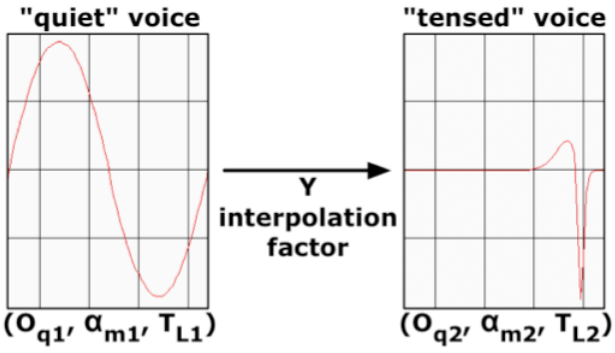


FIGURE 3.21 – *Illustration du passage d'une voix relâchée à tendue (D'Alessandro et al., 2006)*

La régression du contrôle de qualité vocale sur un axe expressif global rend les manipulations de la source vocale possibles par des simples "dessins" (i.e. des formes bi-dimensionnelles + pression). Ce compromis rend cet instrument véritablement intuitif. En fait, comme pour une guitare, l'interprète

n'a besoin que d'une tablette graphique pour jouer. Le contrôleur MIDI (par exemple un pédalier) est uniquement utilisé pour changer de configurations (cf. figure 3.22)

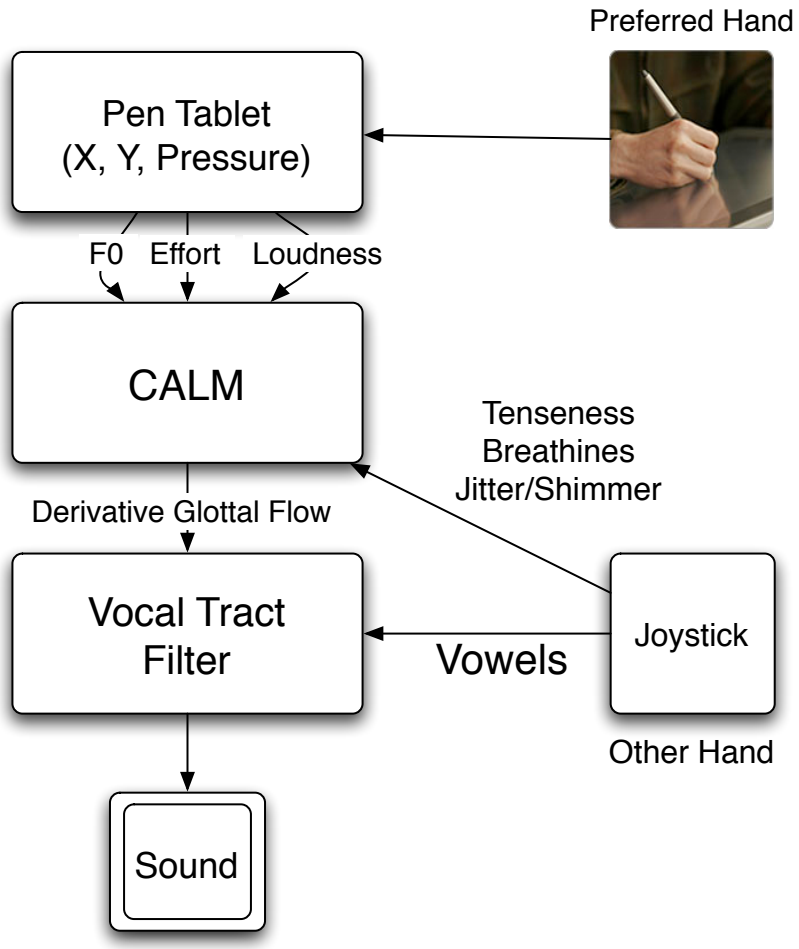


FIGURE 3.22 – *Vue schématique de l'instrument CALM avec tablette et joystick. (d'Alessandro et al., 2006)*

Ce dernier type d'instruments utilise les gestes d'écriture grâce à une tablette graphique. Il est également implémenté dans l'environnement Max/MSP (Max/MSP, 2008). La tablette graphique est conçue selon deux dimensions spatiales planes et une dimension de pression. L'axe X correspond à l'axe mélodique. Il est organisé de gauche à droite pour aller des basses aux aigus de la même manière que pour un clavier musical ou une guitare. L'axe Y correspond à l'effort vocal. Les dimensions d'effort vocal et de pression vocale sont mappés du bas (piano) en haut (forte) selon cet axe. La pression du stylet contrôle le volume global de la voix. Cet instrument se révèle étonnement facile à jouer et expressif. De nombreux effets vocaux sont possibles, comme le *vibrato, portamento, messa di voce, staccato, legato*.

3.4.3 Méta-instrument

Dans le cadre du projet ANR 2PIM, et de l'encadrement du stage de Y. Gaffary, nous avons eu l'opportunité d'explorer les possibilités offertes par un contrôleur comprenant un nombre de capteurs importants, afin de gérer les différents paramètres du synthétiseur CALM. Avant de décrire les spécificités de cette implémentation, nous allons détailler les caractéristiques de ce contrôleur instrument.

Description du Méta-instrument

Le Méta-Instrument est un instrument de musique assistée par ordinateur conçu par Serge de Laubier (de Laubier and Goudard, 2006) et développé par les studios Puce Muse [11]. Trois générations de Méta-instruments ont été développées, la première datant de 1989 ($\mathrm{MI1}$), la seconde de 1996 ($\mathrm{MI2}$) et la troisième et dernière de 2004 ($\mathrm{MI3}$). Nous ne détaillerons ici que les caractéristiques du $\mathrm{MI3}$, qui est la seule version que nous avons utilisée. La figure 3.23 suivante illustre le $\mathrm{MI3}$.

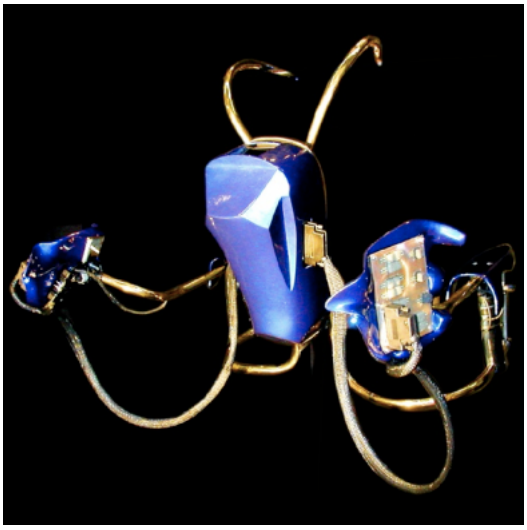


FIGURE 3.23 – *Le Méta-instrument*

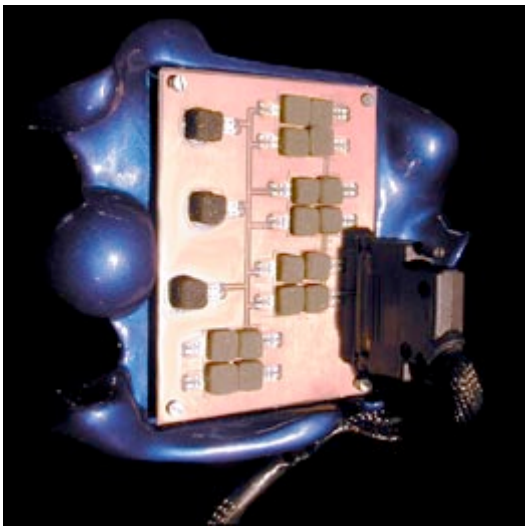


FIGURE 3.24 – *Détails des capteurs de la main droite*

Comme on peut le voir sur la figure de gauche, le Méta-instrument se présente sous la forme d'une armature qui se porte sur le ventre, les deux arceaux supérieurs étant soutenus par les épaules de l'interprète. Les deux bras reposent quant à eux sur les armatures mobiles prévues à cet effet. Les mains vont alors se loger sur les plate-formes dont la figure de droite détaille la disposition.

11. site Web : PuceMuse (http ://pucemuse.com/)

Les capteurs

Le pouce de la main droite est placée sous cette plate-forme, où se trouvent quatre capteurs de pression ainsi qu'une glissière. Sur le dessus de la plate-forme, chaque doigt dispose comme pour le pouce de quatre capteurs de pression, plus un supplémentaire sur lequel repose la phalange (sauf pour le petit doigt). En outre, chaque bras possède 3 degrés de liberté, un mouvement horizontal et un mouvement vertical au niveau du coude et un dernier mouvement rotatif ou niveau du poignet.

On peut donc ainsi distinguer trois catégories de capteurs :

- Les bras et les poignets permettent, en fonction de la force exercée, de contrôler un paramètre assez précisément et de façon naturelle, en ouvrant les bras, en les levant ou en tournant les poignets. Il existe un effort statique de ces mouvements lié à l'armature du Méta-instrument.
- Les capteurs de pression des doigts, les plus nombreux, sont difficiles à contrôler de manière précise, car il n'est pas aisé et assez fatigant de maintenir une pression constante sur ces capteurs. Toutefois, ces capteurs restent suffisamment intuitifs et permettent avec un certain apprentissage d'exprimer sa virtuosité pour l'instrument. En outre, différentes manières de tirer parti de ces capteurs existent, comme nous le verrons un peu plus loin.
- Les potentiomètres, un peu difficiles d'accès et en concurrence avec les capteurs de pression des pouces et la rotation du poignet, permettent de modifier précisément un paramètre qui ne nécessite pas de changement trop rapide ou fréquent.

Le Méta-instrument fournit ainsi 27 capteurs continus pour chaque bras, soit 54 capteurs au total. Ceux-ci sont connectés à une interface Ethersense dont les valeurs sont envoyées à un ordinateur grâce à une connexion Ethernet, avec une résolution maximale de 16 bits à une fréquence de 500 Hz. La précision de la mesure des rotations des poignets est réalisée au $1/20$ de degré et la pression au $1/10$ de gramme.

L'environnement 2PIM

Le studio PuceMuse fournit aux utilisateurs du Méta-instrument un environnement logiciel programmé avec Max/MSP (Max/MSP, 2008) permettant facilement de connecter un ou plusieurs Méta-instruments. Cet environnement 2PIM, développé dans le cadre du projet ANR du même nom, permet notamment de réaliser la calibration du Méta-instrument, de créer des profils et des instruments.

Concernant la calibration, il est possible notamment de réduire la résolution des capteurs, afin d'améliorer la fluidité de la captation (diminution de la bande passante active). En général, une résolution de 13 bits est largement suffisante pour effectuer le contrôle des paramètres de synthèse. En outre, une calibration logicielle permet de gérer indépendamment la réponse de chacun des capteurs, grâce à un gain, un offset, une valeur minimale et une valeur maximale. En effet, au fur et

à mesure de l'utilisation du Méta-instrument, certains capteurs peuvent être plus ou moins bruités ou dériver de leur caractéristiques originelles. Or, comme toutes les données des capteurs sont ensuite toutes normalisées entre 0 et 1 en valeur flottante, il est souhaitable de pouvoir utiliser toute l'étendue de contrôle et ainsi pouvoir intervertir aisément les correspondances capteurs-paramètres, sans perte de dynamique.

La plate-forme $2\mathrm{PIM}$ constitue à la fois un moyen de standardisation et de personnalisation des différents développements logiciels effectués pour le Méta-instrument. Une standardisation d'abord, car l'environnement $2\mathrm{PIM}$ fournit aux utilisateurs une plate-forme identique concernant la partie matérielle et logicielle, grâce aux différentes calibrations, normalisations, émulations (il est possible d'utiliser la $2\mathrm{PIM}$ sans Méta-instrument connecté). Une personnalisation ensuite, car chaque utilisateur peut facilement porter ses propres instruments sur n'importe quel ordinateur. L'utilisateur peut créer son propre profil, et ainsi intégrer ses propres instruments logiciels au sein de la plate-forme logicielle $2\mathrm{PIM}$. Cette intégration est facilité non seulement par la normalisation unitaire de tous les capteurs, mais également grâce à des abstractions Max/MSP fournissant les connexions nécessaires pour récupérer facilement les valeurs des capteurs et également en sortie pour se connecter au système audio dont l'utilisateur dispose (le plus généralement une carte audio stéréo).

Ensuite, l'utilisateur peut à loisir développer son propre instrument, dont les entrées de contrôle seront les valeurs normalisées des différents capteurs du Méta-instrument, et les sorties celles de la synthèse audio réalisée par l'instrument considéré [12]. Dans la prochaine section, nous allons décrire plus précisément les spécificités de l'instrument CALM développé pour le Méta-instrument.

Méta-CALM

L'instrument CALM développé pour le Méta-instrument reprend essentiellement les caractéristiques du synthétiseur CALM décrit précédemment, en particulier concernant le moteur de synthèse constitué avec les objets externes `almPulse~` et `stFilter~`, ainsi que le mapping entre les dimensions vocales et les paramètres de bas niveau du synthétiseur.

Le souci, ici, était donc principalement de profiter des nombreux capteurs disponibles du Méta-instrument pour contrôler les dimensions vocales. L'application visée était plutôt celle de la voix chantée, de part la nature de ce contrôleur. Certaines adaptations ont alors du être apportées pour la réalisation de cet objectif.

En premier lieu, le contrôle mélodique est très difficile a réaliser, voire impossible s'il est effectué directement avec un (ou plusieurs) capteurs de pression. En effet, malgré leur résolution, les

12. Notons toutefois, que la plate-forme a été conçue dans le souci de pouvoir également créer des moteurs de synthèse graphique et vidéo. La carte graphique de l'ordinateur devient alors l'interface de sortie du rendu graphique.

capteurs de pression sont recouverts d'une mousse synthétique dont la course totale est de l'ordre du centimètre. Il n'est donc pas imaginable de pouvoir atteindre précisément une note souhaitée.

Le fonctionnement privilégié est donc celui d'un séquenceur, où la succession des différentes notes de la partition est stockée dans un fichier texte pouvant être chargé dans l'environnement $2\mathrm{PIM}$. Ainsi, l'interprète peut aisément jouer une partition donnée par déclenchement successifs des notes MIDI. Ceci est donc effectué de la manière suivante : l'utilisateur sélectionne un fichier de partition et le programme charge alors la liste des notes MIDI successives. Par pression de l'index gauche sur le capteur correspondant du Méta-instrument, les notes sont jouées dans l'ordre du fichier. Notons qu'un bon nombre de compositions écrites pour le Méta-instrument utilise également ce formalisme, pour la raison expliquée ci-dessus.

Concernant les capteurs de pression, la description fournie précédemment nous montre que pour chaque doigt, quatre capteurs de pression sont disponibles, il est donc possible de tirer parti de cette configuration pour la réalisation de différentes configurations, de la manière suivante :

- Correspondance directe un à un : la pression exercée sur la capteur modifie directement un paramètre. Cette méthode est simple à mettre en œuvre mais présente l'inconvénient qu'il est nécessaire de maintenir la pression pour garder la valeur du paramètre.
- Moyenne de plusieurs capteurs de pression. Etant donné la taille des capteurs de pression, chaque doigt est en contact avec au minimum deux capteurs. Il est donc possible de réaliser la moyenne entre deux ou quatre capteurs, afin d'obtenir une précision accrue.
- Une méthode alternative à celle précédemment citée, consiste à considérer la position d'un doigt selon un axe gauche-droite ou haut-bas, en réalisant une moyenne deux à deux des capteurs. Il est également possible de calculer le barycentre des pressions, en voyant la position du doigt selon un carré.
- Même si la résolution des capteurs est élevée, il peut parfois être nécessaire de faire fonctionner les capteurs comme des boutons $0/1$ pour activer différentes configurations, comme c'est le cas pour les mécanismes $\mathrm{M}1$, $\mathrm{M}2$ par exemple. En revanche, le fait que ces capteurs ne soient pas binaires permet alors de fixer un seuil, suivant que l'on souhaite que le passage soit déclenché par une pression faible ou élevée.
- Enfin, il est possible d'utiliser les capteurs, non plus comme une simple valeur binaire, mais un compteur incrémental ou décrémental. Ce fonctionnement est par exemple utile pour réaliser le séquençage des notes. Cette correspondance présente l'avantage que l'on contrôle implicitement le rythme de déroulement de la partition.

En plus de ces correspondances individuelles, il est évidemment possible de combiner n'importe quel fonctionnement d'un capteur avec un autre. Comme par exemple, le fait de considérer que l'effort vocal n'aura d'effet uniquement si une note a été déclenchée.

De prime abord, le Méta-CALM se révèle relativement difficile à contrôler, essentiellement à cause du nombre de paramètres à contrôler simultanément. Toutefois comme le notent Wanderley

et Depalle (Wanderley and Depalle, 2004), certains auteurs considèrent que les instruments qui sont les plus difficiles à manoeuvrer et donc à jouer, sont également ceux qui offrent le plus de possibilités en termes d'expressivité. Aussi, nous restons assez confiants quant à l'utilisation du Méta-instrument comme contrôleur pour le synthétiseur CALM grâce à des "méta-interprètes" plus chevronnés que nous pouvons l'être pour l'instant.

Le mapping

Le mapping du Méta-CALM est effectué de la manière suivante : la main gauche contrôle principalement les paramètres relatifs à F_0, tandis que le contrôle des paramètres des filtres résonants est réalisé par la main droite.De façon plus détaillée, nous avons :

- Main gauche
    1. L'index contrôle le déclenchement des notes de la mélodie contenu dans le fichier sous la forme d'une série de valeurs MIDI.
    2. L'effort vocal est géré par deux contrôles : le premier par la position verticale du bras gauche lors de l'attaque de note. Ensuite, lorsque la note est déclenchée, le majeur permet de modifier la valeur initiale.
    3. L'annulaire contrôle la fréquence du vibrato, tandis que la rotation du poignet contrôle sa profondeur.
    4. L'auriculaire permet de stopper la note.
    5. Le volume global de la synthèse est connecté au potentiomètre du pouce.
    6. Le capteur de pression du pouce contrôle le degré du souffle de la voix.
- Main droite
    1. La navigation dans le triangle vocalique est effectué à la fois grâce aux phalanges qui contrôlent F_1 et la rotation du poignet, qui elle, contrôle F_2.
    2. Les amplitudes des filtres A_1, A_2, A_3, et A_4 sont respectivement contrôlées par deux capteurs verticaux de l'index et du majeur.
    3. L'annulaire permet de recharger les notes MIDI, lorsque l'on est arrivé en fin de partition.
    4. L'amplitude E de l'ODGD, peut être fixée directement par le potentiomètre ou modifiée par l'auriculaire.
    5. Le changement de mécanisme est géré par le capteur de pression du pouce.

Le fonctionnement typique du Méta-CALM est donc le suivant : on déclenche grâce à l'index droit chaque note. La position du bras définit l'effort vocal initial de la note. Une fois la note déclenchée, il est alors possible d'y ajouter un certain vibrato, de modifier la voyelle prononcée, de moduler l'effort vocal (ce qui correspond grossièrement à un tremolo), ou enfin d'éteindre la note avant d'en déclencher une autre.

Actuellement, le Méta-CALM ne dispose pas de contrôle explicite de l'enveloppe d'amplitude pour chaque note, ce qui rend l'articulation quelque peu artificielle. Il y aurait donc une certaine amélioration a apporter dans ce domaine. En outre, bien que le Méta-CALM soit essentiellement dédié à une application de voix chantée, la description donnée ci-dessus nous montre qu'il est également possible de contrôler toutes les dimensions de qualité vocale, au même titre qu'avec la tablette graphique. Il serait donc intéressant d'explorer la capacité du Méta-instrument à contrôler la qualité vocale, comparativement à la tablette graphique.

La figure 3.25 suivante résume les différentes correspondances de paramètres utilisées pour l'instrument Méta-CALM.

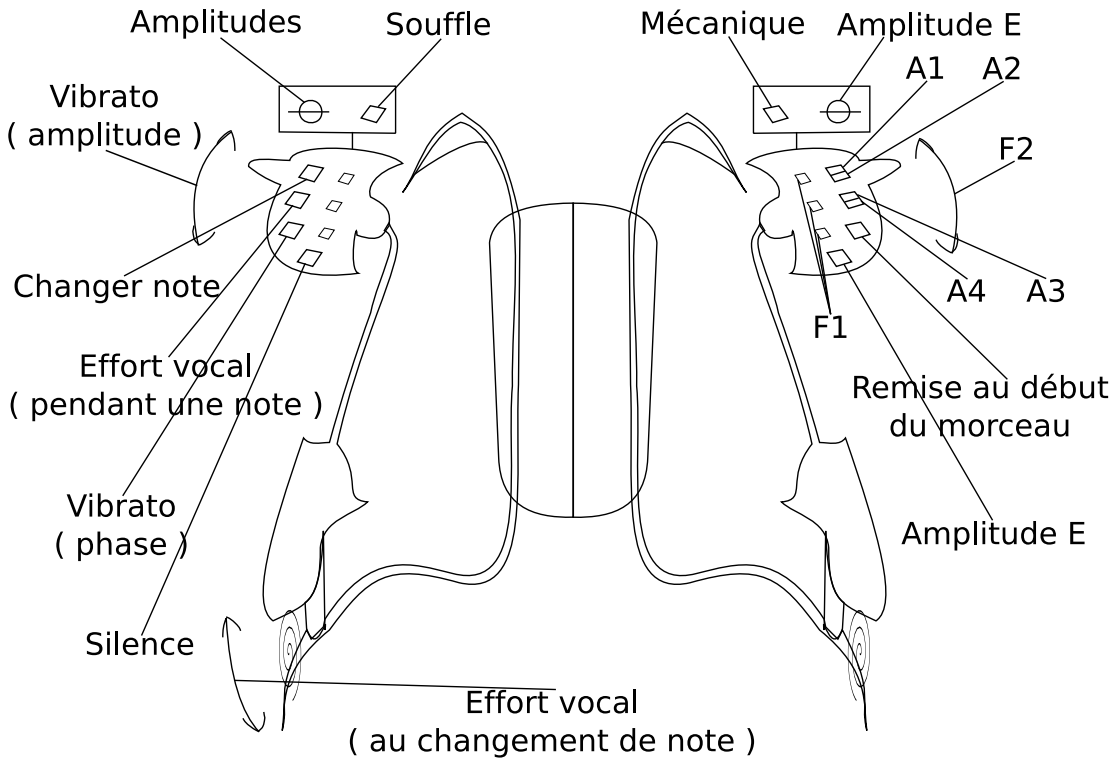


FIGURE 3.25 – *Les correspondances des capteurs du Méta-instrument avec les paramètres du synthétiseur CALM.*

3.4.4 Exploration haptique du phonétogramme

Nous avons eu récemment l'opportunité, grâce au stage d'un autre étudiant N. Cornuet, d'explorer les possibilités offertes par un contrôleur haptique pour le synthétiseur RT-CALM. Le principal but de cette étude était d'observer si le retour haptique permettait, grâce à la proprioception, d'établir un lien

pertinent entre le phonétogramme et la force renvoyé par le bras haptique. En effet, nous avions pu remarquer qu'une des limitations de la tablette graphique était liée aux changements de mécanismes.

Le changement de mécanisme modifie brusquement les caractéristiques laryngées, mais ne constitue par pour autant un phénomène incontrôlable. Les chanteurs entraînés parviennent en effet à adoucir, voire masquer le changement de registre. Or l'arrivée de ce changement est difficilement détectable en synthèse, à la simple écoute de la hauteur du son. Concrètement, avec la tablette graphique, le changement se produit lorsqu'une certaine hauteur est atteinte. On pourrait se dire qu'il suffit de tracer une ligne sur la tablette pour pouvoir visualiser la frontière de ce passage. Cependant, cette hauteur dépend de facteurs à la fois intra et inter individuels. Pour une même personne, ce passage n'est pas réalisé à la même hauteur suivant qu'il s'agit d'un glissando ascendant ou descendant. En outre, pour différentes glissandi ascendants, cette hauteur frontière peut varier, même s'il l'on peut définir une moyenne statistique pour un locuteur donné. De manière peut-être plus primordiale, c'est surtout l'ensemble du phonétogramme, c'est à dire l'ensemble des hauteurs et volumes atteignables par un individu, qui est modifiée d'un individu à l'autre.

L'idée d'utiliser un bras haptique à la place d'une tablette graphique était donc d'observer si le retour de force nous permettait de sentir non seulement les frontières de phonétogramme, c'est-à-dire là où la production vocale s'éteint, mais également le changement de mécanisme lors de glissandi.

Le bras haptique

Le bras haptique que nous avons utilisé est le PHANTOM Omni développé par la socitété SensAble [13], dont on peut voir une illustration sur la figure 3.26 suivante.

13. Site Web : SensAble Technologies

FIGURE 3.26 – *Le bras haptique PHANTOM Omni.*

Les principales caractéristiques du bras PHANTOM, sont les suivantes :

- 6 degrés de liberté : les trois dimensions de l'espace $(\mathrm{x}, \mathrm{y}, \mathrm{z})$ et leurs moments respectifs (roulis, tangage, lacet).
- Deux boutons supplémentaires situés sur le stylet.
- Communication grâce à un port Firewire, avec la possibilité de chaîner deux bras sur un même canal Firewire.
- Résolution spatiale de l'ordre de $0.05\ \mathrm{mm}$.
- Espace de travail de dimensions : $16 \times 12 \times 7\ \mathrm{cm}$.
- Retour de force sur les trois positions spatiales.

La communication logicielle

Pour les besoins de programmation avec le bras PHANTOM, SensAble fournit également un utilitaire de développement logiciel (SDK[14]) sous la forme de librairies $\mathrm{C++}$ permettant une intégration simplifiée du bras haptique avec tout type d'appplication.

Grâce à ce SDK, deux objets externes Max/MSP ont été développés, pour permettre d'une part d'établir la communication avec le bras haptique et récupérer les valeurs de position du bras dans l'espace et d'autre part pour envoyer des valeurs de forces au bras suivant la position du bras dans le phonétogramme. Ces deux objets externes sont disponibles à l'heure actuelle pour Max/MSP sous Mac OSX et Windows.

14. Software Development Kit

Le phonétogramme

Les objets ont été développés dans le souci de pouvoir s'adapter à n'importe quel locuteur. C'est-à-dire qu'il est possible de charger un nouveau phonétogramme, sous la forme d'un fichier texte comportant sur chaque ligne un couple de valeur (F_0 en Hz, A en dB). Cette caractéristique importante nous permet ainsi de pouvoir nous adapter assez facilement à n'importe quel type d'enregistrements de phonétogramme réalisé. Lors de la réalisation d'un phonétogramme, il est en général demandé au sujet-locuteur de produire une voyelle à une fréquence donnée, avec un effort vocal (et par extension une intensité) le plus faible possible pour aller vers les intensités les plus fortes que le locuteur est capable d'accomplir. Pour réaliser cette tâche, le locuteur dispose parfois d'une interface graphique sur un ordinateur pour voir en temps réel l'évolution de l'intensité produite et ainsi vérifier par la même occasion que la fréquence reste la plus fixe possible. En définitive, une fois réalisé, le phonétogramme est constitué d'un ensemble de lignes et/ou de droites plus ou moins espacées à l'intérieur duquel va se trouver le phonétogramme du locuteur considéré.

Ainsi, le fichier texte que nous utilisons définit la zone englobant le phonétogramme, de telle manière que les points (F_0, A) définissent les différentes sommets du polygone ainsi formé.

La visualisation

Le développement accompli jusqu'à aujourd'hui permet de pouvoir visualiser les frontières du phonétogramme et des différents mécanismes, ainsi que la position du stylet du bras haptique dans ce phonétogramme. Même si cette visualisation n'est pas indispensable, elle rend la manipulation bien plus confortable. En effet, en gagnant sur le fait que l'on est désormais capable de situer précisément le changement de mécanisme et les limites du phonétogramme grâce à des frontières bien définies grâce à au retour d'effort du bras haptique, on perd sur la position relative entre ces frontières. L'espace physique définit par le phonétogramme est ici en effet virtuel.

Ainsi, la présence d'un retour visuel aide grandement le déplacement au sein du phonétogramme. Toutefois, une solution alternative existe et consiste à placer une feuille sur la table devant le bras haptique, sur lequel est représenté au choix, soit les différentes limites de fonctionnement des deux mécanismes, soit une représentation plus mélodique avec des notes si l'on vise plutôt une application de voix chantée. La figure 3.27 suivante illustre la visualisation du phonétogramme dans Max/MSP à partir de deux fichiers de phonétogramme pour $M1$ et $M2$.

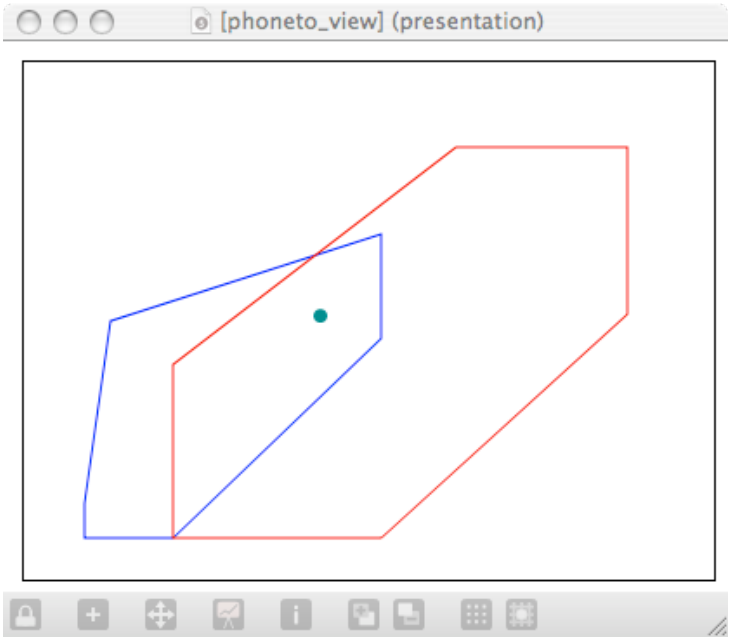


FIGURE 3.27 – *Visualisation dans Max/MSP des mécanismes* M1 *(en bleu) et* M2 *(en rouge), avec la position du stylet du bras haptique dans l'espace.*

Le mapping

Le mapping réalisé ici avec le bras haptique est plus simple que pour le Méta-CALM, au détriment qu'il ne permet pas de contrôler autant de paramètres de synthèse et de dimensions vocales. Par contre, il permet de se focaliser sur les aspects mélodiques et de force de la voix.

Ainsi, comme il paraît désormais clair au vu des explications précédentes, le contrôle de $\mathrm{F_0}$ et de l'effort vocal est en correspondance directe avec la position du stylet du bras haptique dans le phonétogramme virtuel. Comme avec la tablette graphique, le contrôle du vibrato peut être réalisé grâce à un déplacement cyclique du stylet à l'aide du geste manuel, sans avoir recours à une fonction analytique externe. Enfin, afin de pouvoir exécuter des notes détachées, l'un des boutons du stylet permet d'activer et désactiver la sortie audio.

Etant donné que lors d'un déplacement en fréquence important le mécanisme mis en jeu est amené à changer, il est toutefois nécessaire d'attribuer, pour chaque mécanisme des valeurs fixes moyennes pour $\mathrm{O_q}$ et α_m selon le mécanisme considéré. En opérant ainsi, on perçoit alors lors du changement de mécanisme une qualité vocale qui est modifiée de façon abrupte. En outre, du fait des forces exercées par le bras haptique, au voisinage du passage de mécanisme, le stylet est attiré vers l'autre mécanisme. Si bien, que le saut en fréquence est également réalisé simplement par le fait qu'il est difficile de positionner de manière stable le stylet au centre du passage de mécanisme.

En ce qui concerne le contrôle des filtres formantiques, une simple interface graphique permet de se déplacer dans le triangle vocalique grâce à la souris, modifiant ainsi F_1 et F_2 directement. Le calcul de F_3 et F_4 est quant à lui réalisé par pondération, à partir d'une table. Une matrice contient en effet les valeurs des positions des formants, de leurs amplitudes et de leurs bandes passantes respectives pour une douzaine de voyelles. Ces valeurs sont issues de l'analyse spectrale d'un chanteur lyrique. Suivant la position à laquelle on se trouve dans le triangle vocalique, une pondération barycentrique est effectuée avec tous les points présents dans la table, afin de fournir des valeurs pour F_3 et F_4 suffisamment proches de la réalité.

Il est également imaginable d'utiliser en adjonction du bras haptique une autre interface comme une tablette ou un joystick pour permettre de contrôler le reste des dimensions vocales et le triangle vocalique, pour peu que la charge logicielle engendrée ne soit pas trop importante.

3.4.5 Réflexions sur l'adéquation interface/synthétiseur

L'utilisation d'interfaces homme-machine pour le contrôle de la synthèse pose un certain nombre de questions, mais également de problèmes quant à l'évaluation de tels systèmes. En effet, de part les caractéristiques relativement différentes des capteurs (et actionneurs) utilisés, il est très difficile, voire impossible de pouvoir réaliser une évaluation objective de ces différentes interfaces.

Quelques études existent cependant dans ce domaine, qui, à ce titre, ont cherché à caractériser le nombre de degrés de liberté du système (Wanderley and Depalle, 1999), la classification des types de capteurs utilisés (Marshall and Wanderley, 2006; Wanderley et al., 2006), ou encore l'espace de dimensions couvert par l'interface utilisée (Birnbaum et al., 2005). Toutefois, toutes ces études, aussi intéressantes soient-elles, ne nous renseignent que partiellement sur l'efficacité d'un système par rapport à un autre pour accomplir une tâche donnée.

Il reste alors les évaluations subjectives, car ce que traduit la plupart des systèmes développés c'est fondamentalement qu'il n'existe pas de bonne ou de mauvaise interface en soi. Il est uniquement possible de statuer sur le fait qu'une interface sera mieux adaptée à un type d'application ou un synthétiseur particulier. Et cette adéquation peut être abordée, selon deux questions essentielles : (i) est-ce que l'interface est intuitive, agréable à utiliser ? (ii) est-ce que le résultat sonore obtenu est satisfaisant, naturel, expressif ? (ou par extension, offre-t-il de nouvelles sonorités ?)

Au risque de paraître un peu caricatural, on pourra dire que la première question sera principalement relative à la conception physique de l'interface, aux types de capteurs utilisés, et de leur disposition vis-à-vis de l'anatomie humaine. La seconde question sera, quant à elle, plus liée aux types de processus utilisés pour le mapping : le nombre de paramètres qu'il est possible de contrôler, l'utilisation d'un espace perceptif adéquat pour l'application visée, l'utilisation d'un contrôle haut niveau/bas niveau ...

En outre, du point de vue de l'évaluation, la première question fera appel à une évaluation par l'utilisation de l'interface par des sujets pour une tâche déterminée, dépendant de l'application souhaitée, et les questions possibles aux sujets feront référence notamment à la facilité d'utilisation, l'ergonomie, la *jouabilité* ... La seconde question consiste pour sa part, plutôt à réaliser des stimuli que l'on juge suffisamment ressemblant, ou expressifs, et de demander à des auditeurs une classification de ces stimuli, une évaluation du degré de ressemblance avec des stimuli naturels ... Dans tous les cas, l'évaluation réalisée sera largement dépendante du but applicatif visé.

Dans le cadre de notre étude, il est assez évident que notre intérêt était plutôt axé sur la volonté de répondre à la deuxième question. En tout cas, les choix d'interfaces que nous avons réalisés sont en grande partie liés à notre propre expérience vis-à-vis des systèmes utilisés. Aussi, l'utilisation d'une tablette graphique pour le contrôle du synthétiseur CALM a fait suite à des essais avec notamment un clavier MIDI ou un gant de données, qui ne nous donnaient pas entière satisfaction quant à l'ergonomie, principalement. Pour le gant de données, par exemple, le fait qu'il n'existe pas de repère dans l'espace lors de la manipulation ne permet pas de contrôler la dimension mélodique avec la précision requise suffisante. Notre orientation vers la tablette graphique s'est ainsi faite naturellement par notre propre utilisation des différents systèmes.

Concernant les applications ou les tâches réalisables avec notre synthétiseur, deux voix ont été privilégiées : la réalisation d'une synthèse de voix chantée, et la réalisation d'éclats affectifs [15] faisant appel à la qualité vocale.

3.5 Applications

La synthèse de source glottique et son contrôle par des interfaces gestuelles peut se traduire en un certain nombre d'applications, au nombre desquelles et au vu de notre expérience avec le synthétiseur CALM, nous pouvons citer la synthèse de voix chantée, l'exploration de la qualité vocale en recherche, et également la possibilité d'utiliser notre outil pour l'apprentissage de la phonétique.

3.5.1 Voix chantée

Comme nous avons pu le voir au cours de ce chapitre, une bonne partie de notre développement du synthétiseur a été réalisé autour de la synthèse de voix chantée. Cette application a pu nous fournir des résultats intéressants dès les premières implémentations. Toutefois, il nous reste certaines améliorations à apporter afin de pouvoir disposer d'un synthétiseur idéalement contrôlable.

15. affect bursts

L'une des premières pistes à explorer reste celle des études en analyse de voix chantée. Le résultat de synthèse actuel reste relativement impersonnel. A ce titre, l'élaboration d'un conduit vocal de synthèse plus évolué serait nécessaire, non seulement parce qu'il permettrait de pouvoir donner une *couleur* plus marquée à la synthèse, mais également parce que l'utilisation de filtres résonants pour le filtrage pose problème lorsque la fréquence fondamentale de l'onde de débit glottique augmente et entre alors en résonance avec ces filtres, de façon indésirable. Ce filtre de conduit vocal permettrait en outre la réalisation de sons nasalisés pour peu qu'un deuxième filtrage soit implémenté en parallèle.

3.5.2 Synthèse de qualité vocale et génération de stimuli

L'une des autres particularités intéressantes de notre outil de synthèse repose sur le fait qu'il constitue un des rares exemples de synthèse de qualité vocale, avec notamment la prise en compte des apériodicités de la voix. Peu de synthétiseurs permettent en effet, à notre connaissance, la possibilité de contrôler finement et en temps réel, les paramètres de source glottique et notamment le bruit additif et structurel. Cette particularité se révèle particulièrement intéressante dans le but de synthétiser certains types de phénomènes en parole, comme la présence de souffle en fin de phrase dans de nombreuses langues ou l'utilisation d'une qualité de voix pressée ou soufflée pour certaines langues, comme le japonais.

Le japonais constitue en effet un exemple particulièrement intéressant pour approfondir cette étude puisque, contrairement aux langues européennes en général, la qualité de voix constitue un trait prosodique distinctif pour certaines attitudes. Cet état de fait constitue donc un point de départ intéressant pour permettre d'évaluer la capacité de nos outils de synthèse de source glottique à reproduire avec suffisamment de finesse ces différentes qualités de voix pour pouvoir être distinguées. En outre, les études montrent que les locuteurs de langue maternelle japonaise sont capables d'identifier avec suffisamment de certitude ces différentes attitudes uniquement lors d'éclats affectifs, autrement dit avec de simples interjections ne nécessitant qu'une voyelle. Il paraît donc tout à fait imaginable de pouvoir reproduire ces interjections attitudinales grâce à notre synthétiseur.

3.5.3 Apprentissage phonétique

Nous nous sommes rendu compte lors de la présentation de notre outil à des enseignants en logopédie ou phonétique que ce système suscitait un intérêt pour l'apprentissage des sons de la parole et notamment pour découvrir par l'écoute ce qu'est une voix soufflé, pressée, tendue ... selon la terminologie utilisée par les phonéticiens. Non seulement, cela permet à un étudiant de pouvoir se rendre compte auditivement des descriptions anatomiques utilisées couramment en phonétique, mais également, cette découverte peut être réalisée de manière interactive pour se rendre compte

des effets indépendants des différentes dimensions vocales sur la production de parole.

Chapitre 4

Discussion et Perspectives

Sommaire

4.1 Discussion **160**
4.1.1 Modification prosodique 160
4.1.2 Synthèse de source glottique 162
4.2 Perspectives **164**
4.2.1 Evaluation objective de la modification de durée 164
4.2.2 Evaluation des attitudes japonaises 165
4.2.3 Intégration facilitée des interfaces 165
4.2.4 Applications possibles 166
4.2.5 Objectifs à plus long terme 167

4.1 Discussion

4.1.1 Modification prosodique

D'un point de vue théorique, l'approche que nous avons privilégiée pour la modification de la hauteur et de la durée comporte un avantage indéniable : cette méthodologie ne réalise aucune inférence a priori sur la manière dont est modélisée la prosodie de la phrase. Ce faisant, notre approche se situe bien dans une perspective d'analyse par la synthèse, ce qui nous permet d'aborder sous un nouvel angle la modélisation prosodique.

Concernant la modification de hauteur, nos expérimentations successives nous ont permis de valider la possibilité de reproduire une intonation cible, selon différentes modalités (gestuelle et vocale), à partir d'une phrase donnée dont la hauteur était constante. Et cela nous a permis de tirer quelques précieux enseignements. En premier lieu, la modification prosodique par le geste constitue une bonne méthode pour la reproduction intonative, avec des performances obtenues par la modification gestuelle comparables à celles de l'imitation vocale. Deuxièmement, étant donné que les mouvements microprosodiques ne peuvent pas être reproduits par les mouvements gestuels, car trop rapides, les données gestuelles représentent une bonne stylisation de l'intonation, de manière continue et permettent ainsi d'envisager une modélisation dynamique de l'intonation, et de la prosodie au sens large. Troisièmement, les analyses gestuelles effectuées sur les enregistrements des sujets nous révèlent que ceux-ci ont fait appel à leur entraînement issu de l'écriture pour reproduire les intonations cibles. Enfin, il apparaît, à la suite des questionnaires réalisés auprès des sujets, que l'entraînement musical joue un rôle non négligeable sur les performances, mais non pas comme on pouvait s'y attendre dans sa dimension auditive, mais plutôt selon la dimension instrumentale, c'est à dire gestuelle.

Concernant la modification du rythme ou, plus modestement, de la durée, la validation de la possibilité de modifier précisément le déroulement temporel d'une phrase reste encore à réaliser soigneusement. Cependant, les quelques exemples qui ont pu être réalisés par nos soins pour la réalisation de différentes attitudes du français, nous montre qu'il est largement possible d'obtenir des phrases possédant une attitude donnée, qui n'est pas celle de la phrase originale.

Le fait que nous n'ayons pas pu mener d'expérimentations objectives pour la validation de la modification de durée, est essentiellement liée au fait que le cadre de travail dans lequel se place le rythme en prosodie est fondamentalement différent de la modification de hauteur et comporte des problèmes intrinsèquement différents. Nous aurons l'occasion de voir dans les perspectives, des solutions acceptables quant à une future expérimentation à mener. Pour le moment, nous essaierons juste de relever la complexité du paradigme rythmique en comparaison de l'aspect mélodique.

Dans un premier temps, nous avions pensé qu'il serait possible de prendre comme référence

de base de notre étude des phrases isochrones. Cela dit, la réalisation d'une isochronie n'est pas aussi facilement implémentée que l'aplatissement de la mélodie d'une phrase. En outre, si l'on peut raisonnablement penser qu'une phrase avec une hauteur constante moyenne fournit une phrase relativement neutre, il n'en va pas de même avec une phrase isochrone. D'autre part, il est difficilement possible de considérer une modification absolue du rythme. Cette vision impliquerait que l'on dispose d'une valeur de vitesse d'élocution. Or, celle-ci dépend grandement du locuteur considéré ainsi que du contexte dans lequel la phrase est prononcée. Il convient également de se demander quel niveau de granularité est nécessaire pour effectuer une telle modification absolue (phone, diphone, syllabe, mot).

D'un point de vue expérimental, notre étude prosodique nous a amené à développer un outil de modification de la hauteur et de la durée simultanées, en temps réel strict. Comme nous avons déjà pu le souligner dans le chapitre concerné, il n'est pas possible de remonter le temps, et les modifications de durée réalisables sont contraintes non seulement par le temps réel mais également par la qualité synthétique de la phrase résultante. Toutefois, des modifications de qualité satisfaisante peuvent être obtenues d'une vitesse de déroulement moitié à double. Et ces limites paraissent largement raisonnables pour des applications de voix parlée, pour peu que la phrase originale ait été prononcée avec une vitesse d'élocution standard.

Notre outil permet, à l'heure actuelle de générer des résultats convenables, mais pouvant encore être améliorés. La principale amélioration souhaitable serait sans doute de pouvoir disposer d'une analyse des temps d'attaque (ou *p-centers*) permettant ainsi de disposer d'instants d'ancrage pour la modification de durée. Il serait alors possible de modifier le rythme de la phrase soit continûment, soit de manière constante entre deux instants d'attaques consécutifs. Ceci permettrait certainement d'étudier la modification de la durée, selon les différents niveaux de granularité évoqués ci-dessus, en lien avec l'analyse morphologique de la phrase.

La principale originalité de notre système réside néanmoins dans la possibilité de réaliser les modifications prosodiques grâce à une interface gestuelle. Et c'est essentiellement cette caractéristique qui nous a permis d'explorer la question de la modification de la hauteur et de la durée de la parole avec un regard nouveau. Les implications de ce nouveau moyen de synthèse sont nombreuses et nous n'avons pas la prétention d'avoir pu répondre à toutes les interrogations suscitées par cet outil. Malgré tout, nous avons pu observer au cours de nos expérimentations que la réalisation réaliste d'une mélodie donnée était en corrélation avec la pratique musicale. Mais non pas, comme on aurait pu s'y attendre a priori, du fait de l'oreille musicale entraînée des musiciens, mais plus grâce à l'expérience instrumentale. Il convient alors d'admettre en première approche qu'un apprentissage prolongé avec notre outil permettrait d'améliorer les résultats. Il résulte également de cette étude que toute personne sans déficience auditive, est capable de reproduire une mélodie donnée avec une précision suffisamment bonne.

Toutes choses égales par ailleurs, la qualité sonore obtenue par notre système de modification prosodique reste dans les limites acceptables qui sont celles de la synthèse PSOLA. Ce qui signifie qu'il est possible d'imaginer une modification prosodique, au sens hauteur et rythme combinées, par tout autre traitement de parole donnée, pour peu que celui-ci puisse être manipulé en temps réel. Car, c'est bien ici de la manipulation des paramètres de synthèse dont il s'agit et par extension de leur évaluation en termes de mouvements gestuels particuliers.

4.1.2 Synthèse de source glottique

La source glottique, ou de manière plus globale, le larynx est reconnu depuis longtemps pour être responsable des différences de qualité vocale pathologiques ou non, pour un locuteur donné ou entre différents locuteurs. Depuis quelques années seulement, la qualité vocale est aussi reconnue pour avoir une part non négligeable en prosodie, au delà de l'influence certes importante de l'intonation et du rythme.

La qualité vocale souffre néanmoins d'une difficulté technique encore importante aujourd'hui, à savoir celle de son analyse. Même si de nombreuses avancées significatives ont été réalisées ces dernières années dans ce domaine, il reste encore ardu d'extraire aisément et avec suffisamment de fiabilité les caractéristiques de la source vocale. Ces limitations sont liées à deux raisons principales. La première est celle de l'utilisation de "voix de laboratoire", dans la grande majorité des études. Les performances des outils d'analyse sont quasiment exclusivement testées sur des voix enregistrées dans des conditions idéales, ces mêmes performances étant liées en grande partie au niveau de bruit présent dans le contexte d'enregistrement. Par ailleurs, tous les paramètres de source glottique ne sont pas d'égale difficulté à extraire. Si les instants de fermeture glottique, le quotient ouvert et dans une moindre mesure le coefficient d'asymétrie sont assez bien estimés, il n'en va pas de même concernant le tilt spectral par exemple. Or, cette caractéristique vocale se révèle relativement importante en prosodie, puisque le tilt spectral est intimement lié à l'effort vocal, et par extension à l'intensité de la production vocale, représentant elle-même une dimension vocale d'importance (accentuation, forçage vocal, ...).

La seconde difficulté est quant à elle liée à la modélisation de la source vocale. Depuis de nombreuses années, les modèles de source glottique tels que le modèle LF ou KLGLOTT, ont été largement utilisés avec succès dans de nombreuses études. Cependant, ces modélisations de l'onde de débit glottique reposent fondamentalement sur la modélisation source-filtre linéaire de la production vocale. Or, on sait également, principalement grâce aux études aéroacoustiques de la source vocale, qu'une interaction existe entre la source de production vocale, à savoir la larynx et le conduit vocal. Ces interactions sont certes négligeables dans la plupart des cas, mais ne peuvent résolument pas être occultées à plus long terme. Ainsi, la séparation source-filtre réalisée, dans l'écrasante majorité des cas par la méthode de prédiction linéaire, ne nous assure en rien qu'une

partie de la contribution du conduit vocal ne soit encore présente dans le résiduel, après filtrage inverse (et vice-versa).

L'approche que nous avons modestement essayé d'adopter a donc consisté à se focaliser principalement sur les liens existants entre les paramètres des modèles de source glottique et les dimensions vocales associées. Nous n'avons pas essayé non plus de réaliser une modélisation articulatoire complète afin de nous concentrer sur les modes phonatoires de la source vocale. L'idée principale était donc de pouvoir reproduire quelques voyelles avec le plus de naturel et d'expressivité possibles, en se départant des contraintes d'intelligibilité.

L'idée sous-jacente était également qu'en approfondissant la notion de la qualité vocale, il serait toujours possible a posteriori de réintégrer cette synthèse dans un cadre de modification plus complet, avec une extraction de source suffisamment fiable. Nous avons donc concentré nos efforts d'une part sur la synthèse de voix chantée, et d'autre part sur la synthèse d'éclats affectifs, champs d'études pour lesquels l'intelligibilité est de moindre importance.

Du point de vue technique, notre implémentation s'est axée autour de trois directions principales. Tout d'abord, la réalisation du modèle CALM en temps réel. Différentes améliorations successives ont été nécessaires, pour aboutir à une solution directe échantillon par échantillon, qui comporte l'avantage de réduire la latence de la synthèse, et donc sa réactivité, et d'augmenter la fiabilité en terme de manipulation des paramètres.

Deuxièmement, une part importante du travail a consisté à réaliser des correspondances probantes entre les dimensions vocales perceptives, qui nous intéressent et les paramètres bas niveau du modèle CALM. Ces correspondances sont évidemment liées aux connaissances sur les plages de variation de ces paramètres et de leur effet sur la qualité vocale.

Troisièmement, nous avons essayé d'intégrer le couplage présent entre la hauteur de la voix et l'effort vocal, grâce à une modélisation de type phonétogramme. Notre souci était ici de pouvoir adapter, suivant les données analysées, les plages de variation en fréquence et en intensité. Une évaluation reste cependant encore nécessaire pour savoir si l'utilisation d'une interface haptique pour le contrôle de ces paramètres est pertinente ou non.

En outre, les applications pour de la voix chantée comportent un certain nombre de spécificités que nous avons essayé de prendre en compte, telles que les notes tempérées, le vibrato ou encore le formant du chanteur. Cependant, les modèles utilisés à l'heure actuelle restent rudimentaires, et nécessiteraient d'être approfondis pour être plus convaincants.

Par ailleurs, nous avons autant que possible, testé différentes configurations pour les contrôleurs dont nous disposions, tout en gardant à l'esprit l'adéquation nécessaire entre les caractéristiques des

différents capteurs avec l'application visée. Nous avons ainsi utilisé nombre d'interfaces, au nombre desquelles : clavier MIDI, gant de données, tablette graphique, bras haptique ou Méta-instrument.

4.2 Perspectives

4.2.1 Evaluation objective de la modification de durée

Comme nous l'avons déjà fait remarquer au cours du manuscrit, la complexité liée à la notion du rythme de la parole, et par extension la modification de la durée d'une phrase donnée, ne nous a pas permis pour l'instant de réaliser une évaluation objective de notre outil pour de telles modifications. Cependant, le paradigme de synchronie utilisé notamment dans les expériences de Fred Cummins (Cummins, 2003), paraît prometteur comme point de départ de cette évaluation.

Le principe de synchronie consiste, comme l'explique F. Cummins, à demander à deux locuteurs de prononcer un même texte, en même temps. Le protocole se déroule alors ainsi : on place dans une pièce deux personnes avec des casques audio, soit dos à dos, soit face à face. On donne aux sujets le même texte, qu'ils devront lire à haute voix, devant un micro. Dans les casques, le système diffuse sur une oreille sa propre voix, et dans l'autre la voix de l'autre locuteur. La tâche consiste alors pour les deux locuteurs, à accorder leurs rythmes de manière à prononcer le texte de la façon la plus synchrone possible. Cummins note, que si la tâche peut paraître difficile dans un premier temps, l'expérience montre qu'une courte période d'entraînement est nécessaire pour que les deux locuteurs réussissent à réaliser la tâche correctement

Dans un second temps, une fois que les deux voix sont enregistrées, une phase d'analyse est nécessaire pour mesurer les déviations temporelles entre les deux locuteurs, par extraction des p-centers des deux locuteurs. Les études de Cummins ont pu montré que les valeurs standards d'asynchronie (i.e. le décalage temporel entre le même indice temporel chez les deux locuteurs) était environ de $40\ \mathrm{ms}$.

Une expérience possible permettant de statuer des performances atteintes par notre système consisterait alors à enregistrer un texte lu par deux locuteurs différents, de façon indépendante. Ensuite, il suffirait de présenter au sujet, les deux enregistrements dans chacune des deux oreilles, dont l'un des deux serait modifiable temporellement grâce au système que nous avons développé. Les analyses pourraient ainsi nous révéler dans quelle mesure le sujet est capable d'accomplir cette synchronie, en comparaison avec les résultats obtenus avec deux sujets humains réalisant une synchronie.

4.2.2 Evaluation des attitudes japonaises

Outre le fait que notre synthétiseur soit contrôlé par des interfaces gestuelles, l'une des particularités intéressantes est qu'il intègre également la synthèse d'apériodicités structurelles et additives. Ainsi, certaines qualités vocales rarement traitées par la plupart des synthétiseurs (âpreté, souffle, rugosité ...) peuvent ainsi être produites.

L'une des particularités de la prosodie du japonais est que l'expression de certaines attitudes font appel à ces mêmes qualités de voix, et qu'elles permettent justement de distinguer deux expressions différentes. Les études ont montré que les auditeurs de langue maternelle japonaises étaient d'ailleurs capables de différencier ces attitudes de manière fiable sur de simples éclats de voix (c'est-à-dire des voyelles expressives) (Rilliard et al., 2009). Il nous paraît ainsi pertinent de pouvoir évaluer la possibilité de synthétiser tout ou partie de ces attitudes grâce à notre synthétiseur afin de pouvoir mener une étude perceptive de distinction d'attitudes du japonais.

4.2.3 Intégration facilitée des interfaces

L'une des problématiques inhérentes à l'utilisation d'interfaces gestuelles est celle de la connectivité. Lors de l'utilisation et le développement des différents instruments basés sur le synthétiseur CALM, l'un des problèmes récurrents est celui de la connexion des différents capteurs avec les logiciels de synthèse.

Au-delà des problèmes de résolution des capteurs, de calibration, de normalisation (ce dernier étant facilement résolu par l'utilisation d'une échelle entre 0 et 1), de définition d'un espace réduit de dimension contrôlables pertinent, la difficulté majeure liée au contrôle gestuel repose sur le fait que pour chaque nouvel appareil, le nombre de capteurs disponibles est différent, et peut-être soit sous-déterminé soit sur-déterminé (le nombre de contrôleurs étant rarement égal au nombre de dimensions contrôlées). Il convient donc de réfléchir à la manière de pouvoir réaliser des connexions *intelligentes* quelque soit l'interface utilisée.

La situation idéale consisterait à disposer d'un synthétiseur qui puisse prendre des décisions pertinentes quant aux dimensions vocales qui ne seraient pas connectées à un quelconque capteur. Il serait ainsi souhaitable, que du point de vue de la synthèse, le contrôle effectué soit transparent, c'est-à-dire qu'il puisse s'adapter au nombre de contrôleurs effectivement présent. Evidemment, il paraît utopique de pouvoir contrôler le synthétiseur avec un contrôleur unique faisant varier, par exemple la fréquence fondamentale. Toutefois, il paraît envisageable de hiérarchiser les dimensions vocales : la dimension mélodique aurait ainsi une importance plus élevée que le triangle vocalique.

Il paraît également possible de catégoriser les différents contrôleurs, non plus en termes ergonomique, de résolution ou de précision, mais plus exactement grâce aux caractéristiques physiques des différents capteurs qui la composent. Ainsi, on pourrait trouver des capteurs dont la

réponse est linéaire, logarithmique, exponentielle ... Le système pourrait alors, dans une première approche, décider à la place de l'utilisateur quelle dimension vocale la plus pertinente doit être contrôlée par ce capteur. L'apport théorique, et également pratique, de la logique floue et de la fusion de données pourrait, à ce titre se révéler d'une aide importante. Ce cadre de travail a en effet montré son efficacité dans d'autres domaines d'ingénierie pour permettre de prendre des décisions optimales en temps réel, selon le contexte environnemental à un instant donné. Il serait ainsi intéressant de pouvoir adopter ce type d'approche dans le cadre du contrôle de la synthèse musicale.

A ce titre, les travaux effectués par R. Bresin depuis de plus de 15 années maintenant sont remarquables (Bresin et al., 1995). Il fait d'ailleurs à ce propos deux remarques importantes : la première est que la logique floue permet de décrire grâce à l'utilisation de labels des modèles plus efficaces au regard des concepts musicaux, et d'autre part, selon lui, la logique floue conduit à la création de contrôleurs capables ainsi d'expliciter leur propre comportement. Certes, l'application visée, qui est celle de la génération automatique d'interprétation musicale par un ordinateur (Friberg et al., 2006), est différente de la notre, mais le but affiché est de pouvoir un jour ne plus distinguer une performance humaine d'une performance logicielle (Hiraga et al., 2004).

Je pense en effet que l'utilisation de la logique floue, combinée avec la fusion de données, permettrait à la fois de décrire les gestes de l'utilisateur selon une description plus expressive (gestes amples, rapides, lents ...) mais aussi de décider suivant la nature du geste, de la dimension vocale la plus propice à être modifiée. En outre, cela permettrait également d'envisager une utilisation multi-utilisateur du synthétiseur. Les études récentes de N. Rasamimanana (Rasamimanana et al., 2009) sur la co-articulation gestuelle révèlent des profils gestuels temporels proches de courbes habituellement rencontrés pour des variables floues. La prise en compte de ces approches pourraient sans doute mener à une nouvelle manière de contrôler les processus de synthèse, de façon plus expressive.

4.2.4 Applications possibles

Les différentes systèmes développés au cours de cette thèse permettent d'envisager plusieurs applications possibles. Concernant la modification prosodique en temps réel, la première application possible est sans doute celle de la génération de stimuli pour des expériences en recherche en prosodie. Nous avons pu en effet noter un intérêt assez important parmi cette communauté et la volonté de pouvoir utiliser notre outil pour réaliser des modifications prosodiques de phrases expérimentales. Notre système, comporte l'avantage de pouvoir modifier facilement l'intonation et la durée d'une phrase donnée, et cela de manière interactive et sans connaissance particulière en traitement du signal de parole. Les outils de modification prosodique traditionnels font souvent appel à des processus du traitement du signal peu aisés à manipuler et ajuster. Notre système permet en revanche, grâce à un appareillage restreint de pouvoir réaliser rapidement des enregistrements, et

cela de manière intuitive et interactive.

Pour des applications de synthèse à partir du texte, à l'instar des stimuli que nous avons réalisé, il est parfaitement imaginable de modifier la prosodie d'une phrase de synthèse brute, soit pour corriger des jonctions prosodiques entre unités consécutives, soit pour enrichir une base de données avec de nouveaux profils prosodiques.

Sur le terrain des applications commerciales, il est possible d'imaginer une aide à la génération de phrases synthétiques expressives. Notre outil de modification prosodique peut constituer en effet un module complémentaire adéquat pour la transmission de messages textuels expressifs par téléphone ou par internet. L'apparition récente de téléphones mobiles disposant d'une interface tactile paraît prometteur quant à une utilisation de notre outil sur ce type de système.

4.2.5 Objectifs à plus long terme

Comme nous l'avons déjà évoqué au cours de ce manuscrit, l'évaluation qui a pu être effectuée du système *Callliphonie* a surtout servi à statuer sur le fait que l'humain était capable de reproduire avec le geste manuel, avec des performances comparables à l'organe vocal, une mélodie et un rythme donné. Ce que sous-entend ce résultat est premièrement que les mouvements dynamiques des paramètres prosodiques que sont la hauteur et le rythme de la voix possèdent des évolutions temporelles comparables à celles du geste manuel, et par là même, ces dimension prosodiques peuvent être contrôlées de manière pertinente grâce à des interfaces homme-machine, ou des interfaces de captation gestuelle plus généralement. En d'autres termes, le mouvement gestuel est un bon candidat pour la stylisation d'une courbe intonative.

Deuxièmement, par extension, il devient parfaitement imaginable de transposer ce paradigme de modification prosodique à d'autres méthodes de traitement du signal de parole, pour peu que les algorithmes sous-jacents puissent être manipulés et modifiés en temps réel, synchrones à la hauteur. Si la première contrainte (i.e. temps réel) se comprend aisément dans le cadre d'une manipulation interactive de paramètres de synthèse, la seconde remarque sur la nécessité d'utiliser des algorithmes synchrones à la hauteur paraît moins évidente, et nous allons donc expliciter plus avant ce point de vue.

Dans le souci d'obtenir un cadre de modification prosodique global, comprenant également les dimensions de qualité vocale, il paraît souhaitable de pouvoir idéalement remplacer l'onde de débit glottique de parole naturelle par une version synthétique dont on puisse modifier les paramètres. Ainsi, on serait en mesure de modifier toutes les dimensions prosodiques selon un seul et même processus. Or, le remplacement de l'onde de débit glottique ne peut se faire que période par période, d'où la nécessité, ou tout du moins l'avantage, d'utiliser des méthodes de traitement du signal de parole synchrones à la hauteur.

Chapitre 5

Références bibliographiques

Références bibliographiques

Alku, P. and Vilkman, E. (1996). A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers. *Folia Phoniatrica*, 48 :240–254.

Arfib, D., Couturier, J.-M., Kessous, L., and Verfaille, V. (2002). Strategies of mapping between gesture parameters and synthesis model parameters using perceptual spaces. *Organised Sound, Cambridge University Press*, 7 :127–144.

Auto-Tune (2009). http://www.antarestech.com/. *Antares Audio Technologies*.

Bailly, L., Henrich, N., Webb, M., Müller, F., Anna-Katharina, L., and Hess, M. (2007). Exploration of vocal-folds and ventricular-bands interaction in singing using high-speed cinematography and electroglottography. In *19th International Congress on Acoustics*, Madrid, Spain.

Berry, D. A., Herzel, H., Titze, I. R., and Krischer, K. (1994). Interpretation of biomechanical simulations of normal and chaotic vocal fold oscillations with empirical eigenfunctions. *The Journal of the Acoustical Society of America*, 95(6) :3595–3604.

Birnbaum, D., Fiebrink, R., Malloch, J., and Wanderley, M. M. (2005). Towards a dimension space for musical artifacts. In *Proc. of the 2005 International Conference on New Interfaces for Musical Expression - NIME05*, pages 192–195, Vancouver, Canada.

Bloothooft, G., van Wijck, M., and Pabon, P. (2001). Relations between vocal registers in voice breaks. In *Proccedings of Eurospeech*.

Boersma, P. and Weenink, D. (Last checked, December 2008). http://www.praat.org. *doing phonetics by Computer*.

Bozkurt, B. (2004). *Zeros of the z-transform (zzt) representation and chirp group delay processing for the analysis of source and filter characteristics of speech signals*. PhD thesis, Faculté Polytechnique de Mons.

Bresin, R., De Poli, G., and Ghetta, R. (1995). A fuzzy formulation of kth performance rule system. In *Proceedings of the 2nd International Conference on Acoustic and Musical Research - CIARM 95*, pages 433–438, Ferrara, Italy.

Camacho, A. (2008). Detection of pitched/unpitched sound using pitch strength clustering. In *Proceedings of the Ninth International Conference on Music Information Retrieval*, pages 533–537, Philadelphia.

Campbell, N. and Mokhtari, P. (2003). Voice quality : the 4th prosodic dimension. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS'03)*, pages 2417–2420, Barcelona, Spain.

Castellengo, M., Richard, G., and d'Alessandro, C. (1989). Study of vocal pitch vibrato perception using synthesis. In *13th ICA, International Conference on Acoustics*, pages 121–124, Belgrade.

Catford, J. C. (1977). *Fundamental problems in phonetics*. Edinburgh University Press.

Coleman, R. F. (1963). Decay characteristics of vocal fry. *Folia Phoniatrica*, 15 :256–263.

Coleman, R. F. and Wendahl, R. W. (1967). Vocal roughness and stimulus duration. *Speech Monographs*, 34 :85–92.

Cook, P. (1991). *Identification of Control Parameters in an Articulatory Vocal Tract Model, With Applications to the Synthesis of Singing*. PhD thesis, Stanford University.

Cook, P. (1992). Spasm : a real-time vocal tract physical model editor/controller and singer : the companion softwaresynthesis system. *Computer Music Journal*, 17(1) :30–44.

Cook, P. (2005). Real-time performance controllers for synthesized singing. In *Proc. NIME Conference*, pages 236–237, Vancouver, Canada.

Cook, P. R. and Leider, C. (2000). Squeeze vox : A new controller for vocal synthesis models. In *International Computer Music Conference*, Berlin.

Cummins, F. (2003). Practice and performance in speech produced synchronously. *Journal of Phonetics*, 31(2) :139–148.

D'Alessandro, N., d'Alessandro, C., Le Beux, S., and Doval, B. (2006). Real-time calm synthesizer : new approaches in hands-controlled voice synthesis. In *Proc. of New Interfaces for Musical Expression 2006*, pages 266–271, Paris, France.

d'Alessandro, C. (2006). Voice source parameters and prosodic analysis. In Sudhoff, S., Leternovà, D., Meyer, R., Pappert, S., Augurzky, P., Mleinek, I., Richter, N., Schliesser, J., and de Gruyter, W., editors, *Method in Empirical Prosody Research*, pages 63–87. Berlin, New York.

d'Alessandro, C., D'Alessandro, N., Le Beux, S., and Doval, B. (2006). Comparing time domain and spectral domain voice source models for gesture controlled vocal instruments. In *Proceedings of the 5th International Conference on Voice Physiology and Biomechanics*, pages 49–52, Tokyo.

d'Alessandro, C., D'Alessandro, N., Le Beux, S., Simko, J., Cetin, F., and Pirker, H. (2005a). The speech conductor : gestural control of speech synthesis. In *eNTERFACE 2005*, Mons, Belgium. The SIMILAR NoE Summer Workshop on Multimodal Interfaces.

d'Alessandro, C., D'Alessandro, N., Le Beux, S., Simko, J., Cetin, F., and Pirker, H. (2005b). The speech conductor : Gestural control of speech synthesis. In *Proceedings of eNTERFACE'05 Summer Workshop on Multimodal Interfaces*.

d'Alessandro, C., Rilliard, A., and Le Beux, S. (2007). Computerized chironomy : evaluation of hand-controlled intonation reiteration. In *Proceedings of Interspeech 2007*, pages 1270–1273, Antwerpen, Belgium. ISCA.

D'Alessandro, N., Woodruff, P., Fabre, Y., Dutoit, T., Le Beux, S., Doval, B., and d'Alessandro, C. (2007). Realtime and accurate musical control of expression in singing synthesis. *Journal on Multimodal User Interfaces, Springer Berlin/Heidelberg*, 1(1) :31–39.

De Bodt, M. S., Wuyts, F. L., Van de Heyning, P. H., and Croux, C. (1997). Test-retest study of the grbas scale : Influence of experience and professional background on perceptual rating of voice quality. *Journal of Voice*, 11(1) :74–80.

de Cheveigne, A. and Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4) :1917–1930.

de Laubier, S. and Goudard, V. (2006). Méta-instrument 3 : a look over 17 years of practice. In *Proc. of the NIME Conference*, IRCAM, Paris, France.

Disklavier (2009). http://www.yamaha.com/yamahavgn/CDA/Catalog/Catalog_GSXOXX/0,,CTID%25253D201500%252526CNTYP%25253DPRODUCT,00.html. *Yamaha Corporation of America*.

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

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

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

Dudley, H., Riesz, R. R., and Watkins, S. S. A. (1939). A synthetic speaker. *Journal of the Franklin Institute*, 227(6) :739–764.

Dudley, H. and Tarnoczy, T. H. (1950). The speaking machine of wolfgang von kempelen. *The Journal of the Acoustical Society of America*, 22(2) :151–166.

Fant, G. (1960). *Acoustic theory of speech production*. Mouton, The Hague, Netherlands, 2nd edition. 1970 edition.

Fant, G. (1995). The lf-model revisited. transformations and frequency domain analysis.

Fant, G., Liljencrants, J., and Lin, Q. (1985). A four-parameter model of glottal flow.

Fels, S. (1991). Glove-talk : An adaptive interface that uses neural networks. In *Proceedings of Int. Conf. of IEEE Engineering in Medicine and Biology Society*, Orlando, Florida.

Fels, S. and Hinton, G. (1998). Glove-talk ii : A neural network interface which maps gestures to parallel formant speech synthesizer controls. *IEEE Transactions on Neural Networks*, 9(1) :205–212.

Ferrand, C. T. (1995). Effects of practice with and without knowledge of results on jitter and shimmer levels in normally speaking women. *Journal of Voice*, 9(4) :419–423.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 47(6) :381–391.

Flanagan, J. (1965). *Speech Analysis Synthesis and Perception*. Springer-Verlag, New York.

Friberg, A., Bresin, R., and Sundberg, J. (2006). Overview of the kth rule system for musical performance. *Advances in Cognitive Psychology*, 2(2-3) :145–161.

Fuller, B. F. and Horii, Y. (1986). Differences in fundamental, frequency, jitter, and shimmer among four types of infant vocalizations. *Journal of Communication Disorders*, 19 :441–447.

Gerratt, B. R. and Kreiman, J. (2001). Measuring vocal quality with speech synthesis. *The Journal of the Acoustical Society of America*, 110(5) :2560–2566.

Gibet, S., Kamp, J.-F., and Poirier, F. (2004). Gesture analysis : Invariant laws in movement. In *Gesture-based Communication in Human-Computer Interaction*, volume 2915, pages 1–9. LNCS/LNAI, Genova, Italy.

Goebl, W., Dixon, S., De Poli, G., Friberg, A., Bresin, R., and Widmer, G. (2008). *Sound to Sense - Sense to Sound : A state of the art in Sound and Music Computing*, chapter Sense in expressive music performance : Data acquisition, computational studies, and models, pages 195–242. Logos Verlag, Berlin.

Grau, S., d'Alessandro, C., and Richard, G. (1993). A speech formant synthesizer based on harmonic + random formant-waveforms representations. In ESCA, editor, *Proceedings of EUROSPEECH'93, 3rd European Conference on Speech Communication and Technology*, volume 3, pages 1697–1700, Berlin, Germany.

Hamon, C., Moulines, E., and Charpentier, F. (1989). A diphone synthesis system based on time-domain prosodic modification of speech. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 238–241, Glasgow.

Hanson, H. M. and Chuang, E. S. (1999). Glottal characteristics of male speakers : Acoustic correlates and comparison with female data. *Journal of the Acoustical Society of America*, 106(2) :1064–1077.

Henrich, N. (2001). *Etude de la source glottique en voix parlée et chantée : modélisation et estimation, mesures acoustiques et électroglottographiques, perception*. PhD thesis, Université Paris VI.

Henrich, N. (2006). Mirroring the voice from garcia to the present day : Some insights into singing voice registers. *Logopedics Phoniatrics Vocology*, 31 :3–14.

Henrich, N., d'Alessandro, C., Castellengo, M., and Doval, B. (2005). Glottal open quotient in singing : Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency. *Journal of the Acoustical Society of America*, 117 :1417–1430.

Henrich, N., d'Alessandro, C., Doval, B., and Castellengo, M. (2004). On the use of the derivative of electroglottographic signals for characterization of non-pathological phonation. *Journal of the Acoustical Society of America*, 115 :1321–1332.

Hermes, D. J. (1991). Synthesis of breathy vowels : Some research methods. *Speech Communication*, 10 :497–502.

Hermes, D. J. (1998). Measuring the perceptual similarity of pitch contours. *J Speech Lang Hear Res*, 41(1) :73–82.

Hess, D. A. (1959). Pitch,intensity,and cleft palate voice quality. *Journal of Speech and Hearing Research*, 2(113-125).

Hillman, R. E., Oesterle, E., and Feth, L. L. (1983). Characteristics of the glottal turbulent noise source. *Journal of Acoustical Society of America*, 74(3) :691–694.

Hiraga, R., Bresin, R., Hirata, K., and Katayose, H. (2004). Rencon 2004 : Turing test for musical expression. In *Proceedings of the 4th international conference on New Interfaces for Musical Expression (NIME'04)*, pages 120–123, Hamamatsu, Shizuoka, Japan.

Hollien, H. (1974). On vocal registers. *Journal of Phonetics*, 2 :125–143.

Hollien, H., Darnste, H., and Murry, T. (1969). Vocal fold length during vocal fry phonation. *Folia Phoniatrica*, 21 :257–265.

Hollien, H. and Michel, J. F. (1968). Vocal fry as a phonational register. *Journal of Speech and Hearing Research*, 11(1) :600–604.

Hollien, H., Moore, P., Wendahl, R. W., and Michel, J. F. (1966). On the natude of vocal fry. *Journal of Speech and Hearing Research*, 9(1) :245–247.

Hollien, H. and Wendahl, R. W. (1968). Perceptual study of vocal fry. *Journal of the Acoustical Society of America*, 43 :506–509.

Hunt, A. and Kirk, R. (2000). *Trends in Gestural Control of Music*, chapter Mapping Strategies for Musical Performance, pages 231–258. Ircam, Centre Pompidou.

Hunt, A., Kirk, R., and Wanderley, M. (2000). Towards a model for instrumental mapping in expert musical interaction. In *Proc. International Computer Music Conference*.

Ishizaka, K. and Flanagan, J. L. (1972). Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell System 1echnique Journal*, 51 :1233–1268.

Jehan, T. (2008). http://web.media.mit.edu/~tristan/. *Max/MSP Externals*.

Kelso, J. A. S., Vatikiotis-Bateson, E., Saltzman, E. L., and Kay, B. (1985). A qualitative dynamic analysis of reiterant speech production : Phase portraits, kinematics, and dynamic modeling. *Journal of the Acoustical Society of America*, 77(1) :266–280.

Kessous, L. (2004). Gestural control of singing voice, a musical instrument. In *Proceedings of the 2004 Conference on Sound and Music Computing*, IRCAM, Paris, France.

Klatt, D. (1982). The klattalk text-to-speech system. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 1589–1592.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America*, 67(3) :971–995.

Klatt, D. H. and Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*, 87(2) :820–857.

Kounoudes, A., Naylor, P. A., and Brookes, M. (2002). The dypsa algorithm for estimation of glottal closure instants in voiced speech. In *Proceedings of IEEE International Conference on Acoustics Speech Signal Processing*, volume I, pages 349–352.

Kreiman, J., Gerratt, B., Kempster, G., Erman, A., and Berke, G. (1993). Perceptual evaluation of voice quality : Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research*, 36 :21 – 40.

Kreiman, J. and Gerratt, B. R. (1998). Validity of rating scale measures of voice quality. *The Journal of the Acoustical Society of America*, 104(3) :1598–1608.

Lacquaniti, F., Terzuolo, C. A., and Viviani, P. (1983). The law relating the kinematic and figural aspects of drawing movements. *Acta Psychologica*, 54 :115–130.

Ladefoged, P. (1971). *Preliminaries to linguistic phonetics*. University of Chicago Press.

Larkey, L. S. (1983). Reiterant speech : An acoustic and perceptual validation. *The Journal of the Acoustical Society of America*, 73(4) :1337–1345.

Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge Universtiy Press.

Le Beux, S., Rilliard, A., and d'Alessandro, C. (2007). Calliphony : A real-time intonation controller for expressive speech synthesis. In *6th ISCA Workshop on Speech Synthesis*, Bonn, Germany. ISCA.

Lukaszewickz, K. and Kajalainen, M. (1987). Microphonemic method of speech synthesis. In *Proceedings of IEEE International Conference on Acoustics Speech Signal Processing*, pages 1426–1429, Dallas. IEEE.

MacKenzie, I. S. and Buxton, W. (1992). Extendign fitts' law to two dimensional tasks. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 219–226, New York. ACM.

Maoz, U., Portugaly, E., Flash, T., and Weiss, Y. (2006). Noise and the two-thirds power law. *Advances in Neural Information Processing Systems*, 17.

Marshall, M. T. and Wanderley, M. M. (2006). Evaluation of sensors as input devices for computer music interfaces. In Kronland-Martinet, R., Voinier, T., and Ystad, S., editors, *CMMR 2005 - Proc. of Computer Music Modeling and Retrieval 2005 Conference*, pages 130–139. Berlin Heidelberg : Springer-Verlag.

Max/MSP (2008). www.cycling74.com. *Cycling'74*.

McLean, A. and Wiggins, G. (2008). Vocable synthesis. In *Proceedings of the ICMC'08*, SARC, Belfast. International Computer Music Association.

Mocquereau, A. (1927). *Le nombre musical grégorien*. Desclée, Paris.

Monsen, R. B. and Engebretson, A. M. (1977). Study of variations in the male and female glottal wave. *Journal of Acoustical Society of America*, 62 :981–993.

Moore, P. and von Leden, H. (1958). Dynamic variations of the vibratory pattern in the normal larynx. *Folia Phoniatrica*, 10 :205–238.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6) :453–467.

Moulines, E. and Laroche, J. (1995). Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication*, 16(2) :175–205.

Nelson, W. L. (1983). Physical principles for economies of skilled movements. *Biological Cybernetics*, 46 :135–147.

Pausewang Gelfer, M. and Fendel, D. M. (1995). Comparisons of jitter, shimmer, and signal-to-noise ratio from directly digitized versus taped voice samples. *Journal of Voice*, 9(4) :378–382.

Perrier, P. and Fuchs, S. (2008). Speed-curvature relations in speech production challenge the one-third power law. *Journal of Neurophysiology*, 100(9) :1171–1183.

Pfitzinger, H. (2006). Five dimensions of prosody : Intensity, intonation, timing, voice quality, and degree of reduction. In Hoffmann, R. ; Mixdorff, H., editor, *Speech Prosody Abstract Book*, Studientexte zur Sprachkommunikation, Band 40, pages 6–9, Dresden. TUDpress.

Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation*. PhD thesis, MIT.

Prudon, R. and d'Alessandro, C. (2001). A selection/concatenation text-to-speech synthesis system : databases development, system design, comparative evaluation. In *4th ISCA/IEEE International Workshop on Speech Synthesis*. ISCA.

Puckette, M. (1996). Pure data. In *Proceedings, International Computer Music Conference*, pages 269–272, San Francisco. International Computer Music Association.

Rasamimanana, N., Kaiser, F., and Bevilacqua, F. (2009). Perspectives on gesture-sound relationships informed from acoustic instrument studies. *Organised Sound*, 14(2) :208–216.

Richard, G. and d'Alessandro, C. (1996). Analysis/synthesis and modification of the speech aperiodic component. *Speech Communication*, 19 :221–244.

Rilliard, A., Shochi, T., Martin, J.-C., Erickson, D., and Aubergé, V. (2009). Multimodal indices to japanese and french prosodically expressed social affects. *Language and Speech*, 52(2/3) :223–243.

Rodet, X. (1984). Time-domain formant wave function synthesis. *Computer Music Journal*, 8(3) :9–14.

Rosenberg, A. E. (1971). Effect of glottal pulse shape on the quality of natural vowels. *The Journal of the Acoustical Society of America*, 49(2B) :583–590.

Rothenberg, M. (1986). Vocal closure patterns and source-tract acoustic interaction. *The Journal of the Acoustical Society of America*, 79(S1) :S83.

Roubeau, B., Henrich, N., and Castellengo, M. (2009). Laryngeal vibratory mechanisms : The notion of vocal register revisited. *Journal of Voice*, 23(4) :425–438.

Rovan, J. B., Wanderley, M., Dubnov, S., and Depalle, P. (1997). Instrumental gestural mapping strategies as expressivity determinants in computer music performance. In *Proceedings of the Kansei - The Technology of Emotion Workshop*, Genova, Italy.

Sanguineti, V., Laboissière, R., and Ostry, D. J. (1998). A dynamic biomechanical model for neural control of speech production. *Journal of the Acoustical Society of America*, 103(3) :1615–1627.

Schoentgen, J. and de Guchteneere, R. (1995). Time series analysis of jitter. *Journal of Phonetics*, 23 :189–201.

Scott, S. K. (1998). The point pf p-centres. *Psychological Research*, 61(1) :4–11.

Shiller, D. M., Laboissière, R., and Ostry, D. J. (2002). Relationship between jaw stiffness and kinematic variability in speech. *Journal of Neurophysiology*, 88 :2329–2340.

Sorensen, D. N. and Horii, Y. (1983). Frequency and amplitude perturbation in the voices of female speakers. *Journal of Communication Disorders*, 16 :57–61.

Stevens, K. N. (1960). Toward a model for speech recognition. *Journal of the Acoustical Society of America*, 31(1) :47–55.

Stevens, K. N. and Hanson, H. M. (1995). *Vocal fold physiology : voice quality control*, chapter Classification of glottal vibration from acoustic measurements, pages 147–170. Singular Publishing Group, San Diego, California.

Sundberg, J. (2001). Level and center frequency of the singer's formant. *Journal of Voice*, 15 :176–186.

Tarling, J. (2004). *The Weapons of Rethoric*. Corda Music Pub., London.

Tasko, S. M. and Westbury, J. R. (2004). Speed-curvature relations for speech-related articulatory movement. *Journal of Phonetics*, 32 :65–80.

Tatham, M. and Morton, K. (2004). *Expression in Speech : Analysis and Synthesis*. Oxford University Press, Oxford, second edition 2006 edition.

Titze, I. R. and Strong, W. J. (1975). Normal modes in vocal cord tissues. *The Journal of the Acoustical Society of America*, 57(3) :736–744.

Traunmüller, H. and Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *Journal of the Acoustical Society of America*, 107(6) :3438–3451.

Veldhuis, R. N. J. (1996). An alternative for the lf model. In van Amelsfort, U., Janse, M. D., Hermes, D. J., de Ridder, H., and van Overveld, W. M. C. J., editors, *IPO Annual Progress Report 31*, pages 100–108. The Institute for Perception Research, Eindhoven.

Viviani, P. and Terzuolo, C. (1983). *Language production*, volume II, Development, writing and other language processes, chapter The organization of movement in handwriting and typing, pages 103–146. Academic Press, New York.

Wanderley, M. M. and Depalle, P. (1999). *Interfaces Homme-Machine et Creation Musicale*, chapter Contrôle Gestuel de la Synthèse Sonore. Hermès Science Publishing, Paris.

Wanderley, M. M. and Depalle, P. (2004). Gestural control of sound synthesis. In Johannsen, G., editor, *Special Issue on Engineering and Music - Supervisory Control and Auditory Communication*, volume 92, pages 632–644.

Wanderley, M. M., Malloch, J., Birnbaum, D., Sinyor, E., and Boissinot, J. (2006). Sensorwiki.org : A collaborative resource on transducers for researchers and interface designers. In *Proceedings of the 2006 International Conference on New Interfaces for Musical Expression (NIME06)*, pages 180–183, Paris, France.

Chapitre 6

Articles

Sommaire

Modification prosodique **179**
Interspeech 2007 179
Speech Synthesis Workshop 2007 184
Synthèse de source glottique **191**
eNTERFACE 2005 191
NIME 2006 202
ICVPB 2006 209
eNTERFACE 2006 214
JMUI 2008 225
Autres travaux : Le projet ORA **235**
ICMC 2009 235
Smart Graphics 2009 240

Computerized chironomy: evaluation of hand-controlled Intonation reiteration

Christophe d'Alessandro, Albert Rilliard, Sylvain Le Beux[1]

[1] LIMSI-CNRS, BP 133, F-91403, Orsay, France

{cda, rilliard, slebeux}@limsi.fr

Abstract

Chironomy means in this paper intonation modeling in terms of hand movements. An experiment in hand-controlled intonation reiteration is described. A system for real-time intonation modification driven by a graphic tablet is presented. This system is used for reiterating a speech corpus (sentences of 1 to 9 syllables, natural and reiterant speech). The subjects also produced vocal imitation of the same corpus. Correlation and distances between natural and reiterated intonation contours are measured. These measures show that chironomic reiteration and vocal reiteration give comparable, and good, results. This paves the way to several applications in expressive intonation synthesis and to a new intonation modeling paradigm in terms of movements.

Index Terms: prosodic modeling, prosodic perception, gestures, prosodic synthesis

1. Introduction

Although various intonation models have been proposed for a variety of languages, the question of expressive intonation representation is still wide open. Phonological models of intonation are focusing on contrastive (often tonal) structures: they are not designed for description of expressive intonation variations. Phonetic description and stylization of intonation often describes melodic patterns in terms of "movements", "contours" or "target points". The approach defended in this paper is based on the hypothesis that intonation shares a lot of common features with other types of expressive human movements or gestures (like face and hand gestures). Then, addressing the question of intonation representation in terms of movements, like e.g. hand movements, could bring new insights in intonation research. The analogy between intonation and movements seems promising. A first application is direct hand-controlled expressive synthesis: this could be used for corpora enrichment in concatenative speech synthesis, or stimuli generation in expressive speech analysis. A second more fundamental application could be intonation modeling in terms of movement representation (movement speed, direction, height). A main advantage of such a modeling is that intonation and rhythm are dealt with in a unified framework.

Expressive intonation description in terms of hand gestures is known since antiquity under the term "chironomy" (cf. [9], part 1, p. 103). This term comes from the Greek "chiro" (hand) and "gnomos" (rule). The term appears first in the fields of rhetoric for describing co-verbal hand movement that reinforce expression of the discourse [11]. Another meaning appears in medieval music, where chironomy is meant for the hand gestures of the conductor that indicates the tones to the choir in Gregorian chant ([9], part 2, p. 683).

Music and speech are forms of human communication by the mean of expressive sound control. Music, contrary to speech, developed the usage of external "instruments" for sound production and sound control. Instrumental music is produced by hand-, breath-, or feet-controlled "interfaces". As new interfaces for musical expression recently received a lot of attention, resources like real-time sound programming languages, control devices, modification algorithms are available in the electronic music community (cf. [4], [7]). Along this line a system for computerized chironomy, i.e. real-time melodic control by hand-driven movements is presented and evaluated in this paper. Among the devices available for controlling hand movement, hand-writing (graphic tablet) has been preferred. This is because hand writing allows for the most accurate and intuitive intonation control. The main questions addressed in the present experiment are:

1. How well can handwriting movements reproduce intonation movements?
2. How do handwriting and vocal intonation stylization compare?
3. In both cases, how close are natural intonation contours and stylized contours?

These questions are addressed using an intonation reiteration paradigm. The task of the subjects was to reproduce intonation patterns by vocal mimicking and hand-control movements. Both speech and reiterant (i.e. "mamama") speech sampled of various sizes were proposed. Distance measures between the original and reiterated speech are used as performance assessment.

The paper is organized as follows. The experimental apparatus, test paradigm and analysis procedures are described in Section 2. Results in terms of performance for intonation imitation are given in Section 3. Section 4 discusses the results obtained, and gives some conclusions.

2. Experiments

2.1. Prosodic control system

2.1.1. Gestural pitch shifter

A new system was developed in order to control the pitch of speech utterances by means of handwriting gestures. The system, to some extent similar to the one described in [1], is based on the Max/MSP programming environment, and uses a real time version of the TDPSOLA algorithm. It deals with two inputs: (1) a recorded speech utterance with a flattened fundamental frequency (to e.g. 120Hz for a male speaker), and (2) the output of a gesture control device such as a graphic tablet. The value of one parameter of the graphic tablet (controlled by handwriting movements) is mapped to the pitch value of the spoken utterance, resulting in a direct control by the gesture of the output utterance pitch. Hence, this system allows one operator to precisely control the pitch of a previously recorded utterance, using only a pen on a graphic tablet.

2.1.2. *Prosodic imitation interface*

In order to test whether the control of prosody by handwriting movements can realistically reproduce natural prosody, a specific computer interface has been developed (cf. figure 1) under the Max/MSP platform. It is intended to allow subjects of the experiment to imitate the prosody of natural speech either vocally or by handwriting movements. Each subject listens to a natural sentence by clicking on a button with the mouse pointer, and therefore has to imitate the prosody he has just heard by two means: vocally by recording his own voice, and by using the gestural controller of prosody.

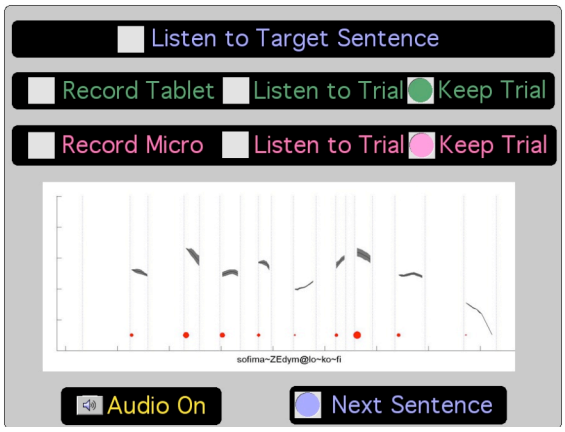


Figure 1: *interface of the experiment. Buttons allow to listen to the original sentence, record his speech or the graphic tablet, listen to a performance and save it if it is satisfactory. The image represents the prosody of the current sentence.*

The interface displays some control buttons (cf. figure 1): (1) to record the voice or the graphic tablet imitation, (2) to replay the recorded imitations and (3) to save the satisfactory ones. It also displays a graphic representation of the prosodic parameters of the original sound, as it will be described latter.

As the aim of the experiment is to investigate how close to the original the imitations can be, subjects are able to listen the original sound when they need to, and to perform imitation until they are satisfied. Several performances can be recorded for each original sound.

Finally, subjects go on to the next sound. As the test typically lasts several minutes per sentence, subjects are instructed to take rest from time to time.

2.2. Corpus

These experiments are based on a dedicated corpus constructed on 18 sentences, ranging from 1 to 9 syllables length (cf. table 1). Each sentence was recorded in its lexicalized version, and also in a delexicalized version, replacing each syllable by the same /ma/ syllable, in order to obtain reiterant speech [8]. When constructing the corpus, words were chosen with respect to two criterions (use of CV syllable structure and no plosive consonant at the beginning of the words), in order to obtain easily comparable prosodic patterns amongst the sentences and to avoid important micro-prosodic effect due to plosive bursts.

Two speakers (a female and male, native speakers of French) recorded the corpus. They have to produce each sentence in a random order, and according to three different consigns: (1) using a declarative intonation, (2) performing an emphasis on a specific word of the sentences (generally the verb) and (3) using an interrogative intonation. The speakers were instructed to read the sentence and then to produce it using the current intonation style. Once the sentence is recorded in its lexicalized version, they have to reproduce it by using the same prosody, but in its reiterated version. Speakers were able to make as many trials as needed in order to obtain a satisfactory pair of sentences.

108 sentences were thus recorded and directly digitalized on a computer (41kHz, 16bits) for each speaker, using an USBPre sound device connected to an omnidirectional AKG C414B microphone placed 40 cm to the speaker mouth, and performing a high-pass filtering of frequency under 40Hz plus a noise reduction of 6dB.

2.3. Subjects

Until now, 4 subjects have completed the experiment on a subset of 9 sentences ranging from 1 to 9 syllables, either lexicalized or reiterated, and with the three prosodic conditions (declarative, emphasized, interrogative), for the male speaker. All subjects are involved in this work and completely aware of its aims and are therefore familiar with prosody. Three out of the four subjects are trained musicians. One of the four subjects is the male speaker of the original corpus, who has therefore imitate his own voice vocally and by handwriting movements.

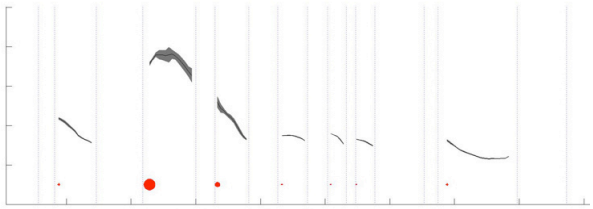


Figure 2: *prosodic parameters of a 7-syllable length sentence from our corpus.*

2.4. Prosodic contours measurements

All the sentences of the corpus were manually analyzed in order to extract their prosodic parameters: fundamental frequency (in semitones), syllabic duration, and intensity thanks to Matlab (the yin script [3]) and Praat [2] programs.

Table 1: *The 18 sentences of the corpus, from 1 to 9-syllable length.*

Nb syllable	Sentence	Phonetic	Sentence	Phonetic
1	Non.	[nɔ̃]	L'eau	[lo]
2	Salut	[saly]	J'y vais.	[ʒi vɛ]
3	Répétons.	[ʁepetɔ̃]	Nous chantons.	[nu ʃɑ̃tɔ̃]
4	Marie chantait.	[maʁi ʃɑ̃tɛ]	Vous rigolez.	[vu ʁigole]
5	Marie s'ennuyait.	[maʁi sɑ̃nɥijɛ]	Nous voulons manger.	[nu vulɔ̃ mɑ̃ʒe]
6	Marie chantait souvent.	[maʁi ʃɑ̃tɛ suvɑ̃]	Nicolas revenait.	[nikola ʁəvənɛ]
7	Nous voulons manger le soir.	[nu vulɔ̃ mɑ̃ʒe lə swaʁ]	Nicolas revenait souvent.	[nikola ʁəvənɛ suvɑ̃]
8	Sophie mangeait des fruits confits.	[sofi mɑ̃ʒɛ de fʁɥi kɔ̃fi]	Nicolas lisait le journal.	[nikola lizɛ lə ʒuʁnal]
9	Sophie mangeait du melon confit.	[sofi mɑ̃ʒɛ dy məlɔ̃ kɔ̃fi]	Nous regardons un joli tableau.	[nu ʁəgaʁdɔ̃ ɛ̃ ʒoli tablo]

For all the sentences, graphics were displayed to depict the prosody of original sound in order to facilitate the task of subjects of the experiment (cf. figure 2). These graphics represents the smoothed F0 of the vocalic segments (manually aligned), with the line thickness representing the voicing strength. The voicing strength was calculated from the intensity (in dB) of the signal at the point of F0 analysis. The locations of the Perceptual Centers [10] are represented by red circles, the diameter of which is related to the mean intensity of the vocalic segment. Vertical dotted lines represent the phonemes' boundaries.

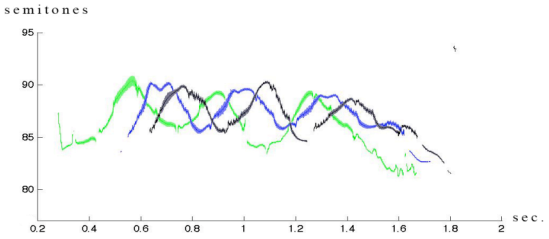


Figure 3: *raw F0 value (in tones) for an original sentence (gray) and the two vocal imitations of one subject. Stimuli are not time-aligned.*

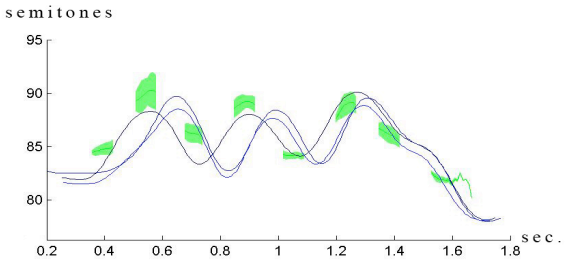


Figure 4: *stylized F0 of an original sentence (the same as in fig. 3 – gray curve, smoothed values for the vocalic segment expressed in tones), and the value of the pitch parameter controlled by the graphic tablet for all the imitations performed by one subject. Stimuli are time-aligned.*

2.5. Prosodic distances and correlation

In order to evaluate the performance of the imitation (either vocal or gestural), two physical measures of the distance between the fundamental frequency extracted from the imitation and the one from the original sentence were set of, on the basis of the physical dissimilarity measures introduces by Hermes [6]: the correlation between the two F0 curves, and the root-mean-square difference between theses two curves. As already noted by Hermes, the correlation is a measure of the similarity of the two sets of F0 parameters, whereas the RMS difference is a dissimilarity measure, but both give an idea of the similarity of the compared F0 curves. However, correlation test the similitude between the shapes of the two curves, without taking into account their mean distances: e.g. one can reproduce an F0 curve an octave lower than the original, if the shape is the same, the correlation will be very high. On the contrary, the RMS distance will give an idea of the area between the two curves, and is sensitive to differences between the two mean F0 levels.

Using a similar procedure as the one described in [6], the two prosodic distances were applied with a weighting factor in order to give more importance to the phonemes with a higher sound level. The weighting factor used is the intensity, as a measure of the local strength of voicing.

These two dissimilarity measures were automatically calculated for all the gestural imitations recorded by the four subjects for each of the 54 sentences. Then only the closest gestural imitation (according to first the weighted correlation and then the weighted RMS difference) was kept for the result analysis.

This part of the work can be completely automated, as there is no change in the duration of the output of the gestural controller of speech (only F0 is controlled). This is not the case for the oral imitations, which have to be manually labeled in order to calculate such distances. The distance computation supposes segments of the same length, a condition not met for vocal imitations. Therefore, only the distances between the original sentences and the gestural imitations have been calculated so far.

Graphics with the raw F0 value of both the original and the vocal imitations have been produced in order to visually compare the performances of gesture vs. vocal imitations. Graphic with the stylized F0 of the original sentences (smoothed F0 for the vocalic segments) superimposed with the course of the pen on the graphic tablet were also produced in order to compare the two imitations modalities (fig. 3 & 4).

3. Results

3.1. Prosodic distances and correlation

The physical distances between stimuli produced by handwriting movements are summarized in table 2. In analyzing the results of the experiment, the relative influence of each controlled parameter will be detailed.

Table 2: *mean distances for each subject and for all 54 sentences imitated by handwriting movements.*

Subject	R	RMS
CDA	0.866	3.108
BD	0.900	3.079
SLE	0.901	3.091
AR	0.898	4.728
Total	0.891	3.502

3.1.1. Effect of subjects

There is no important difference between the results obtained by all subjects: all correlations are comparable and around .9, showing that subjects are able to perceive and reproduce the shape of the intonation curve by means of handwriting movement. The only noticeable difference is the RMS distance obtained by subject AR (4.7) compared to the score of other subjects (around 3.1). This difference indicates an F0 curve closer to the original one for the three other subjects than for AR. This can be explained by the fact that AR is the only subject without a musical education, and therefore he is not trained to reproduce a given melody as the others are. However, as the correlations are quite the same, it does not imply difficulty to reproduce the pitch variation, but only the pitch height.

3.1.2. Effect of sentence length

The sentence length has a more noticeable effect on the distances. As shown in the figure 5, the dissimilarity measures increase as the sentences length grows: correlation continuously decrease when sentence length increase, and except for some accident for the 3 and 7-syllable length sentences, RMS difference grows according to sentence length. The two accidents could be explained by high RMS distances obtained by two subjects for this stimulus, and by the fact that this measure is particularly sensitive to small differences between curves. The effect of sentence length could be an artifact, because computation of correlation does

not take into account any weighting for length compensation. More analyses would be needed before concluding on a sentence length effect.

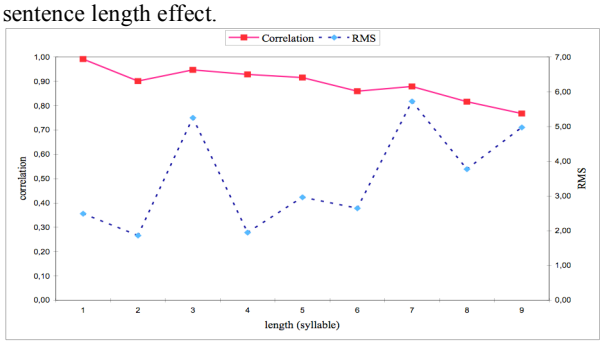


Figure 5: *evolution of the two distances measures with the sentence's length. X-axis: length of stimuli, left Y-axis: correlations (plain line), right Y-axis: RMS difference (dotted line).*

3.1.3. Effect of the prosodic style and type of stimuli

Declarative, emphasized or interrogative sentence modalities give similar results according to the correlation measure, but RMS distance is smaller for declarative curves (2.0) than for emphasized or interrogative ones (respectively 4.2 and 4.4). It can be linked to the preceding result: subjects are able to reproduce the shape of all intonation contours, but the precise perception the pitch level is harder when the curve present a glissando (e.g. during emphasis or interrogation) than a more flat curve, like for declarative intonation.

Finally, to imitate a reiterant sentence is nor easier nor harder than to imitate a lexicalized one: distances are the same for both kinds of stimuli.

4. Discussion and conclusions

4.1. Performance level and feasibility

Good performance levels are achieved in terms of correlation and distances between original and reiterated intonation contours. Of course, it must be pointed out that the best reiterated utterance has been selected for each sentence. However, the amount of training of each subject was not very heavy. The task seemed not particularly difficult, at least compared to other intonation recognition tasks, like e.g. musical dictation.

4.2. Gestures and vocal modalities

A remarkable and somewhat striking result is that the performance levels reached by hand written and vocal reiterated intonation are very comparable. This could suggest that intonation, both on the perceptual and motor production aspects, is processed at a relatively abstract cognitive level, as it seems somehow independent of the modality actually used. This fact was already inferred by orators in the ancient world, because description of the expressive effect of co-verbal gestures (i.e. multimodal expressive speech) has been remarked even in early roman rhetoric treatises [11]. Then one can hypothesize that intonation control can be achieved by other gestures than pitch gestures with comparable accuracy.

4.3. Intonation and gestures

It seems that micro-prosodic variations have been neglected for almost all sentences. Writing is generally slower than speaking, and then hand gestures are not able to follows fine grained intonation details like micro-prosody [5]. Moreover, results for delexicalized speech and normal speech are comparable, although micro-prosody is almost neutralized in delexicalized speech. Then the hand gestures correspond rather to prosodic intonation movements. The specific gestures used by different subjects for achieving the task at hand have not been analyzed in great detail for the moment. Some subject used rather circular movements, other rather linear movements. This point will be addressed in future work.

4.4. Conclusion and future work

This paper presents a first evaluation of computerized chironomy, i.e. hand-driven intonation control. The results show that vocal intonation reiteration and chironomic intonation reiteration give comparable intonation contours in terms of correlation and RMS distance. Applications and implications of these finding are manifold. Chironomic control can be applied to expressive speech synthesis. It can also be used for expressive speech analysis, as expressive contours can be produced and represented by the hand-made tracings. Future work will address the question of gesture control of rhythm and voice quality parameters. An auditory evaluation of the reiterated intonation contours is also planned. Finally, this work can also serve as a basis for intonation modeling in terms of movements. This could form a unified framework for expressive gesture representation, using common features like velocity, target position, rhythmic patterns etc.

5. References

[1] D'Alessandro, N., d'Alessandro, C., Le Beux, S. & Doval, B. (2006). "Real-time CALM synthesizer: new approaches in hands-controlled voice synthesis". Proc. of NIME2006, 266-271, Paris, France, June 4-8.

[2] Boersma, P. & Weenink, D. (2006): Praat: doing phonetics by computer (Version 4.5.05) [Computer program]. Retrieved 12/2006 from http://www.praat.org/

[3] de Cheveigné, A. & Kawahara, H. (2002). "YIN, a fundamental frequency estimator for speech and music". JASA, 111, 1917-1930.

[4] Cook, P. (2005). "Real-Time Performance Controllers for Synthesized Singing". Proc. NIME 2005, 236–237, Vancouver, Canada, May 2005.

[5] Fels, S. & Hinton, G. (1998). "Glove-Talk II: A Neural Network Interface which Maps Gestures to Parallel Formant Speech Synthesizer Controls". IEEE Transactions on Neural Networks, 9 (1), 205–212.

[6] Hermes, D.J. (1998). "Measuring the Perceptual Similarity of Pitch Contours". J. Speech, Language, and Hearing Research, 41, 73-82.

[7] Kessous, L. (2004). "Gestural Control of Singing Voice, a Musical Instrument". Proc. of Sound and Music Computing 2004, Paris, October 20-22.

[8] Larkey, L.S. (1983). "Reiterant speech: an acoustic and perceptual validation". JASA, 73(4), 1337-1345.

[9] Mocquereau, A. (1927). "Le nombre musical grégorien". Desclée, Paris.

[10] Scott, S.K. (1993). P-Centers in speech – an acoustic analysis. PhD thesis, University College London.

[11] Tarling, J. (2004). "The Weapons of Rethoric", Corda Music, Pub. London.

Calliphony: A real-time intonation controller for expressive speech synthesis

Sylvain Le Beux, Albert Rilliard, Christophe d'Alessandro

LIMSI-CNRS, BP 133, F-91403, Orsay, France

{slebeux, rilliard, cda}@limsi.fr

Abstract

Intonation synthesis using a hand-controlled interface is a new approach for effective synthesis of expressive prosody. A system for prosodic real time modification is described. The user is controlling prosody in real time by drawing contours on a graphic tablet while listening to the modified speech. This system, a pen controlled speech instrument, can be applied to text to speech synthesis along two lines. A first application is synthetic speech post-processing. The synthetic speech produced by a TTS system can be very effectively tuned by hands for expressive synthesis. A second application is data-base enrichment. Several prosodic styles can be applied to the sentences in the database without the need of recording new sentences. These two applications are sketched in the paper.

Index Terms: prosodic modeling, prosodic perception, gestures, prosodic synthesis

1. Introduction

As speech synthesizers attain acceptable intelligibility and naturalness, the problem of controlling prosodic nuances emerges. Expression is made of subtle variations (particularly prosodic variations) according to the context and to the situation. In daily life, vocal expressions of strong emotions like anger, fear or despair are rather the exception than the rule. Then a synthesis system should be able to deal with subtle continuous expressive variations rather than clear-cut emotions.

Expressive speech synthesis may be viewed from two sides: on the one hand is the question of expression specification (what is the suited expression in a particular situation?) and on the other hand is the question of expression realization (how is the specified expression actually implemented). The first problem (situation analysis and expression specification) is one of the most difficult problems for research in computational linguistics, because it involves deep understanding of the text and its context. In this paper, only the second problem is addressed. The goal is to modify speech synthesis in real time according to the gestures of a performer playing the role of a "speech conductor" [1]. The Speech Conductor adds expressivity to the speech flow using Text-to-Speech (TTS) synthesis, prosodic modification algorithms and gesture interpretation algorithms.

This work is based on the hypothesis that human expressivity can be described in terms of movements or gestures, performed through different media, e.g. prosodic, body or facial movements. This question is closely related to musical synthesis, a field where computer based interfaces are still subject of much interest and development [2]. It is not the case for speech synthesis, where only a few interfaces are available for controlling in real time expressivity of spoken utterances. Existing gesture-controlled interfaces for speech production are either dealing with singing synthesis (cf. [3], [4]) or with full speech synthesis [5], but with a sound quality level insufficient for expressivity generation.

In this paper a new system for real-time control of intonation is presented, together with application to text-to-speech synthesis. This system maps hand gestures to the prosodic parameters, and thus allows the user to control prosody in a cross-modal way. As a by-product, the cross-modal approach of prosody generation represents a new way to generate and describe prosody and may therefore shed a new light on the fields of prosody systems and prosody description.

The paper is organized as follows. The real-time intonation controller is described in Section 2. The performances of the controller for real-time intonation modification are evaluated in section 3. Applications to expressive text-to-Speech synthesis are sketched in section 4. Section 5 discusses the results obtained so far, proposed future work and gives some conclusions.

2. Real-time intonation controller

2.1. Principle

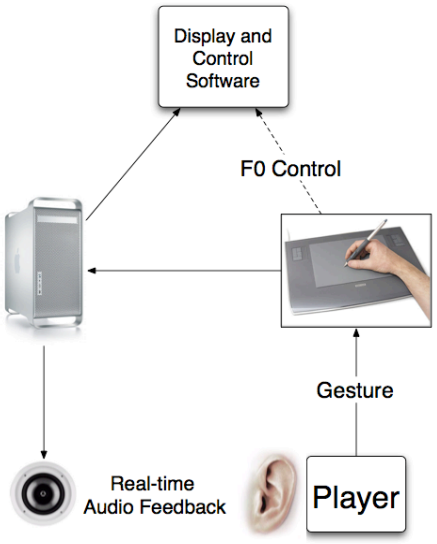


Figure 1: *Generic diagram of the system*

The real-time intonation controller operates in principle like a musical instrument. The loop between the player and the instrument is depicted in Figure 1. The player's hand movements are captured using an interface, and these movements are mapped on the input controls of the synthesizer. The sound is modified accordingly, played, and the player, who modifies his gestures as a function of the perceived and intended sounds, perceives this audio feedback.

2.2. Gesture interface: writing movements

Many devices, among which MIDI keyboard, Joystick and data glove; have been tested for capturing gestures with intonation control in mind.

Keyboard is not well fitted because it allows only discrete scales, although in speech a continuous control is mandatory. An additional pitch-bend wheel proved not very convenient from an ergonomic point of view.

As for the joystick and data glove, the precision in terms of position seemed insufficient: it proved too difficult to reach accurately a given target pitch. Such devices seem better suited for giving directions (as in a flight simulator) than precise values.

The graphic tablet has been chosen because it presents a number of advantages: its sampling frequency is high (200 Hz) and its resolution in terms of spatial position is sufficient for fine-grained parameter control (5080 dots per inches). Moreover, all the users are trained in writing since childhood, and are 'naturally" very much skilled in pen position control. Scripture, like speech, is made of a linguistic content and a paralinguistic, expressive content (in this case called "calligraphy"). There is a remarkable analogy between pitch contour and scripture. This analogy between drawing and intonation is very effective and intuitive from a performance point of view. Untrained subjects proved to be surprisingly skilled for playing with intonation using the pen on the graphic tablet, even at the first trial. For intonation control, only one axis of the tablet is necessary. The vertical dimension (Y-axis) is mapped on the F0 scale, expressed in semi-tones. The x-scale is not used: it means that very different gestures can be used for realizing a same intonation pattern: some players were drawing circle-like movements, when others preferred vertical lines or drawing similar to pitch contours. The second spatial dimension of the tablet will be used later for duration control in a second stage. Other degrees of freedom are still left in the tablet (pressure, switch) and will be use for controlling additional parameters, e.g. parameters related to voice quality.

Taking these observations into account, we decided to opt for a Wacom graphic Tablet, A4 size and we based our platform on a Power PPC Apple G5 Mac, 2.3 GHz bi-processor.

2.3. Real-time software

Real-time processing of information is a key point of the Calliphony system: as the user adapts his hand movement to perceived pitch at the output of the system, the delay has to remain inaudible. Calliphony is elaborated under the Max/MSP[1] software ([6], [7]), which is a graphical development environment intended to processes sound in real-time and which has already proven several years of reliable experience in real-time sound processing. Concerning the modification of speech pitch, we used a TD-PSOLA [8] Pitch-Shifter external provided by Tristan Jehan for Max/MSP environment [9].

As described on Figure 2, Calliphony takes as inputs the Y-axis position of the pen on the graphic tablet, and a recorded sound to be modified. It then maps the pitch value of the sound output to a value corresponding to the Y-axis value. This mapping is done on a logarithmic scale, such as the metric distance of each octave is the same. This corresponds analogously to the perception of the pitch by the human ear.

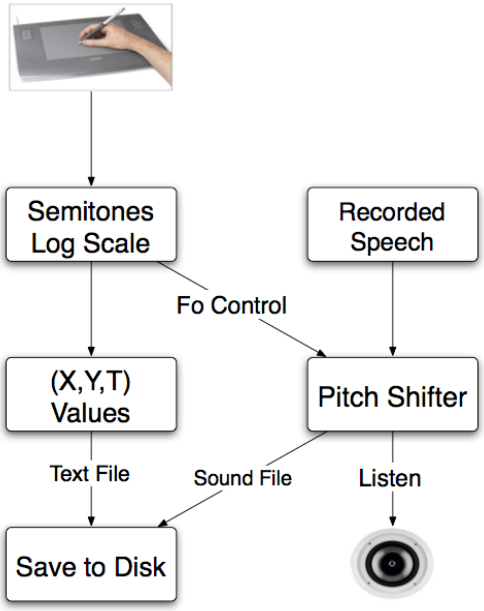


Figure 2: *"Calliphony" system description*

3. Evaluation of the controller

The use of handwriting movement to control pitch is not a priori straightforward. An evaluation procedure has therefore been developed, in order to assess the ability of a human to perform real-time control of speech prosody. The principle of this evaluation procedure is to measure the ability of the Calliphony player to imitate as closely as possible the prosody of an original sentence. The handwriting imitation performances are compared to the oral ability of the same user to imitate the same sentences. This work is described in more detail in a companion paper (cf. [10])

3.1. Prosodic imitation interface

A specific interface (cf. fig. 3) was developed to allow the subjects of the experiment to easily perform their imitation task. This interface encapsulate the Calliphony system, so that the user can listen to an original sentence, and then imitate the prosody both on a F0 flattened version of the sentence and vocally by recording his own voice.

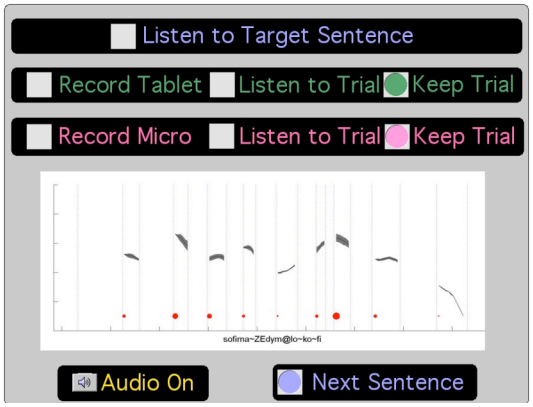


Figure 3: interface used for the handwriting imitation of prosody. Buttons allow to listen to the original sentence, record its own speech or the graphic tablet, listen to a recorded performance and save it. The current sentence's F0 is displayed.

[1]It is noticeable however that Max/MSP software is not multithreaded and consequently did not allows taking full advantage of the multi-processors architectures.

Table 1: *The 18 sentences of the corpus, from 1 to 9-syllable length.*

Nb syllable	Sentence	Phonetic	Sentence	Phonetic
1	Non.	[nɔ̃]	L'eau	[lo]
2	Salut	[saly]	J'y vais.	[ʒi vɛ]
3	Répétons.	[ʁepetɔ̃]	Nous chantons.	[nu ʃɑ̃tɔ̃]
4	Marie chantait.	[maʁi ʃɑ̃tɛ]	Vous rigolez.	[vu ʁigole]
5	Marie s'ennuyait.	[maʁi sɑ̃nɥijɛ]	Nous voulons manger.	[nu vulɔ̃ mɑ̃ʒe]
6	Marie chantait souvent.	[maʁi ʃɑ̃tɛ suvɑ̃]	Nicolas revenait.	[nikola ʁəvənɛ]
7	Nous voulons manger le soir.	[nu vulɔ̃ mɑ̃ʒe lə swaʁ]	Nicolas revenait souvent.	[nikola ʁəvənɛ suvɑ̃]
8	Sophie mangeait des fruits confits.	[sofi mɑ̃ʒɛ de fʁɥi kɔ̃fi]	Nicolas lisait le journal.	[nikola lizɛ lə ʒuʁnal]
9	Sophie mangeait du melon confit.	[sofi mɑ̃ʒɛ dy məlɔ̃ kɔ̃fi]	Nous regardons un joli tableau.	[nu ʁəgaʁdɔ̃ ɛ̃ ʒoli tablo]

As the aim of the evaluation is to investigate how close to the original the imitation can be, subjects are able to listen to the original sound when they need to, and to perform imitation until they are satisfied. Several performances can be recorded for each original sound.

3.2. Evaluation paradigm

3.2.1. Corpus

The evaluation procedure is based on a dedicated corpus constructed on 18 sentences, ranging from 1 to 9 syllables length (cf. table 1). Each sentence was recorded in its lexicalized version, and also in a reiterant delexicalized version, replacing each syllable by the same /ma/ syllable. Constraints on the corpus construction were: the use of CV syllable structure and absence of plosive consonant at the beginning of each word. Such constraints aimed at obtaining easily comparable prosodic patterns amongst the sentences and at avoiding important micro-prosodic effects due to plosive bursts.

Two native speakers of French recorded the corpus (a female and a male), according to three consigns: (1) to perform a declarative prosody, (2) to make an emphasis on one specific word of each sentence (generally on the verb) and (3) to perform an interrogative prosody. This results in 108 sentences, directly digitalized on a computer (41kHz, 16bits) for each speaker, using an USBPre sound device connected to an omni directional AKG C414B microphone placed 40 cm from the speaker mouth, and performing a high-pass filtering of frequency under 40Hz plus a noise reduction of 6dB.

3.2.2. Calliphony players

4 users have completed the experiment on a subset of 9 sentences ranging from 1 to 9 syllables, either lexicalized or reiterated, and using the three prosodic conditions (declarative, emphasized, interrogative), for the male speaker. All subjects are involved in this work and completely aware of its aims and are therefore familiar with prosody. Three out of the four subjects are trained musicians. One of the four subjects is the male speaker of the original corpus, who has therefore imitated its own voice vocally and by handwriting movements.

3.2.3. Prosodic parameters and distances measures

In order to evaluate the objective distance between the original and the imitated sentences, their pitch values have to be carefully extracted and computed. All the sentences of the corpus were manually analyzed. Their prosodic parameters were automatically extracted: fundamental frequency for vocalic segments (in semitones) and the corresponding voicing strength (calculated from intensity), syllabic duration and intensity thanks to Matlab (the yin script [11]) and Praat [12] programs.

The objectives distances between the prosody of the original sentence and the imitated prosody were calculated on the basis of the physical dissimilarity measures introduces by Hermes [13]: the correlation between the two F0 curves, and the root-mean-square (RMS) difference between theses two curves. The voicing strength was used (as suggested by [13]) as a weighting factor in the calculation of these two distances measures.

Objective distances between the original sentence and each repetition at the output of the Calliphony system were automatically calculated by using 10 ms spaced vector of F0 values for each vocalic segment. Then only the closest imitation, according to the weighted correlation measure and then the weighted RMS distance, was kept for the result analysis. This part of the work can be completely automated, as there is no duration change between the output of Calliphony and the original sentence. This is not the case for the oral imitations, which have to be labeled prior to extract F0 values for vocalic segments.

Moreover the distance computation supposes segments of the same length, a condition not met for vocal imitations. Therefore, only the distances between the original sentences and the gestural imitations have been calculated so far.

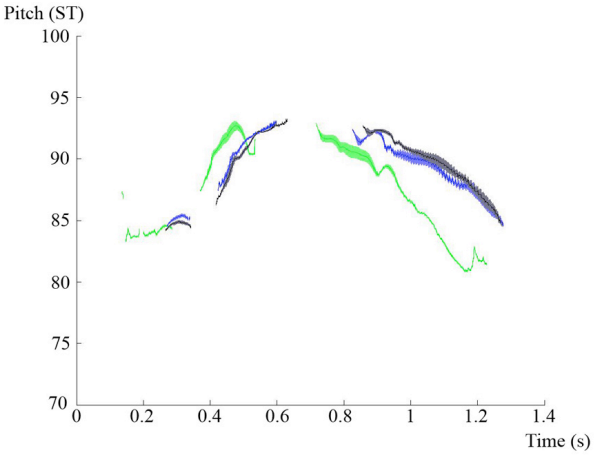


Figure 4: raw F0 value (in tones) for an original sentence (gray) and the two vocal imitations of one subject. Stimuli are not time-aligned.

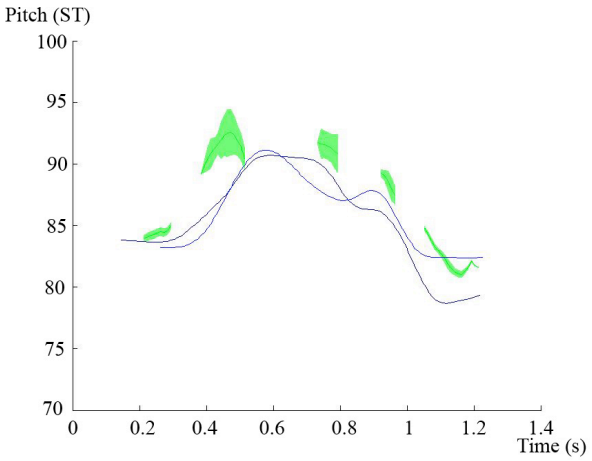


Figure 5: stylized F0 of an original sentence (the same as in fig. 4 – gray curve, smoothed values for the vocalic segment expressed in tones), and the value of the pitch parameter controlled by the graphic tablet for all the imitations performed by one subject. Stimuli are time-aligned.

Graphics with the raw F0 value of both the original and the vocal imitations have been produced in order to visually compare the performances of gesture vs. vocal imitations. Graphics with the stylized F0 of the original sentences (smoothed F0 for the vocalic segments) superimposed with the course of the pen on the graphic tablet were also produced in order to compare the two imitations' modalities (fig. 4 & 5).

3.3. Results

The mean objective distances are summarized in Table 2. There is no major difference between the four users, except for a higher RMS distance for AR, the only non-musician amongst the users (for a discussion about this issue cf. [10]).

Table 2: *mean distances for each subject and for all sentences imitated by handwriting movements.*

Subject	***R***	***RMS***
CDA	0.866	3.108
BD	0.900	3.079
SLE	0.901	3.091
AR	0.898	4.728
Total	0.891	3.502

The prosodic condition (declarative, emphasized, interrogative prosody) did not have a significant impact on the users' performances. The reiterant speech condition neither did.

The most influential factor in the experiment is the length of the sentence, as correlations continuously decrease while the number of syllable increase (cf. figure 6). This result can be explained either by an increasing difficulty of the user's task, or by an artifact due to the sentence length, because computation of correlation does not take into account any weighting for length compensation. More analyses would be needed before concluding on a sentence length effect.

Finally, the most important result of this evaluation procedure is the high overall correlation and low RMS distance obtained by all users. This result generally validates the ability of human users to imitate very closely an original prosody by using handwriting movements. Moreover, the observation of the imitated F0 curves shows a complete smoothing of any micro-prosodic variations: this indicates that users only reproduce prosodic movement at the level of the syllable or above, and that the task adequately matches prosody imitation and generation purposes.

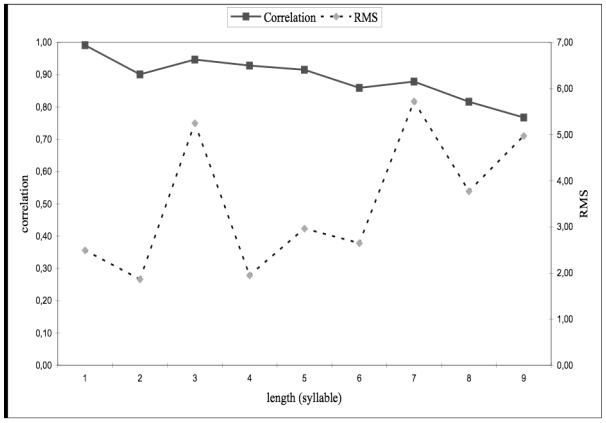


Figure 6: evolution of the two distances measures with the sentence's length. X-axis: length of stimuli, left Y-axis: correlations (plain line), right Y-axis: RMS difference (dotted line).

4. Application to expressive speech synthesis

Since the adequacy of a hand-driven interface to control speech prosody is validated, this section will explore some possible applications of this interface.

4.1. Intonation post-processing

A first application of the Calliphony system is directly derived from the scheme developed for the evaluation of the system: to allow a user to directly change the pitch of a spoken utterance. Such application can be useful in the field of speech synthesizers: as such devices have already reached a high degree of naturalness, they are now seeking for expressivity. The major problem is then to record and adequately model the huge corpora needed to be able to face any kind of expressivity for any sentences.

Our proposal is to give the end user the possibility to directly add the expressivity he needs on the output of his speech synthesizer thanks to the Calliphony system. This system is easy to use and only need little practice. Someone could then easily add e.g. a focalization on a desired word.

4.1.1. Assessment procedure

In order to assess the ability of our system to add such kind of expressivity to synthetic speech, a validation procedure has been set up, and is reported hereafter. It is based on exactly the same principle as the validation of the Calliphony system for prosody imitation reported above, with the only difference being that flattened speech (the input of the Calliphony system) has been replaced here with a synthetic sentence, produced with the Selimsi TTS system [14]. The player of Calliphony hears an original sentence from our corpus, carrying either a focalization on one word or an interrogative prosody. He has then to reproduce the pitch contour of the original sentence on the synthetic sentence, on a similar task that the one described above.

The major difference between the two experiments concerns the segmental duration of the modified stimuli: for the preceding evaluation, the segmental durations are exactly the same as the original, as it is only a flattened version of the natural stimulus, whereas the synthetic sentence has his proper

segments' durations. It induces two major differences between the two protocols. The first one concerns the modification procedure: it is harder to perform an imitation when important lengthening is present in the original sentence. The second one concerns the distance measure between the original and the reproduced pitch contours. As the pitch values are compared for vowel only, and as synthetic and natural vowels don't necessarily have the same duration, instead of extracting one value of pitch for each 10 ms, 10 values for each vowel were calculated, regularly spaced along the vowel. These 10 values per vowel are then used to calculate correlation and RMS distance using the same formulae as those presented above.

Table 3: mean distance scores obtained for focalized sentence, interrogative sentences and for all sentences.

	Correl	RMS
Focalization	0,92	3,18
Interrog.	0,86	4,14
Mean	0,89	3,66

4.1.2. Results and analyses

The results obtained for this assessment are quite similar to those already exposed.

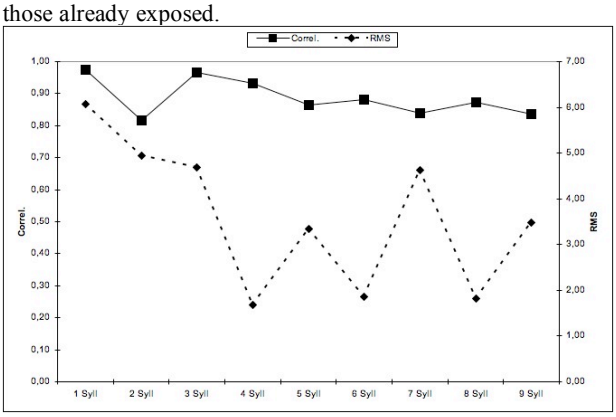


Figure 7: mean distances (correlation and RMS distance) obtained for sentences of all length, from 1 to 9 syllables.

Mean correlation and RMS distances are good (cf. tab 3), and indicate a close stylization of the pitch curve on the synthetic stimuli, even if there are durational differences. Mean score obtained for focalization vs. interrogation sentences are quite similar, with slightly better score for focalization. About the effect of the sentence's length (cf. fig. 7), the effect is a bit more complicated than the one observed with natural speech: if correlation decreases gradually with the sentence length, as it has already been observed, the RMS distance did not have any particular tendency, except for the 1-, 2- and 3-syllables length's sentences, that receive high RMS distances scores, contrary to natural speech.

The objective distance between modified prosodic parameters at the output of Calliphony and the original natural prosody is rather small, giving a very good idea of the system's performances at producing expressive speech.

However, it must be noted that the duration parameter is not dealt with in this first version of Calliphony. This is not satisfactory for high quality expressive synthesis, where durations' modification is mandatory. In addition, the sound quality is better for natural speech modification compared to synthetic speech modification. In our current implementation Calliphony results in two successive modifications of the signal (concatenation and PSOLA modification), a situation that is not optimal indeed. More work is still needed before to obtain a better sounding system, but we think that the ability of players to add expressivity to synthesizers has been convincingly demonstrated.

Considering the databases that are not previously tagged, the system can still be used online in a slightly different manner. When the purpose is only to produce some expressive sentences (for various perceptive experiments for example) then one is able to modify online the synthesized sentences and to record them directly after modification.

This gives the opportunity for someone not necessarily familiar with speech synthesis and processing, to produce expressive sentences in a convenient manner, without having to buy an expensive system or to acquire deep knowledge in speech processing. Moreover, one can use synthesized sentences from any TTS engine publicly available or can directly record sentences on its owns with a simple microphone and recording software, before achieving its modification thanks to our system.

As an extent, thanks to the good quality of expressive modifications on synthetic speech, it could find applications in various research and development situations, going from advertising to industrial mass media, or even animated cartoon characters voice synthesis.

4.2. Data-base enrichment

Another application of the Calliphony system to TTS concerns specifically data-driven speech synthesis. Synthesis systems based on selection/concatenation of non-uniform units need large corpora of recorded speech. Our system can be used for enrichment of the speech database, prior to synthesis. In this case, natural speech is modified, and a same sentence can be given several prosodic variations, as depicted in Fig.8

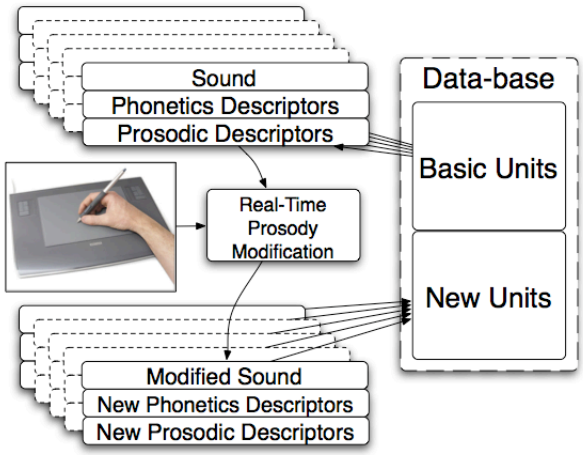


Figure 8: Enrichment of Data-Base with Non-Uniform Units

There are several steps to achieve this enrichment and it can be applied to various types of databases. There are no constraints on the content of the database. The system can then be used to add new expressions that were not recorded, or to have more utterances of a less represented expression.

Then the prosodic content of the database can be extended and/or improved without the need of new recordings. This is independent of the TTS system itself, because it is only a matter of database pre processing.

This application is in a preliminary stage: no formal evaluation of the synthetic speech obtained is available for the moment.

5. Discussion and conclusion

Speech instruments have been an important part of the history of speech synthesis, but have played only a marginal role in

speech synthesis application or research. We think that high quality real-time speech modification algorithms and new high precision interfaces have the potential for dramatically changing the current situation.

In this paper, we explore the ability of handwriting movement for expressive speech synthesis. The system has been called “calliphony”, i.e. expressive speech beyond phonemes by analogy with “calligraphy”, i.e. expressive writing beyond graphemes. The results indicate even untrained players are almost as skilled for vocal imitation as for written imitation of expressive prosody.

Then, the system can be applied to TTS post processing and database pre-processing. TTS post-processing can be a useful extension of a TTS system for tuning synthetic speech utterance output without the need of deep engineering or expensive recordings. The quality obtained is basically the quality of the TTS system itself.

We are currently exploring the quality reached by database enrichment, a pre-processing for augmenting the prosodic content of a selection/concatenation TTS system, without recording new sentences.

Future work will be devoted to duration and tempo modifications. Our experiments show (or confirm) that changing intonation without changing duration or tempo is not enough in many situations. Changing voice quality is also required for more realistic prosodic modifications. Additional control parameters will then be needed.

Another path of research for future work is the interface itself. We are currently pursuing the study of the range of possibility offered by an ad-hoc controller called the Meta-Instrument. This controller offers up to 54 continuous controllers simultaneously, supervised by the fingers and the arms (see [15]).

6. References

[1] D'Alessandro, C., et al. (2005) “*The speech conductor : gestural control of speech synthesis*.” in eNTERFACE 2005. The SIMILAR NoE Summer Workshop on Multimodal Interfaces, Mons, Belgium.

[2] http://www.nime.org/

[3] Cook, P., (2005) *"Real-Time Performance Controllers for Synthesized Singing,.* Proc. NIME Conference, 236Ð237, Vancouver, Canada.

[4] Kessous, L. , (2004) *"Gestural Control of Singing Voice, a Musical Instrument"*. Proc. of Sound and Music Computing, Paris.

[5] Fels, S. & Hinton, G. , (1998) *“Glove-Talk II: A Neural Network Interface which Maps Gestures to Parallel Formant Speech Synthesizer Controls.”* IEEE Transactions on Neural Networks, 9 (1), 205Ð212.

[6] Puckette, M. (1991). *“Combining Event and Signal Processing in the MAX Graphical Programming Environment”*. Computer Music Journal 15(3): 68-77.

[7] http://www.cycling74.com/

[8] E. Moulines and F. Charpentier, (1990) *"Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,"* Speech Communication, vol. 9, pp. 453–467.

[9] http://web.media.mit.edu/tristan/

[10] d’Alessandro, C., Rilliard, A. & Le Beux, S. (Submitted). *“Computerized chironomy: evaluation of hand-controlled intonation reiteration.”* Proc. of InterSpeech 2007.

[11] de Cheveigné, A., Kawahara, H., (2002) *“YIN, a fundamental frequency estimator for speech and music”,* J. Acoust. Soc. Am. 111, 1917-1930.

[12] Paul Boersma & David Weenink, (2001)*“PRAAT, a system for doing phonetics by computer.”* Glot International 5(9/10): 341-345.

[13] Hermes, D.J. (1998). *“Measuring the Perceptual Similarity of Pitch Contours”*. Journal of Speech, Language, and Hearing Research, 41, 73-82.

[14] Prudon, R. and C. d'Alessandro. (2001) “*A selection/concatenation text-to-speech synthesis system: databases development, system design, comparative evaluation*.” in 4th ISCA/IEEE International Workshop on Speech Synthesis.

[15] Serge de Laubier, Vincent Goudard, (2006) *"Méta-Instrument 3 : a look over 17 years of practice",* in Proc. of the NIME Conference, IRCAM, Paris, France .

The Speech Conductor: Gestural Control of Speech Synthesis

Christophe d'Alessandro (1), Nicolas D'Alessandro (2), Sylvain Le Beux (1), Juraj Simko (3),
Feride Çetin(4), Hannes Pirker (5)
(1) LIMSI-CNRS, Orsay, France, (2) FPMS, Mons, Belgium (3) UCD, Dublin, Ireland,
(4) Koç Univ.,Istanbul, Turkey, (5), OFAI, Vienna, Austria

Abstract

The Speech Conductor project aimed at developing a gesture interface for driving ("conducting") a speech synthesis system. Four real-time gesture controlled synthesis systems have been developed. For the first two systems, the efforts focused on high quality voice source synthesis. These "Baby Synthesizers" are based on formant synthesis and they include refined voice source components. One of them is based on an augmented LF model (including an aperiodic component), the other one is based on a Causal/Anticausal Linear Model of the voice source (CALM) also augmented with an aperiodic component. The two other systems are able to utter unrestricted speech. They are based on the MaxMBROLA and MidiMBROLA applications. All these systems are controlled by various gesture devices. Informal testing and public demonstrations showed that very natural and expressive synthetic voices can be produced in real time by some combination of input devices/synthesis system

***Index Terms*—speech synthesis, glottal flow, gesture control, expressive speech.**

I. PRESENTATION OF THE PROJECT

A. Introduction

Speech synthesis quality seems nowadays acceptable for applications like text reading or information playback.

However, these reading machines lack expressivity. This is not only a matter of corpus size, computer memory or computer speed. A speech synthesizer using several times more resources than currently available will probably improve on some points (less discontinuities, more smoothness, better sound) but expression is made of real time subtle variations according to the context and to the situation. In daily life, vocal expressions of strong emotions like anger, fear or despair are rather the exception than the rule. Then a synthesis system should be able to deal with subtle continuous expressive variations rather than clear cut emotions. Fundamental questions concerning expression in speech are still unanswered, and to some point even not stated. Expressive speech synthesis is the next challenge. Expressive speech synthesis may be viewed from two sides: on the one hand is the question of expression specification (what is the suited expression in a particular situation?) and on the other hand is the question of expression realisation (how is the specified expression actually implemented). The first problem (situation analysis and expression specification) is one of the most difficult problems for research in computational linguistics because it involves deep understanding of the text and its context. Without a deep knowledge of the situation defining an adequate expression is difficult, if not impossible.

It is only the second problem that has been addressed in this workshop. Given the expressive specifications, produced and controlled in real time by a "speech conductor", given the intended expression, or an "expression score" for a given speech utterance, how to "interpret" the speech produced according to this intended expression?

The Speech Conductor project aims at developing and testing gesture interfaces for driving ("conducting") a speech or voice synthesis system. The goal is to modify speech synthesis in real time according to the gestures of the "Speech Conductor". The Speech Conductor adds expressivity to the speech flow using speech signal synthesis and modification algorithms and gesture interpretation algorithms. This multimodal project involves sounds, gestures and text.

B. Domains and challenges

The main goal of this project was to test various gesture interfaces for driving a speech synthesiser and then to study whether more "natural" expressive speech (as compared to rule-based or corpus-based approaches) could be produced. The problems addressed during the workshop were:

1. Identify the parameters of expressive speech and their relative importance. All the speech parameters are supposed to vary in expressive speech. In time domain a list of speech parameters would encompass: articulation parameters (speed of articulation, formant trajectories, articulation loci, noise bursts, etc.) phonation parameters (fundamental frequency, durations, amplitude of voicing, glottal source parameters, degree of voicing and source noise etc.). Alternatively, physical parameters (sub glottal pressure, larynx tension) or spectral domain parameters could be used.
2. Signal processing for expressive speech. Techniques for parametric modification of speech: fundamental frequency, duration, articulation, voice source.
3. Domain of variation and typical patterns for expressive speech parameters, analysis of expressive speech.

4. Gesture capturing and sensors. Many types of sensor and gesture interfaces were available. The most appropriates have been selected and tried.
5. Mapping between gestures and speech parameters. The correspondence between gestures and parametric modifications is of paramount importance. This correspondence can be more or less complex (one to many, many to one, one to one). A physiologically inspired model for intonation synthesis has been used.
6. Different types of vocal synthesis have been used. Parametric source/filter synthesis proved useful for accurately controlling voice source parameters. Diphone based concatenative speech synthesis proved useful for more unrestricted speech synthesis applications, but allowed for less fine grained controls. Of course real time implementations of the synthesis systems were needed.
7. Expression, emotion, attitude, phonostylistics. Questions and hypotheses in the domain of emotion research and phonostylistics, evaluation methodology for expressive speech synthesis have only marginally been addressed because of the short time available. For the same reason preliminary evaluation of the results obtained took place on an informal basis only.

C. *Gesture Control Devices*

Several devices, whose controllers and ranges are quite different, where used. At first, we used two keyboards, one Roland PC-200, with 49 keys, a Pitch Bend /Modulation Wheel and one fader. The range of the keyboard is by default between 36 and 84 but can be shifted in order to change the frequency register. The Pitch Bend/Modulation wheel sends values between 0 and 127 according to the MIDI protocol. Thus, these several controllers are respectively sending values on dedicated Note On/Off, Pitch Bend and Control Change channels.

The second keyboard was a Edirol PCR-50 which features 8 knobs and 8 faders in addition to the controls mentioned before. Similarly, in this keyboard the values are set between 0 and 127 and it sends data on several Control Change channels.

In addition to the Roland keyboard we also used an Eobody controller to have some extra knob controls in order to drive the MaxMBROLA Text-To-Speech synthesizer. This sensor interface converts any sensor raw data to MIDI protocol, but as a matter of fact we only used the inbox knobs. We were also able to use a MIDI foot controller providing ten switches in ten different banks and two expression pedals.

A P5 Glove with five flexion sensors linked to the fingers that could bend when fist clench was also employed. The sensors send data in range 0 to 63. Thanks to an Infrared sensor, the glove offers the ability to track the hand position in three spatial dimensions (x,y,z) within a continuous range roughly equal to [–500,+500].

The glove does not actually use MIDI protocol but Open Sound Control (OSC) instead. Contrary to MIDI which sets data in a serial way, under OSC the values are sent in parallel, allowing a fixed rate for every controller.

D. *Overview of the work done*

The work has been organized along two main lines: text-to-speech synthesis and parametric voice quality synthesis. As for text-to-speech synthesis two different configurations have been produced. For one of the systems the only parameter controlled in real time is fundamental frequency. Phonemes and durations are computed automatically by the text-to-speech engine (we used Mary (Schröder & Trouvain, 2003) for English) and then produced by the MBROLA diphone system (Dutoit & al., 1996). For the second system, syllables are triggered by the player. Then durations, fundamental frequency and intensity are controlled using the MidiMBROLA synthesis system (D'Alessandro & al. 2005).

As for parametric voice quality synthesis, coined herein the "Baby Synthesizers" also two different approaches have also been implemented. Both are based on a parametric description of the voice source. In one system, the well-known LF model (Fant & al. 1985, Fant 1995) of the glottal flow derivative has been used, and augmented with an aperiodic component. The other system is based on a spectral approach to glottal flow modelling, the Causal/Anticausal Linear Model, CALM (Doval & al. 2003). This model has also been augmented with an aperiodic component.

In the remaining of this paper, the four systems developed during the workshop will be described in more detail.

II. REAL TIME CONTROL OF AN AUGMENTED LF-MODEL.

A. *The voice source model in the time domain*

In the linear acoustic model of speech production, the effect of the voice source is represented by the time-varying acoustic flow passing through the glottis. When the vocal folds are regularly oscillating (voiced speech), the glottal flow can be represented using a glottal flow model, the most widely used being the Liljencrants-Fant (LF) model (Fant & al. 1985). The glottal flow is the air stream coming from the lungs through the trachea and pulsed by the glottal vibration. All the glottal flow models are pulse like, positive (except in the case of ingressive speech), quasi-periodic, continuous, and differentiable (except at closure). Acoustic radiation of speech at the mouth opening can be approximated as a derivation of the glottal flow. Therefore, the glottal flow derivative is often considered in place of the glottal flow itself. The form of the glottal flow derivative can often be recognized in the speech waveform, with additional formant ripples. The time-domain glottal flow models can be described by equivalent sets of 5 parameters (Doval & d'Alessandro, 1999):

- A_v: peak amplitude of the glottal flow, or amplitude of voicing.
- T_0: fundamental period (inverse of F_0)
- O_q: open quotient, defined as the ratio between the glottal open time and the fundamental period. This

quotient is also defining the glottal closure instant at time O_q*T_0.

- A_m: asymmetry coefficient, defined as the ratio between the flow opening time and the open time. This quotient is also defining the instant T_m of maximum of the glottal flow, relative to T_0 and O_q ($T_m = A_m*O_q*T_0$). Another equivalent parameter is the speed quotient S_q, defined as the ratio between opening and closing times, $A_m = S_q / (1 + S_q)$.
- Q_a: the return phase quotient defined as the ratio between the effective return phase duration (i.e. the duration between the glottal closure instant, and effective closure) and the closed phase duration. In case of abrupt closure $Q_a = 0$.

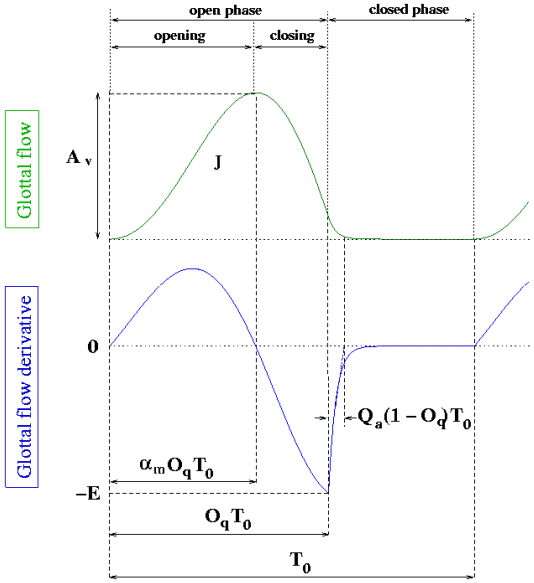


Figure 1: Time domain models of the glottal flow and glottal flow derivative (LF-model),after Henrich & al. 2002.

When considering the glottal flow derivative, the peak amplitude is generally negative, because the closing phase is generally shorter than the opening phase. So the descending slope of the glottal flow is steeper, and its derivative larger. All the time-domain parameters are equivalent for the glottal flow and its derivative except this amplitude parameter:

- E: peak amplitude of the derivative, or maximum closure speed of the glottal flow. Note that E is situated at O_q*T_0, or glottal closure instant. It is often assumed that E represents the maximum acoustic excitation of the vocal tract

E and A_v are both representing a time domain amplitude parameter. One or the other can be used for controlling amplitude, but E appears more consistently related to loudness and should probably be preferred for synthesis. The waveform and derivative waveform of the LF model are plotted in Figure 1.It must be pointed out that an aperiodic component must also be added to the periodic LF model. Two types of aperiodicities have to be considered: structural aperiodicities (jitter and shimmer) that are perturbations of the waveform periodicity and amplitude, and additive noise.

Note that compared to the LF model new parameters are added for controlling the aperiodic component. Shimmer and Jitter are perturbation of T_0 amplitude of the LF model (structural aperiodicities). Filtered white noise is also added to the source for simulating aspiration noise in the voice source. The voice source waveform is then passed in a vocal tract filter to produce vowels. The initial formant transitions have been designed to produce a voiced stop consonant close to /d/. This time-domain "baby synthesizer" based on the augmented LF model is presented in Figure 2. The circles indicate those parameters than can be controlled in real time by the gesture captors

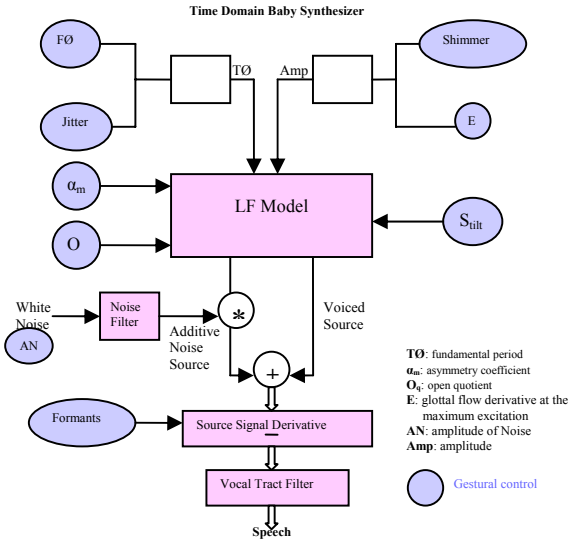


Figure 2. The time-domain "baby synthesizer" implemented in the project, LF model of the source, source aperiodicities and vocal tract filter.

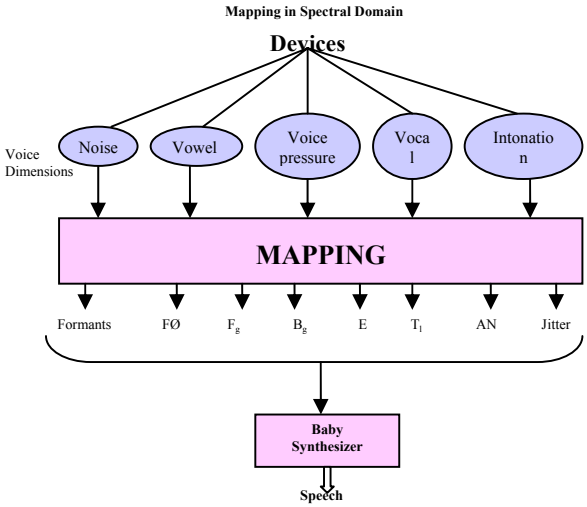
B. *Mapping*

There is no one-to-one correspondence between voice quality and glottal flow parameters. Their relationships are the subject of a large body of work. They can be sketched as follows (d'Alessandro, forthcoming). F_0 describes melody. A very low F_0 generally signals creaky voice and a high F_0 generally signals falsetto voice. O_q describes mainly the lax-tense dimension. O_q is close to 1 for a lax voice, and may be as low as 0.3 for very pressed or tense phonation. As A_v represents the maximum flow, it is an indication of flow voice, and it may help for analysis of the vocal effort dimension. E correlates well with the sound intensity. Q_a

correlates also with the effort dimension. When Q_a =0 the vocal cords close abruptly. Then both E the asymmetry A_m are generally high, and so is vocal effort. Conversely, large values of Q_a (0.05-0.2) give birth to a smooth glottal closure –the vocal effort is low. The asymmetry coefficient A_m has an effect on both the lax-tense dimension (asymmetry is close to 0.5 for a lax voice, and higher for a tense voice) and the vocal effort dimension (asymmetry generally increases when the vocal effort increases). Therefore some sort of mapping between raw voice source parameters and voice quality dimensions is needed.

For controlling of the baby synthesizers, voice quality dimensions are mapped onto voice source acoustic parameters. These voice quality dimensions are then controlled by the gesture captors, as explained in Figure 3.

Figure 3. Mapping in Time domain



C. Gestural control

The augmented LF model has been implemented entirely in the Pure Data environment. The implementation is based on the normalized LF model worked out in (Doval & d'Alessandro 1999).

The way controllers have been mapped to the various synthesizers was somewhat arbitrary. It must be pointed out that controllers could practically be driving any of the several synthesizers we implemented. For the augmented LF model Baby synthesizer the configuration was settled as follows:

- The Edirol MIDI keyboard was driving three voice dimensions. The keys from (from left to right) define the vocal effort, and the velocity of the pressed key was linked to the glottal pressure.
- In order to be able to have a dynamic mapping of these two dimensions we chose to have the possibility to change the parameters driving these dimensions. So that we could easily set the mid value and the span of asymmetry, open quotient and closing phase time, these parameters were each set by two knobs.
- The Pitch Bend/Modulation wheel was respectively controlling Frequency and Volume in such a way that no sound is produced the wheel is released.
- In addition to this, we used the pedal board to switch between the different presets of the vocal tract formants of different predefined vowels (a,e,i,o,u).
- Finally, one expression pedal of this pedal board was use to add noise to the signal generated.

III. Real time control of a Causal/Anticausal Linear Spectral Model

A. The voice source model in the spectral domain

Modelling the voice source in the spectral domain is interesting and useful because the spectral description of sounds is closer to auditory perception. Time-domain and frequency domain descriptions of the glottal flow are equivalent only if both the amplitude and the phase spectrum are taken into account, as it is the case in this work.

The voice source in the spectral domain can be considered as a low-pass system. It means that the energy of the voice source is mainly concentrated in low frequencies (recall that only frequencies below 3.5 kHz were used in wired phones) and is rapidly decreasing when frequency increases. The spectral slope, or spectral tilt, in the radiated speech spectrum (which is strongly related to the source derivative) is at most -6 dB/octave for high frequencies. As this slope is of +6 dB/octave at frequency 0, the overall shape of the spectrum is a broad spectral peak. This peak has a maximum, mostly similar in shape to vocal tract resonance peaks (but different in nature). This peak shall be called here the "glottal formant". This formant is often noticeable in speech spectrograms, where it is referred at as the "voice bar", or glottal formant below the first vocal tract formant.

Spectral properties of the source can then be studied in terms of properties of this glottal formant. These properties are:

1. the position of the glottal formant (or "frequency");
2. the width of the glottal formant (or "bandwith");
3. the high frequency slope of the glottal formant, or "spectral tilt";
4. the height of the glottal formant, or "amplitude".

One can show that the frequency of the glottal formant is inversely proportional to the open quotient O_q (Doval et al. 1997). It means that the glottal formant is low for a lax voice, with a high open quotient. Conversely, a tense voice has a high glottal formant, because open quotient is low.

The glottal formant amplitude is directly proportional to the amplitude of voicing. The width of the glottal formant is linked to the asymmetry of the glottal waveform. The relation is not simple, but one can assume that a symmetric waveform (a low S_q) results is a narrower and slightly lower glottal formant. Conversely, a higher asymmetry results in a broader and slightly higher glottal formant

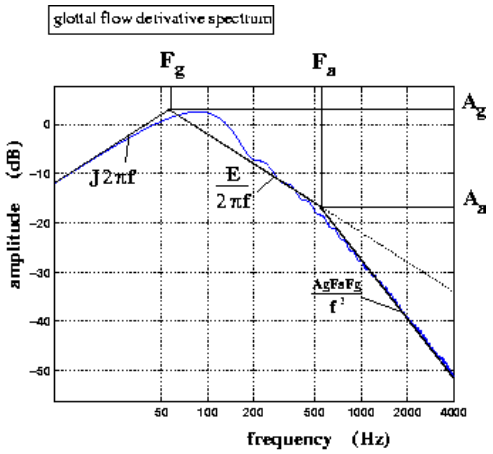


Figure 4. Glottal flow derivative spectrum (after Henrich &al. 2002)

Around a typical value of the asymmetry coefficient (2/3) and for normal values of open quotient (between 0.5 and 1), the glottal formant is located slightly below or close to the first harmonic ($H_1 = f_0$). For O_q=0.4 and A_m=0.9, for instance, it can then reach the fourth harmonic

Up to now, we have assumed an abrupt closure of the vocal folds. A smooth closure of the vocal folds is obtained by a positive Q_a in time domain. In spectral domain, the effect of a smooth closure is to increase spectral tilt. The frequency position where this additional attenuation starts is inversely proportional to Q_a. For a low Q_a, attenuation affects only high frequencies, because the corresponding point in the spectrum is high. For a high Q_a, this attenuation changes frequencies starting at a lower point in the spectrum.

In summary, the spectral envelope of glottal flow models can be considered as the gain of a low-pass filter. The spectral envelope of the derivative can then be considered as the gain of a band-pass filter. The source spectrum can be stylized by 3 linear segments with +6dB/octave, -6dB/octave and -12dB/octave (or sometimes -18dB/oct) slopes respectively. The two breakpoints in the spectrum correspond to the glottal spectral peak and the spectral tilt cut-off frequency

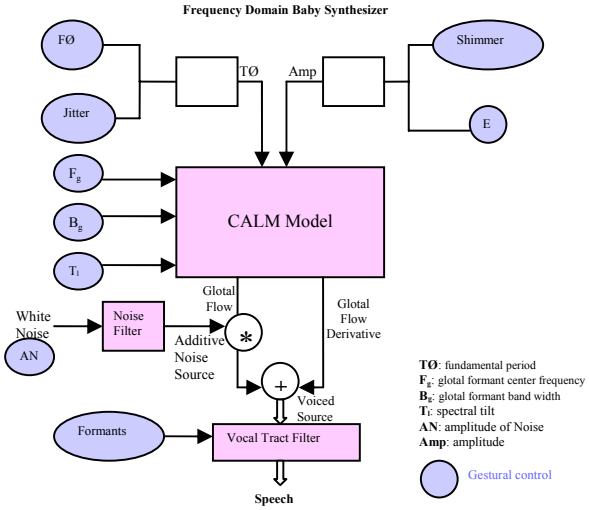
An example displaying linear stylization of the envelope of the glottal spectrum in a log representation is given in Figure 4.

For synthesis in the spectral domain, it is possible to design an all-pole filter which is comparable to e.g. the LF model. This filter is a 3rd order low-pass filter, with a pair of conjugate complex poles, and a simple real pole. The simple real pole is given directly by the spectral tilt parameter. It is mainly effective in the medium and high frequencies of the spectrum. The pair of complex-conjugate poles is used for modeling the glottal formant. If one wants to preserve the glottal pulse shape, and then the glottal flow phase spectrum, it is necessary to design an anticausal filter for this poles pair. If one wants to preserve the finite duration property of the glottal pulse, it is necessary to truncate the impulse response of the filter. The spectral model is then a Causal (spectral tilt) Anti-causal (glottal formant) Linear filter Model (CALM, see Doval & al. 2003). This model is computed by filtering a pulse train by a causal second order system, computed according to the frequency and bandwidth of the glottal formant, whose response is reversed in time to obtain an anti-causal response. Spectral tilt is introduced by filtering this anti-causal response by the spectral tilt component of the model. The waveform is then normalized in order to control accurately the intensity parameter E.

An aperiodic component is added to this model, including jitter, shimmer and additive filtered white noise. The additive noise is also modulated by the glottal waveform.

Then the voice source signal is passed through a vocal tract formant filter to produce various vowels. Figure 6 presents an overview of the spectral "Baby synthesizer.

Figure 6. CALM Model



B. Mapping

This global spectral description of the source spectrum shows that the two main effects of the source are affecting the two sides of the frequency axis. The low-frequency effect of the source, related to the lax-tense dimension is often described in terms of the first harmonic amplitudes H_1 and H_2 or in terms of the low frequency spectral envelope. A pressed voice has a higher H_2 compared to H, and conversely a lax voice has a higher H_1 compared to H. The effort dimension is often described in terms of spectral tilt. A louder voice has a lower spectral tilt, and spectral tilt increases when loudness is lowering.

Then the vocal effort dimension is mainly mapped onto the spectral tilt and glottal formant bandwidth parameters (asymmetry), although the voice pressure dimension depends mostly on the glottal formant centre frequency, associated to open quotient.

Other parameters of interest are structural aperiodicities (jitter and shimmer) and additive noise.

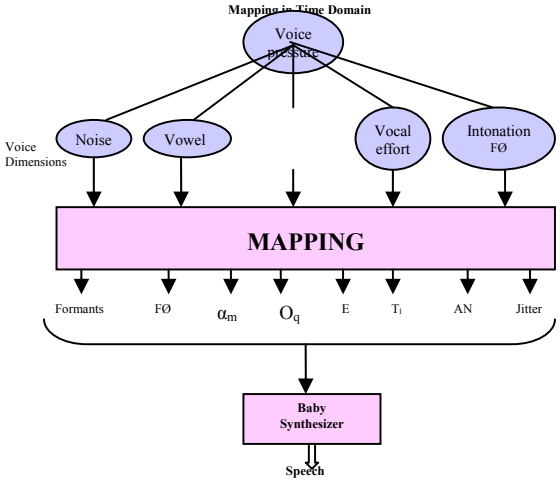


Figure 7. Mapping in Spectral domain

C. *Gestural control of the spectral Baby Synthesizer*

For this synthesizer, a P5 data glove is used. This input device allows driving 8 continuous variable parameters at once: 3 spatial position x, y, z associated with the movement of the glove relative to a fixed device on the table and 5 parameters associated with bending of the five fingers. Several keys on the computer keyboard are controlling vowels. The glove was driving the spectral-domain glottal source model. Only the two horizontal spatial dimensions (x,z) were used as follows: the x variable was linked to intensity E and the z variable was linked to fundamental frequency. All the fingers but the little finger were used to control respectively (beginning from the thumb) noise ratio, Open Quotient, Spectral Tilt and Asymmetry. This mapping is most reliable and effective (compared to the keyboard used in the first experiment). Only a short training phase was sufficient to obtain very natural voice source variations. The computer keyboard was used for changing values of the formant filters for synthesizing different vowels, and then basic vocal tract articulations.

IV. Real time control of F0 in a Text-To-Speech System using MaxMbrola

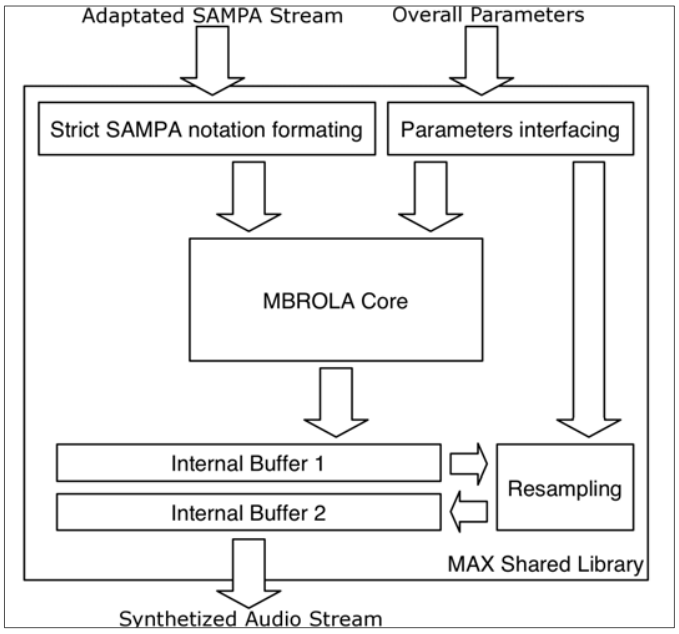
A. *Max/MSP Graphical Programming Environnement*

The Max graphical development environment and its MSP audio processing library (Zicarelli & al., 2004) are widely used the computer music community. This software is a powerful tool in many fields of electronic music like real-time sound processing, control mapping, composition, enhancement of performance abilities etc. It is a rare example of an intuitive interface (design of personalized modules by the building of graphs of simple functions, called *objects*) and a high level of flexibility (functions accepting and modifying numbers, symbols, audio and video stream, etc) at the same time. The capabilities of that software increase every day due to the help of an active developer community providing new *external* objects (or *externals*).

B. *MaxMBROLA~ external object: MBROLA inside Max/MSP*

This section explains how the MBROLA technology has been integrated inside the Max/MSP environment (D'Alessandro & al. 2005). Max/MSP objects work as small servers. They are initialized when they are imported into the workspace. They contain a set of dedicated functions (methods) which are activated when the object receives particular messages. These messages can be simple numbers, symbols or complex messages with a header and arguments. Considering that real-time request-based protocol of communication between objects, a Max/MSP external object containing the MBROLA algorithm has been developed and a particular set of messages (header and arguments) has been formalized to communicate with the synthesizer.

Figure 8. Internal structure of the MaxMBROLA~ external object (after D'Alessandro & al. 2005).



As shown in Figure 8, we can separate the possible requests in two main channels. On one side, there is parameter modification, which influences the internal state of the synthesizer. On the other side, there is the phonetic/prosodic stream, which generates speech instantaneously.

C. *Available actions of the object*

1) *Internal state modifications*

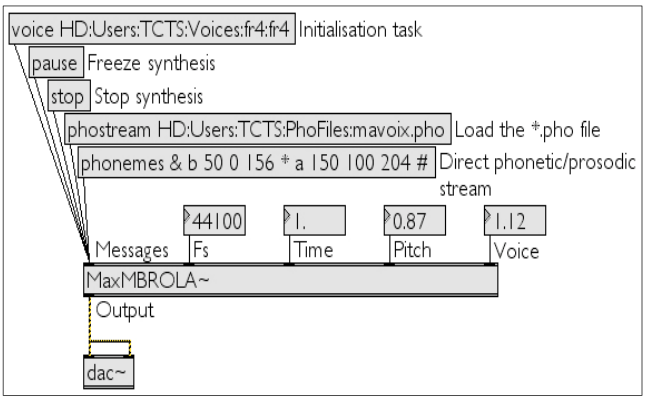
Specific modifications of the internal state of the MBROLA synthesizer can be applied with Max/MSP requests. Here follows a description of the supported actions. The labels are used to name inlets (from left to right: *Messages*, *Fs*, *Time*, *Pitch* and *Voice)* and examples of the supported messages are illustrated on Figure 9.

The synthesizer always starts with the initialization task (*Messages* inlet). Thit function starts the MBROLA engine loads the requested diphone database and set all the internal

parameters to their default values. All the existing MBROLA databases are compatible with this application.

The stream provided by the external can be frozen (*Messages* inlet). It means that the phonetic/prosodic content stays in memory but the MBROLA engine stops the synthesis task.

Figure 9. Supported messages of the MaxMBROLA~ external object.



The MBROLA engine can also be stopped (*Messages* inlet). That function flushes the phonetic/prosodic content, stops the synthesis process and sets all the internal parameters to their default values. The diphone database remains loaded.

Fs inlet receives a floating point number. It controls the output sampling rate. Indeed, the original sampling rate depends on the database (16000Hz or 22050Hz). Linear interpolation is performed allowing the use of that external object with all possible sampling rates.

The inlets *Time*, *Pitch* and *Voice* each receive a floating point number. These values are respectively the time ratio (deviation of the reference speed of speech), the pitch ratio (deviation of the reference fundamental frequency of speech) and voice ratio (compression/dilation ratio of the spectrum width). For each inlet, 1.0 is the default value. The object doesn't transmit values lower than 0.01 (means "100 time lower than the default value").

2) Phonetic/prosodic stream processing

The requests for generating speech in the Max environment are described. All the following messages are sent into the *Messages* inlet.

A loading request allows to use a standard *.pho file (which include the list of phonemes to be produced and the target prosody) to perform synthesis. Examples are available together with MBROLA voices and complete explanations about standard SAMPA (Speech Assessment Methods Phonetic Alphabet). SAMPA is a machine-readable phonetic alphabet used in many speech synthesizers. (Cf. the SAMPA-page http://www.phon.ucl.ac.uk/home/sampa/home.htm).

We developed a function that directly accepts SAMPA streams inside Max messages to provide user control to interactive speech production. The standard SAMPA notation has been modified to fit to the Max message structure. For example, the following stream:

phonemes & b 50 0 156 * a 150 100 204 #

begins by initializing the synthesizer, then produces a syllable /ba/ of 200 (50 + 150) milliseconds with a fundamental frequency increasing from 156Hz to 204Hz (two pitch points). Finally, it flushes the phoneme buffer.

D. Adding Text-to-Phoneme capabilities to MaxMBROLA

MaxMBROLA requires a phonemic specification as input just like it is used in mbrola .pho files, i.e. a transcription in SAMPA with optional information on duration and pitch. MaxMBROLA, just as mbrola, is not intended to be a fully fledged text-to-speech system. Anyway, it is obviously advantageous to combine it more directly with some kind of text-to-phoneme preprocessing in order to increase the flexibility of the system.

It was thus decided to use the text-to-phoneme capabilities provided by the TTS-system Mary (Schröder & Trouvain, 2003).

Mary is a Text-To-Speech system available for German and English. One of its attractive properties is that it offers full access to the results of intermediate processing steps. It provides an XML representation that contains not only the phonemes, their durations and pitch, but also a straightforward encoding of the full prosodic hierarchy which comprises phrases, words and syllables.

As there are applications of MaxMBROLA where the speech is to be synthesized syllable-wise, the latter information is most valuable.

A collection of simple Perl-scripts for parsing and converting Mary-XML format as well as standard mbrola .pho files to the input format required by MaxMBROLA was produced.

Max/MSP provides a "shell"-object which allows the execution of shell-commands, including piping, within a patch. This made the smooth integration of the text-to-phoneme processing rather straightforward.

As Mary is implemented as server-client architecture, as a special treat Mary was currently not installed locally but was accessed via Internet from within Max/MSP.

E. Gestural control of the Text-to-Speech system

Only one parameter, namely fundamental frequency (F0), was controlled by the glove in the MaxMbrola + mary text-to-Speech system. The phoneme stream and segment durations were computed by the TTS system. A flat pitch MBROLA signal was computed according to this data. Then F0 movements were computed by a PSOLA post-processing module receiving the flat MBROLA synthesized speech as input. F0 was modulated in real time, according to the distance between the glove and a fixed device on the table. This very simple control scheme was very effective. Very realistic and expressive prosodic variations were produced almost immediately because controlling F0 this way proved very intuitive.

V. Real time control of F0, durations and intensity in a syllable based Speech synthesis System using MidiMbrola

A. *MIDI-MBROLA: The First MaxMBROLA-based MIDI Instrument*

A Max/MSP musical instrument, called MIDI-MBROLA, has also been developed around the MaxMBROLA external object (D'Alessandro & al. 2005). This tool has a full MIDI compatible interface. MIDI *control changes* are used to modify the internal parameters of the MBROLA synthesizer. *Events* from a MIDI keyboard are used to compute the prosody, which is mixed with the phonetic content at the time of performance. As a standard module of the Max/MSP environment, the MIDI-MBROLA digital instrument automatically allows polyphony. Indeed, many voices can readily be synthesized simultaneously because the MBROLA synthesis doesn't utilize many CPU resources. It can also be compiled as a standalone application or as a VST instrument ("Virtual Studio Technology", a digital effect standard developed by Steinberg) instrument. That tool is publicly available.

B. *Gestural control of MIDI-MBROLA*

The MIDI-MBROLA instrument has been linked to the Roland keyboard and the three knobs of the Eobody Controller. The input text consisted of a syllabic sliced phonetic transcription of the speech utterance. Syllables were triggered by the keyboard. F0 was modulated by the keyboard and pitch-bend. Note that the keyboard has been divided in 1/3 of semitone between to adjacent keys. The Pitch Bend allowed for even smaller pitch excursions. The three knobs were controlling the overall speed, the mid-pitch and the vowel length. But it should be noticed that only the pitch control was effectively driving a parameter in real time whereas the three others were only sampled at syllables frequency (his means that once triggered a syllable was played with a given speed, without variation within the syllable). The configuration used is showed in Figure 9. With this configuration, the output speech had a singing character which sounded rather unnatural for speech. This was because the pitch variations were limited by the discrete nature of the keyboard.

VI. Fujisaki Intonation Modelling

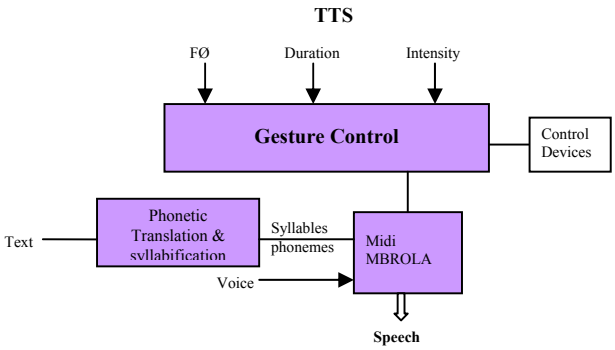
Another strand of development dealt with the implementation of the Fujisaki model of intonation (Fujisak &Hirose, 1984) in the Pure Data environment. This model aims to take physical and physiological processes involved in the production of F0 into account. The main idea is to model the intonation contour by superimposing the results of two different processes. On the one hand there is the phrase component that models the phenomenon of slowly declining global pitch baseline throughout a prosodic phrase. The accent component, on the other hand, is responsible for modeling the local excursions in the F0 contour used for marking pitch accents.

Fujisaki's model has proved its descriptive adequacy in capturing F0 contours for a variety of different languages.

As the input parameters of the model that have to be dynamically controlled can be basically reduced to the triggering of phrase commands, accent commands and their respective amplitudes, it seemed worthwhile to investigate its applicability in a real-time system.

An implementation of the Fujisaki in PureData was produced. In a first experiment the parameters where controlled by a MIDI-keyboard, where attack, release and velocity map quite straightforwardly to the timing and the amplitude of both accent- and phrase commands.

Figure 10. TTS Control Model



VII. Discussion and conclusion

A. *Summary of software produced.*

Four different software projects have been produced during eNTERFACE:

1. the time-domain Baby Synthesizers. A LF model based vowel formant synthesizer, written in Pure Data, and mainly tested with keyboard, joystick and pedal-board real-time interfaces.
2. the spectral domain Baby synthesizer. A CALM model based vowel formant synthesizer, written in Max/MSP, and mainly tested with a digital glove real-time interface.
3. the Mary TTS in English with real-time intonation control, using a digital glove.
4. the MIDI-MBROLA speech synthesizer in French with a real-time control of intonation, duration and intensity using a keyboard with pitch bend.

B. *Comparing patch programming Environments*

Baby Synthesizers were developed using the real-time graphical environments Pure Data (PD) and Max/MSP. PD is an Open Source platform developed and maintained by Miller Puckette and includes code written by wide community of programmers. Max/MSP is commercial software developed by Cycling'74 company.

During this process we also tested some limits of these closely related platforms, and learnt lessons which we share.

Graphical environment

Being a commercial product, Max/MSP environment is better designed and user friendlier. However, simpler PD user

interface wasn't causing any problems in development process.

Stability

No stability issues with Max platform were encountered during the development. On the other hand, Pure Data programmers experienced several challenging problems, when some objects kept changing their behavior, disappearing and reappearing randomly. In general, stability issues were less serious for MacOS then for Windows platform; even system reboot didn't always help…

Richness

PD proved to be slightly more flexible when it came to coding more complex mathematical functions on sound wave in real time. Unlike Max/MSP, it allows a wide variety of mathematical operations to be performed in real-time directly on the sound signal with one very simple universal object. Similar operations had to be coded in C, compiled and imported to MAX/MSP.

Despite of the limitations mentioned above, both of these closely related environments proved to be suitable for sound processing applications of the kind we were developing.

C. Towards expressivity evaluation

Up to now no formal evaluation of the different variants of synthesizers has been performed. As a matter of fact, the evaluation of the "quality" of a speech synthesis system is not a trivial task in general, and is even more complicated when it comes to the evaluation of expressivity.

Usually synthesis systems are evaluated in terms of intelligibility and "naturalness". For the former there exist a number of established tests (Gibbon et al. 1997). Typically samples of isolated syllables or nonsense words are presented and it is possible to perform a quantitative evaluation of correctly perceived samples. When evaluating the "naturalness" of synthesized speech, an objective measure is less straightforward. In the simplest case, a comparison between two systems or between two variants of a system by forced preference choice can be performed. Another method is the rating of the "adequacy" of a synthesized sample for a given context. But again it is difficult to impossible to come up with an objective independent evaluation.

In the field of the synthesis of expressive speech, the predominant evaluation method is to synthesize sentences with neutral meaning and encode a small set of "basic" emotions (typically joy, fear, anger, surprise, sadness). Subjects are then asked to identify the emotional category.

A competing evaluation model is to use more subtle expressive categories: use test sentences with non-neutral semantics, and let again rate the adequacy of the sample for a given context.

In the context of the Speech-Conductor project it was only possible to perform an informal comparison of the two synthesizers that implemented glottal source models. At the current state the CALM based model gives much better "impression" than the time-domain model. On the other hand there are still a number of slight differences in the actual implementation of these two models; e.g. the differences in the modeling of jitter and shimmer or the automatic superimposing of micro-prosodic variations, that have a strong impact on the perceived "quality" of the models.

A more interesting evaluation would be a rating test for the recognizably of perceptual voice quality measures such as laxness/tenseness, vocal effort etc. Though this would be probably a promising method of evaluating the current state, it is not easy to perform, as it would rely on the availability of independent "expert" listeners with a certain amount of phonetic experience.

In this context it would thus be interesting to further investigate whether it is possible to get reliable ratings on voice quality factors from so called "naive listeners".

For the MaxMBROLA system different evaluation methods have to be taken into account, as this is basically a classical diphone-synthesis system which allows for the real-time control of prosodic features, most prominently pitch. Thus the evaluation methods used for "normal" concatenative synthesis systems could easily applied. One of the peculiarities of this system is that inevitable the virtuosity of the person "conducting" the synthesizer is a strong factor in the quality of the output.

A straightforward evaluation would be a rating test of different input devices (e.g. Data Glove vs. Keyboard), but apart from the "human factor", currently still too many differences in the underlying synthesis scenarios exist to allow a real comparison.

D. Conclusion

Devices:

The glove performed much better than the keyboard or joysticks for controlling intonation and expressivity. However, the tested glove model had some performance limitations (it proved too slow for real time). However, the glove wasn't tested for its capacity to reproduce the intended gesture precisely and reliably.

Keyboard on the contrary allows for exact reproducibility of gestures. When combined with TTS synthesizers the produced speech had somewhat singing quality, as pitch changes are directly linked to syllable onsets.

Synthesizers:

In general, voice source models produced much more expressive vocal utterances than TTS models. For TTS, better results were reached when speech was generated using pre-computed segment durations and intensity and we only controlled F_0. So, surprisingly, less control can in some situations yield better results. In any case, it's clear that to add a real expressivity, flexible control of all of the voice source parameters is needed.

To our best knowledge, this project was the first attempt to implement real-time system of gestural control of expressive speech. The results proved really encouraging, and opened a new avenue for expressive speech synthesis research.

ACKNOWLEDGEMENTS

Hannes Pirker states that his research is carried out within the Network of Excellence Humaine (Contract No. 507422) that is funded by the European Union's Sixth Framework Programme with support from the Austrian Funds for Research and Technology Promotion for Industry (FFF 808818/2970 KA/SA). This publication reflects only the author's views. The European Union is not liable for any use that may be made of the information contained herein.

REFERENCES

(Bozkurt et al., 2005) B. Bozkurt, B. Doval, C. D'Alessandro, and T. Dutoit, "Zeros of Z-Transform Representation With Application to Source-Filter Separation in Speech" IEEE Signal Processing Letters, Vol. 12, No. 4, April 2005, p 344-347

(d'Alessandro, 2005) C.d'Alessandro, "*Voice source parameters and prosodic" analysis,* in Methods in Experimental prosody research, Mouton de Gruyter (in press)

(d'Alessandro & Doval, 2003) C. d'Alessandro, B. Doval, "Voice quality modification for emotional speech synthesis", Proc. of Eurospeech 2003, Genève, Suisse, pp. 1653-1656

(D'Alessandro et al., 2005)N. D`Alessandro, B. Bozkurt, T. Dutoit, R. Sebbe, 2005, "MaxMBROLA: A Max/MSP MBROLA-Based Tool for Real-Time Voice Synthesis", Proceedings of the EUSIPCO'05 Conference, September 4-8, 2005, Antalya (Turkey).

(Doval & d'Alessandro, 1999) B. Doval, C. d'Alessandro, 1999. *The spectrum of glottal flow models.* Notes et Documents LIMSI 99-07, 22p.

(Doval & d'Alessandro, 1997) B.Doval and C.d'Alessandro. *Spectral correlates of glottal waveform models: an analytic study.* In International Conference on Acoustics, Speech and Signal Processing, ICASSP 97, pages 446--452, Munich, avril 1997. Institute of Electronics and Electrical Engineers

(Doval et al., 2003) B. Doval, C. d'Alessandro, and N. Henrich, "The voice source as a causal/anticausal linear filter," in *Proc. ISCA ITRW VOQUAL 2003*, Geneva, Switzerland, Aug. 2003, pp. 15–19

(Dutoit et al., 1996) T. Dutoit, V. Pagel, N. Pierret, F. Bataille and O. van der Vrecken, ``The MBROLA project: towards a set of high quality speech synthesizers free of use for non commercial purposes" Proc ICSLP, Philadelphia, pp. 1393-1396, 1996.

(Fant et al., 1985) Fant G., Liljencrants J. and Lin Q. (1985) "A four-parameter model of glottal flow". STL-QPSR 4, pp. 1-13.

(Fant, 1995) G. Fant, "The LF-model revisited. Transformation and frequency domain analysis," *Speech Trans. Lab. Quarterly .Rep., Royal Inst. of Tech. Stockholm*, vol. 2-3, pp. 121-156, 1995.

(Fels, S. 1994) Fels, "Glove talk II: Mapping hand gestures to speech using neural networks," Ph.D. dissertation, Univ. Toronto, Toronto, ON, Canada, 1994.

(Dutoit, 1997) Dutoit T. An Introduction to Text-To-Speech Synthesis. Kluwer Academic Publishers, 1997.

(Fujisaki & Hirose, 1984) H. Fujisaki, K. Hirose Analysis of voice fundamental frequency contours for declarative sentences of Japanese Journal of Acoustic Society. Jpn. (E) 5, 4. 1984

(Gibbon et al., 1997) Gibbon, D., Moore, R. & Winsky, R. (Eds) Eagles handbook of Standards and Resources for Spoken Language Systems (1997) Mouton de Gruyter

(Henrich et al. 2002) N. Henrich, C. d'Alessandro, B. doval. `` Glottal flow models: waveforms, spectra and physical measurements". Proc. Forum Acusticum 2002, Séville 2002

(MIDI, 1983) "MIDI musical instrument digital interface specification 1.0," Int. MIDI Assoc., North Hollywood, CA, 1983.

(Schröder, 2004) M. Schröder "Speech and emotion research", Phonus, Nr 7, june 2004, ISSN 0949-1791, Saarbrücken

(Schröder & Trouvain, 2003) M. Schröder & J. Trouvain (2003). The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. International Journal of Speech Technology, 6, pp. 365-377.

(Zicarelli et al., 2004b) Zicarelli, G. Taylor, J. K. Clayton, J. and R. Dudas, MSP 4.3 Reference Manual. and Max 4.3 Reference Manual. Cycling'74/Ircam, 1990-2004.

(Wanderley & Depalle; 2004) M. Wanderley and P. Depalle, "Gestural Control of Sound Synthesis", Proc. of the IEEE, 92, 2004, p. 632-644.

http://mary.dfki.de

http://tcts.fpms.ac.be/synthesis/maxmbrola/

http://www.disc2.dk/tools/SGsurvey.html

Real-time CALM Synthesizer
New Approaches in Hands-Controlled Voice Synthesis

N. D'Alessandro
TCTS Lab (FRIA Researcher)
Faculté Polytechnique de Mons
B-7000 Mons, Belgium
nicolas.dalessandro@fpms.ac.be

C. d'Alessandro, S. Le Beux, B. Doval
LIMSI - CNRS
Université Paris Sud XI
F-91403 Orsay, France
{cda, slebeux, boris.doval}@limsi.fr

ABSTRACT

In this paper, a new voice source model for real-time gesture–controlled voice synthesis is described. The synthesizer is based on a causal-anticausal model of the voice source, a new approach giving accurate control of voice source dimensions like tenseness and effort. Aperiodic components are also considered, resulting in an elaborate model suitable not only for lyrical singing but also for various musical styles playing with voice qualities. The model is also tested using different gestural control interfaces : data glove, keyboard, graphic tablet, pedal board. Depending on parameter-to-interface mappings, several instruments with different musical abilities are designed, taking advantage of the highly expressive possibilities of the synthesis model.

Keywords

Singing synthesis, voice source, voice quality, spectral model, formant synthesis, instrument, gestural control.

1. INTRODUCTION

Remarkable achievements have been recently reached in singing voice synthesis. A review of state of the art can be found in [1]. Technology seems mature enough for replacing vocals by synthetic singing, at least for backing vocals [2] [3]. However, existing singing synthesis systems suffer from two restrictions : they are aiming at mimicking singers rather than creating new instruments, and are generally limited to MIDI controllers.

We think it worthwhile to extend vocal possibilities of voice synthesizers and design new interfaces that will open new musical possibilities. On the one hand, a voice synthesizer should be able to reproduce several voice quality dimensions, resulting in a wide variety of sounds (e.g. quasi-sinusoidal voice, mixed periodic aperiodic voice, pressed voice, various degrees of vocal effort, etc.). On the other hand, vocal instrument being embodied in the singer, multidimensional control strategies should be devised for externalizing gestural controls of the instrument.

In this paper, a new elaborate voice source model able to produce various voice qualities is proposed. It is based on spectral modelling of voice source [4]. Links between spectral parameters and auditory effects are relatively straightforward. Then playing instruments based on spectral modelling seems very intuitive. Another key point is gesture-to-parameter mapping. Following the pioneering work by Fels [5], we found data glove particularly well suited to vocal expression. Recent work on hand-controlled vocal synthesis include series of instruments presented by Cook [6] and the Voicer by Kessous [7]. It must be pointed out that musical possibilities offered by an instrument strongly depend on mapping and interfaces. Then, depending on intended musical aims, different instruments are proposed. This paper is organized as follows. In section 2, the voice synthesis model is reviewed. In section 3, control devices and mapping of voice quality dimensions onto control parameters are discussed. Section 4 presents two musical instruments built on basis of synthesis model and vocal dimensions. Section 5 presents a discussion of results obtained and proposes directions for future works.

NIME 06, June 4-8, 2006, Paris, France

2. VOICE SYNTHESIS MODEL

In this section, we first give an overview of mechanisms involved in voice production. Then, we focus on the glottal source and present the causal-anticausal linear model developed by d'Alessandro/Doval/Henrich in [4]. We also explain the nature of non-periodical components we introduced in the model. Finally, we describe structure and possibilities of the real-time glottal flow synthesizer based on CALM (RT-CALM) we developed and integrated in following singing instruments.

2.1 Voice production

Voice organ is usually described as a "source/filter" system. Glottal source is a non-linear volume velocity generator where sound is produced by complex movements of vocal folds (larynx) under lungs pressure. A complete study of glottal source can be found in [8]. Sounds produced by the larynx are then propagated in oral and nasal cavities which can be seen as time-varying filtering. Finally, the volume velocity flow is converted into radiated pressure waves through lips and nose openings (cf. Figure 1).

In the context of signal processing applications, and particulary in singing synthesis, some simplifications are usually accepted. First, lips and nose openings effect can be seen as derivative of the volume velocity flow. It is generally processed by a time-invariant high-pass first order linear filter [9]. Vocal tract effect can be modelized by filtering of glottal signal with multiple (4 or 5) second order resonant linear filters (formants).

Contrary to this "standard" vocal tract implementation, plenty of models have been developed for represen-

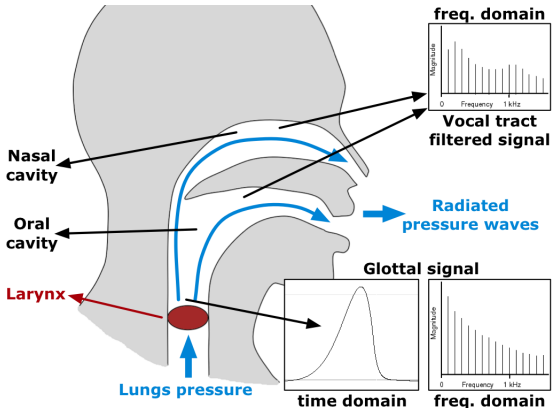


Figure 1: Voice production mechanisms : vocal folds vibrations, vocal tract filtering and lips/nose openings radiation.

tation of glottal flow, with differences in accuracy and flexibility. Usual models are KLGLOTT88 [10], R++ [11], Rosenberg-C [12] and LF [13] [14]. We present now the causal-anticausal linear model (CALM) [4], explain why we worked with this spectral approach and propose adaptations of the existing algorithm to ease real-time manipulation.

2.2 CALM : causal-anticausal linear model

We have seen that modelling vocal tract in spectral domain (with resonant filters central frequency, amplitude and bandwidth) is very powerful in term of manipulation because spectral description of sounds is close to auditory perception. Traditionally, glottal flow has been modelized in time domain. A spectral approach can be seen as equivalent only if both amplitude and phase spectra are considered in the model.

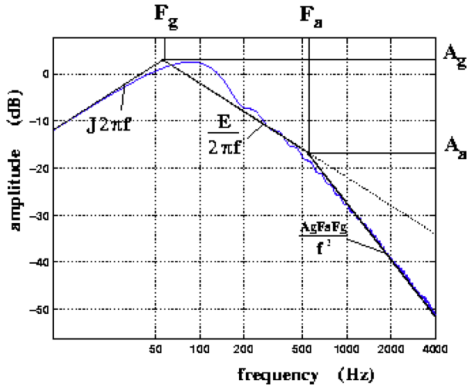


Figure 2: Amplitude spectrum of the glottal flow derivative : illustration of glottal formant (F_g, A_g) and spectral tilt (F_a, A_a).

For amplitude spectrum, two different effects can be isolated (cf. Figure 2). On the one hand, an amount of energy is concentrated in low frequencies (i.e. below 3 kHz). This peak is usually called "glottal formant". We can see that bandwidth, amplitude and position of the glottal formant change with voice quality variations. On the other hand, a variation of spectrum slope in higher frequencies (called "spectral tilt") is also related to voice quality modifications.

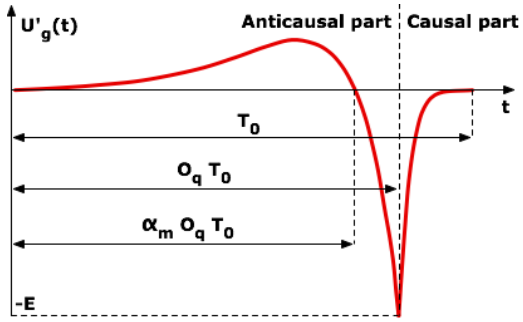


Figure 3: Time-domain representation of derivated glottal pulse : anticausal part and causal part.

Considering both "glottal formant" and "spectral tilt" effects, two cascading filters are implemented. A second order resonant low-pass filter (H_1) for glottal formant, and a first order low-pass filter (H_2) for spectral tilt. But phase information indicates us that this system is not completely causal. Indeed, as it is illustrated on Figure 3, glottal pulse is a combination of a "increasing" (or active) part and a "decreasing" (or passive) part. The decreasing part, called the return phase, mainly influences the spectral tilt and hence is causal. And we can also show that the second order low-pass filter has to be anticausal in order to provide a good phase representation.

A complete study of spectral features of glottal flow, detailed in [4], gives us equations linking relevant parameters of glottal pulse (F_0 : fundamental frequency, O_q : open quotient, α_m : asymetry coefficient and T_l : spectral tilt, in dB at 3000Hz) to H_1 and H_2 coefficients. Note that expression of b_1 has been corrected. [4] also contains equations linking this time-domain parameters with spectral-domain parameters.

Anticausal second order resonant filter :

$$H_1(z) = \frac{b_1 z}{1 + a_1 z + a_2 z^2}$$

where :

$$a_1 = -2e^{-a_p T_e} cos(b_p T_e),\ a_2 = e^{-2a_p T_e}$$

$$b_1 = \frac{E}{b_p} e^{-a_p T_e} sin(b_p T_e)$$

$$a_p = -\frac{\pi}{O_q T_0 tan(\pi \alpha_m)},\ b_p = \frac{\pi}{O_q T_0}$$

Causal first order filter :

$$H_2(z) = \frac{b_{T_L}}{1 - a_{T_L} z^{-1}}$$

where :

$$a_{T_L} = \nu - \sqrt{\nu^2 - 1},\ b_{T_L} = 1 - a_{T_L}$$

$$\nu = 1 - \frac{1}{\eta},\ \eta = \frac{e^{\frac{1}{-T_L/10ln(10)}} - 1}{cos(2\pi \frac{3000}{Fe}) - 1}$$

2.3 Non-periodical components

As described theoretically in [4], the glottal flow is a deterministic signal, completely driven by a set of parameters. Adding naturalness involves the use of some random components we propose to describe.

Jitter

Jitter is a natural unstability in the value of fundamental frequency. It can be modelized by a random value (gaussian distribution, around 0 with variance depending on the

amount of jitter introduced), refreshed every period, added to the stable value of fundamental frequency.

Shimmer
Shimmer is a natural unstability in the value of the amplitude. It can be modelized by a random value (gaussian distribution, around 0 with variance depending on the amount of shimmer introduced), refreshed every period, added to the stable value of amplitude.

Turbulences
Turbulences are caused by additive air passing through vocal folds when glottal closure is not complete. It can be modelized by pink noise filtered by a large band-pass (tube noise), modulated in amplitude by glottal pulses.

We can note here that we kept a direct control on irregularities (based on *Jitter*, *Shimmer* and *Turbulences* rates). Other models were developed, involving granular synthesis coupled with self-organizing dynamic systems [15], and could be considered in further works.

2.4 RT-CALM framework

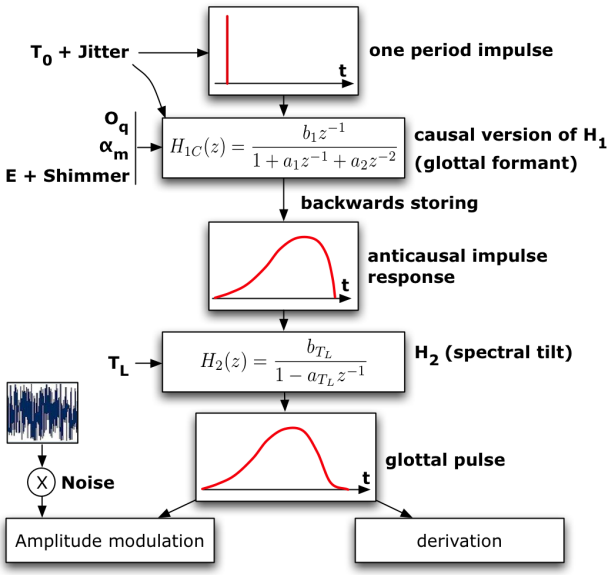


Figure 4: Framework of RT-CALM algorithm, allowing real-time synthesis of glottal pulses based on causal-anticausal linear model.

Full anticausal processing is only possible offline, by running algorithms backwards on data buffers. Anyway, in this context, we can take advantage of physical properties of glottis to propose a real-time algorithm. Indeed, glottal pulse corresponds to opening/closing movements of vocal folds. It means that impulse responses generated by H_1 and H_2 filters can't overlap. Thus, if ranges of parameters are correctly limited, impulse responses can be stored backwards and truncated period-synchronously without changing too much their spectral properties.

To achieve the requested waveform, impulse response of causal version of H_1 (glottal formant) is computed, but stored backwards in the buffer. This waveform is truncated at a length corresponding to instantaneous fundamental frequency (F_0+*Jitter*). Then the resulting period is filtered by H_2 (spectral tilt). Coefficients of H_1 and H_2 are calculated from equations described in subsection 2.2 and [4]. Thus, both time-domain and spectral-domain parameters can be sent. On the one hand, glottal pulses are derivated to produce pressure signal (cf. Figure 3). On the other hand, it is used to modulate the amount of additive noise. Complete RT-CALM algorithm is illustrated at Figure 4.

3. VOICE QUALITY DIMENSIONS

Voice synthesis model is driven by a set of low-level parameters. In order to use these parameters in singing, they must be organized according to musical dimensions. Mappings between parameters and dimensions, and between dimensions and controllers are essential parts of instrument design. In this section, we describe main musical dimensions for voice source (cf. Figure 5) and vocal tract (cf. Figure 6).

3.1 Glottal source

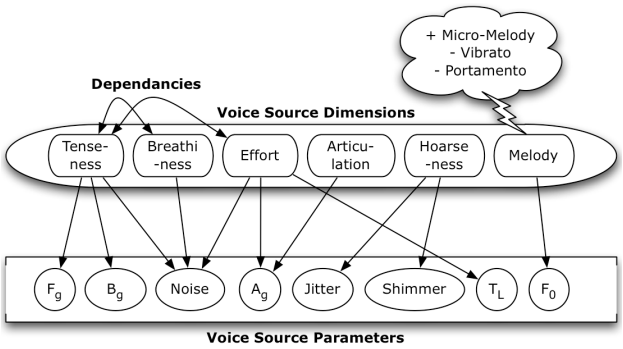


Figure 5: Mapping of the vocal source

Melodic dimension
For singing, this dimension can be decomposed into two parts. On the one hand, it seems important to sing in tune i.e. to make use of notes with well-defined pitches. On the other hand, micro-melodic variations are essential for expressive and natural singing (portamento, vibrato, etc.). Two different controls seem necessary for melodic dimension. This dimension mainly depends on parameter F_0. Anyway, a more precise vibrato synthesis should also involve amplitude variations.

Hoarseness dimension
This dimension is linked to structural aperiodicities in voice source, like *Jitter* and *Shimmer*.

Breathiness dimension
This dimension is linked to aspiration noise in voice source. It controls the relative amount of voicing vs. whispering, using the *Noise* parameter.

Pressed/lax dimension
This dimension is mainly linked to the position of the glottal formant F_g and its bandwidth B_g. It is often linked to breathiness and vocal effort. The pressed/lax dimension is used in some styles of singing e.g. Japanese noh theater or belt singing.

Vocal effort dimension
This dimension is linked to spectral tilt T_l and of course to gain parameter A_g.

3.2 Vocal tract

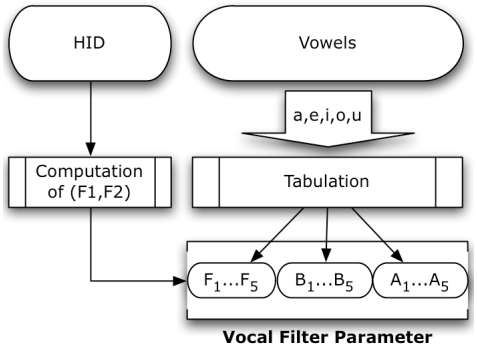


Figure 6: Mapping of the vocal tract

Vocalic space
This space is defining vocal genre (male/female/child), phonemes, and other expressive features (lips rounding, lips spreading, tongue position). This space can also be used for harmonic singing. The vocalic space is defined by formant parameters F_1, B_1, A_1, F_2, B_2, A_2,... F_N, B_N, A_N.

Articulation dimension
Finally, notes attacks and decays are controlled by an articulation dimension. "Articulation" is taken here in its musical meaning i.e. transitions between notes. It is essentially controlled by gain parameter A_g.

3.3 Musical control of vocal dimensions

Playing with melody
Melodic playing usually requires precise pitches. Then "selection" gestures are needed using e.g. a keyboard. However, natural vocal note transitions are generally slow, with more or less portamento and vibrato. Small and controlled pitch variations are therefore needed, and the "selection" gesture must be accompanied by a "modification" gesture, using e.g. hand position in one dimension of space. Another elegant solution offering accurate pitch control and smooth micro-melodic variation is using a graphic tablet. A virtual guitar board can be emulated this way. Well tuned pitches are not required in some singing styles imitating speech, like Sprechgesang (parlar cantando). Then only one control gesture is needed, that can be achieved by position of hand in one spatial dimension.

Playing with timbre : vocalic space
Playing with vocalic timbre is often used on slow moving melodies e.g. harmonic singing. The basic vocalic space needs two dimensions for contrasting vowels e.g. a joystick or a graphic tablet. One dimension is sufficient for harmonic singing (moving only second formant frequency), using a slider or position of hand in one spatial dimension. But a third dimension would be needed for signaling facial movements like lips spreading or rounding, using e.g. a data glove.

Playing with timbre : noise and tension
Some musical styles are also playing with noise and tension. These parameters are moving relatively slowly, on a limited scale, and gestures must not be extremely precise. They can be naturally associated to flexion of fingers in a data glove.

Playing with articulation and phrasing The data glove proved also useful for articulation (in the musical meaning of note attack and release) and phrasing. Hand movements in space are well suited to phrasing and finger flexions are well suited to articulation.

4. CALM-BASED INSTRUMENTS

This section describes two setups we realised. Main purpose of this work was to realize extensive real-time tests of our CALM synthesis model and voice quality dimensions mappings. No dedicated controllers were designed for this purpose. Only usual devices such as tablets, joyticks or keyboards were used.

4.1 Instrument 1

In this first instrument implementation, we use a keyboard to play MIDI notes in order to trigger the vowels at different tuned pitches. Thus, by using keyboard, we are able to set glove free for fine tuning of F_0 so as to achieve vibrato, portamento of other types of melodic ornaments. Accurate control of F_0 by glove position alone proved difficult because well tuned notes references were missing, due to approximative nature of hand gestures.

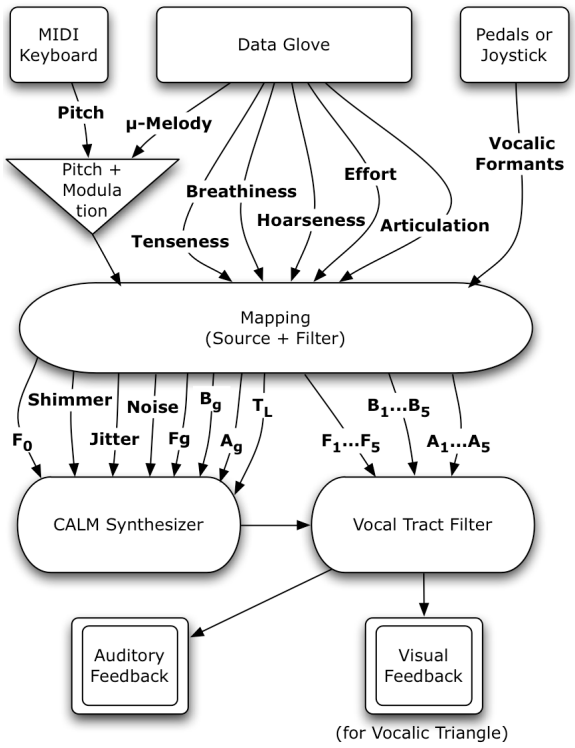


Figure 7: Structure of Instrument 1

Tempered notes (or other conventions) delivered by keyboard can be modified to a certain extent, thanks to tracking of glove position along a certain axis (transversal axis gives better ergonomics as one don't have to fold the elbow to achieve vibrato). General gain is mapped onto longitudinal axis of the glove. Then both vibrato and amplitude envelopes of sound can be produced by circular hand movements. Other vocal dimensions are controlled by flexion of

data glove fingers. First finger controls vocal effort (spectral tilt), second finger controls breathiness (linked to additive noise), third finger control the pressed/lax dimension (linked to the glottal formant), fourth finger controls hoarseness (linked to jitter and shimmer). Voice quality modifications are achieved by closing/opening movements of whole hand or selected fingers. Preset vowels are associated to keys of computer keyboard. Vowel formants can also be modified by additional devices, like pedal board or joysticks.

In summary, for this first instrument :

1. left-hand controls the keyboard (tempered notes)
2. right-hand movements control both fine pitch modulation, and note phrasing.
3. right-hand fingers control tension, effort, hoarseness, and breathiness.

In this implementation, note phrasing results of relatively large hand movements. An alternative solution is to couple effort and note phrasing in fingers movements, and to keep one dimension of hand movement for controlling another vocal dimension (e.g. breathiness). Then, phrasing is controlled by smaller and quicker finger movements. Overall description of this instrument and its various components is illustrated on Figure 7.

4.2 Instrument 2

The key point of this second instrument is simplicity of learning and using. Different choices have been made to achieve that result. First, we decided to focus on voice quality. Vocal tract control would be limited to vowel switching. Then, we took advantage of our natural writing abilities to map all glottal flow features only on tree dimensions of a graphic tablet (x axis, y axis and pressure).

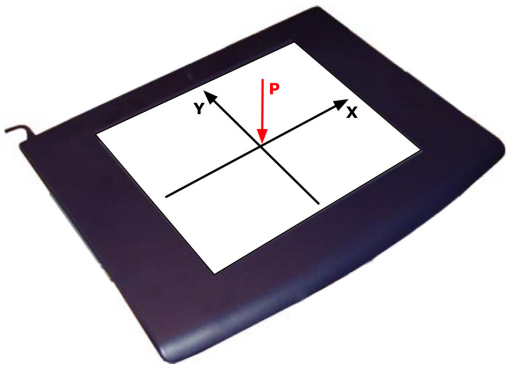


Figure 8: Mapping on the graphic tablet. X axis : fundamental frequency, Y axis : pressed/lax and vocal effort dimensions, Pressure (P) : general volume.

As described on Figure 8, horizontal axis is mapped to fundamental frequency. Tests have been made showing that, after a few training, 2 or 3 (even 4) octaves can be managed on a *Wacom Graphire* tablet. Anyway, transposition and surface scaling features have been implemented. Vertical axis control both pressed/lax and vocal effort dimensions. Mapping is made by using Y value as an interpolation factor between two different configurations of parameters O_q, α_m and T_L, from a "quiet" voice to a "tensed" voice (cf. Figure 9). Finally, pressure parameter is mapped to the gain (E).

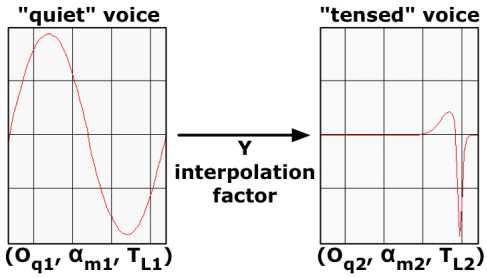


Figure 9: Interpolation between "quiet" voice and "tensed" voice made by Y axis of the graphic tablet.

Regression of voice quality control on an overall expressive axis makes main manipulations of voice source possible with simple "drawings" (i.e. bidimensional + pressure shapes). This compromise makes this instrument really intuitive. Indeed, as it can be done e.g. with a guitar, interpreter only needs graphic tablet to play. MIDI controller (e.g. pedal board) is just used for changing presets (cf. Figure 10).

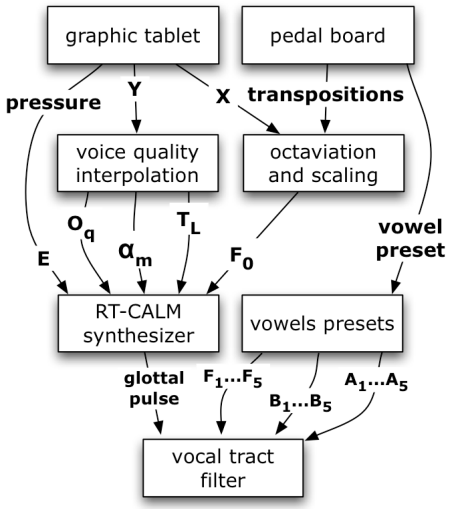


Figure 10: Structure of Instrument 2.

5. CONCLUSIONS AND FUTURE WORK

The two instruments implemented so far are suitable for musical use. Instruments have a truly human sound, and new possibilities offered by gestures to sound mapping enable intuitive playing. Compared to other voice synthesis systems, more emphasis is put on voice quality controls. It is then possible to play with expressive musical dimensions inherent to wind instruments, like effort, pressure and noise. These dimensions are exploited in acoustic instruments like saxophones and brass, and of course voice, but are generally ignored in singing synthesis. Hand movements in space and hand/fingers closures/openings are intuitively associated to such dimensions as effort or voice pressure.

Another challenging point for singing synthesis is accurate yet flexible F_0 control, like in fretless string instruments. This has been implemented in two ways in our

instruments (graphic tablet and glove controlled F_0). This flexible F_0 control enables the player all possible types of intonation, from singing to speech. Melodic ornaments like e.g. vibrato or portamento are easily controlled.

Spectral processing of voice quality proved also useful for "spectral" singing styles. Overtone singing, formant melodies, various types of throat singing are easily produced and controlled in this framework.

Instruments can also be considered as tools for studying singing, because they produce very natural sounding and controlled signals. Then they can be used for investigating musical gestures involved in singing.

Apart from the two instruments presented here, we are also investigating other types of data gloves and elaborated 3D joysticks for refining control of the synthesizer. However, this will not change the nature and number of useful vocal dimensions, but improve precision and ergonomics.

Of course, singing is an instrument that mixes together music and language. Thus, our next challenge is to control the "speech" part of singing. This point has been only marginally considered in the present research and will be the object of future work. Addition of speech articulations would drive us to more accurate modelization of vocal tract, eventually based on existing databases. Considering interfaces, syntaxic abilities of controllers have to be determined in order to achieve syllables, words or sentences synthesis.

6. ACKNOWLEDGMENTS

This work originated from the Speech Conductor project, a part of the eNTERFACE'05 workshop organized by Prof. Thierry Dutoit (Faculté Polytechnique de Mons) within the SIMILAR Network of Excellence (European Union - FP6). We would like to thank all these institutions that provided excellent working conditions.

7. REFERENCES

[1] M. Kob, "Singing Voice Modelling As We Know It Today," *Acta Acustica United with Acustica*, Vol. 90, pp. 649–661, 2004.

[2] Virsyn Corporation, "The Cantor Singing Synthesis Software," 2005-present, url : http ://www.virsyn.de/

[3] Yamaha Corporation, "The Vocaloid Singing Synthesis Software," 2003-present, url : http ://www.vocaloid.com/

[4] B. Doval, C. d'Alessandro and N. Henrich, "The Voice Source as an Causal-Anticausal Linear Filter," *Proc. ISCA ITRW VOQUAL'03*, Geneva, Switzerland, August 2003, pp. 15–19.

[5] S. Fels and G. Hinton, "Glove-TalkII : A Neural Network Interface which Maps Gestures to Parallel Formant Speech Synthesizer Controls," *IEEE Transactions on Neural Networks*, Vol 9, No. 1, pp. 205–212, 1998.

[6] P. Cook, "Real-Time Performance Controllers for Synthesized Singing," *Proc. NIME 2005*, Vancouver, Canada, May 2005, pp. 236–237

[7] L. Kessous, "Gestural Control of Singing Voice, a Musical Instrument," *Proc. of Sound and Music computing 2004*, Paris, October 20-22, 2004.

[8] N. Henrich. *"Etude de la source glottique en voix parlée et chantée."* Thèse de doctorat à l'Université Paris VI, 2001.

[9] G. Fant. *"Acoustic Theory of Speech Production"*, Gravenhage, 1960.

[10] D. Klatt and L. Klatt, "Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers," *J. Acoust. Soc. Am.*, 87(2) :820–857, 1990.

[11] R. Veldhuis, "A Computationally Efficient Alternative for the Liljencrants-Fant Model and its Perceptual Evaluation," *J. Acoust. Soc. Am.*, 103 :566–571, 1998.

[12] A. E. Rosenberg, "Effect of Glottal Pulse Shape on the Quality of Natural Vowels," *J. Acoust. Soc. Am.*, 49 :583–590, 1971.

[13] G. Fant, J. Liljencrants, and Q. Lin, "A Four-Parameter Model of Glottal Flow," *STL–QPSR*, 85(2) :1–13, 1985.

[14] G. Fant, "The LF-Model Revisited. Transformations and Frequency Domain Analysis," *STL–QPSR*, 2–3, 119–56, 1995.

[15] E. R. Miranda, "Generating Source-Streams for Extralinguistic Utterances". *Journal of the Audio Engineering Society (AES)*, Vol. 50 N°3, March 2002.

Comparing time domain and spectral domain voice source models for gesture controlled vocal instruments

Christophe d'Alessandro[1], Nicolas D'Alessandro[2], Sylvain Le Beux[1], Boris Doval[1]

[1]LIMSI-CNRS, BP 133 — F-91403 Orsay, France {cda;Sylvain.le.beux,doval}@limsi.fr

[2] TCTS - FPMs, Mons, Belgium, nicolas.dalessandro@fpms.ac.be

Abstract

Three real-time gesture controlled vocal instruments are presented. They are based on a time domain (LF) and a spectral domain (CALM) model of the glottal pulse signal. Gestural control is able to add expression to the synthetic voices, enabling simulation of various vocal behaviors. Expressive vocal instruments are demonstrated for musical and research purposes.

1. Introduction

Gesture controlled vocal instruments provide new tools for vocal studies. Taking advantage of the fast growing field of music technology, accurately controlled real time voice synthesis systems open new domains of investigation. On the one hand, analysis by synthesis has been recognized since several decades as a much powerful paradigm for speech and voice analysis. On the other hand, most of the research on voice source synthesis until now considered mainly fine grained voice source parameters rather than the domains of variation and co-variation of these parameters. The time span of voice source model is typically one pitch period, although the time span for voice quality perception encompasses several pitch periods or even a full speech sentence or musical phrase. When using real time vocal instruments, the questions of parameter variation and co-variation can no more be ignored or underestimated. Examples of such domains of variation are voice source mechanisms, the phonetogram, co-variation of voice open quotient and noise, vibrato and notes transitions and so one. Gesture controlled voice instruments are also useful for direct perceptive assessment of voice source model as the sounds produced are easily controlled by the "player" and evaluated by both the "player" and the "audience". Finally, real time voice instruments can serve as new musical instruments.

Following the "Speech Conductor" [1] project, this research aims at developing and testing gesture interfaces for driving ("conducting") voice and speech synthesis systems. Gestural control is able to add expression to the synthetic voices, enabling simulation of various vocal behaviours [8]. Then expressive vocal instruments can be designed, for musical and research purposes. In this communication, after a review of voice quality dimensions (section 2) two voice source models for real-time gesture–controlled voice synthesis are compared (section 3). Three vocal instruments are presented (section 4)and discussed (section 5).

2. Voice quality dimensions

No general agreement is currently available on the dimensions of phrase-level voice quality variations (i.e. variations of the vocal activity above the level of the pitch period). However dealing with these dimensions is of the utmost importance for voice analysis, synthesis and perception. A possible framework can be sketched using five main prosodic dimensions:

A. The voice register dimension: Voice registers depend on the underlying voice mechanisms. The different voice mechanisms are corresponding to different voice parameter settings. Changes between mechanisms are usually corresponding to voice "breaks", i.e. sudden voice parameter changes..

B. The noise dimension The noise dimension represents the relative amount of noise in the speech signal, an indication of breathiness or hoarseness. Noise is an important phrase level feature of the voice source.

C. The pressed/lax dimension The pressed/tense dimension corresponds mainly to changes in open quotient. It is a stylistic feature in speech and singing related to relaxed or strangled voices.

D. The effort dimension. The vocal effort dimension corresponds to the nuance *piano* or *forte* in singing. In speech it is used for signalling accentuation. Vocal effort seems relatively independent of vocal pressure.

E. The melodic dimension. The voice has very specific melodic patterns. These patterns are including in singing *vibrato*, *portamento*, note transitions and in speech intonation patterns.

3. Glottal pulse models

3.1 Augmented LF- model

The most widely used glottal flow model is the Liljencrants-Fant (LF) model [2]. This time-domain glottal flow model can be described by equivalent sets of 5 parameters: 1: A_v: peak amplitude of the glottal flow, or amplitude of voicing; 2: T_0: fundamental period (inverse of F_0); 3: O_q: open quotient, defined as the ratio between the glottal open time and the fundamental period. This quotient is also defining the glottal closure instant at time O_q*T_0. 4: A_m: asymmetry coefficient, defined as the ratio between the flow opening time and the open time. This quotient is also defining the instant T_m of maximum of the glottal flow, relative to T_0 and O_q ($T_m= A_m*O_q*T_0$). Another equivalent parameter is the speed quotient S_q, defined as the ratio between opening and closing times, $A_m = S_q / (1 +S_q)$; 5: Q_a: the return phase quotient defined as the ratio between the effective return phase duration (i.e. the duration between the glottal closure instant, and effective closure) and the closed phase duration. In case of abrupt closure $Q_a = 0$.

For realistic voice synthesis, an aperiodic component must also be added to the periodic LF model (called

then Augmented LF model, or A-LF). Two types of aperiodicities have to be considered: structural aperiodicities (jitter and shimmer) that are perturbations of the waveform periodicity and amplitude, and additive noise.

3.2 A-LF parameters and phonation dimensions

There is no one-to-one correspondence between voice qualities and glottal flow parameters. They can be sketched as follows: F_0 describes melody. A very low F_0 generally signals creaky voice and a high F_0 generally signals falsetto voice. O_q describes mainly the lax-tense dimension. O_q is close to 1 for a lax voice, and may be as low as 0.3 for very pressed or tense phonation. As A_v represents the maximum flow, it is an indication of flow voice, and it may help for analysis of the vocal effort dimension. Q_a correlates also with the effort dimension. When Q_a =0 the vocal cords close abruptly. Then the asymmetry A_m is generally high, and so is vocal effort. Conversely, large values of Q_a (0.05-0.2) give birth to a smooth glottal closure –the vocal effort is low. The asymmetry coefficient A_m has an effect on both the lax-tense dimension (asymmetry is close to 0.5 for a lax voice, and higher for a tense voice) and the vocal effort dimension (asymmetry generally increases when the vocal effort increases). Therefore some sort of mapping between raw voice source parameters and voice quality dimensions is needed.

3.3 Spectrum of the voice source

Modelling the voice source in the spectral domain is interesting and useful because the spectral description of sounds is closer to auditory perception. Time-domain and frequency domain descriptions of the glottal flow are equivalent only if both the amplitude and the phase spectrum are taken into account, as it is the case in this work.

The voice source in the spectral domain can be considered as a low-pass system. It means that the energy of the voice source is mainly concentrated in low frequencies and is rapidly decreasing when frequency increases. The spectral slope, or spectral tilt, in the radiated speech spectrum (which is strongly related to the source derivative) is at most -6 dB/octave for high frequencies. As this slope is of +6 dB/octave at frequency 0, the overall shape of the spectrum is a broad spectral peak. This peak has a maximum, mostly similar in shape to vocal tract resonance peaks (but different in nature). This peak shall be called here the “glottal formant'”. This formant is often noticeable in speech spectrograms, where it is referred at as the “voice bar”, or glottal formant below the first vocal tract formant.

Spectral properties of the source can then be studied in terms of properties of this glottal formant. These properties are: 1: the position of the glottal formant (or “frequency”); 2: the width of the glottal formant (or “bandwith”); 3: the high frequency slope of the glottal formant, or “spectral tilt”; 4:the height of the glottal formant, or “amplitude”.One can show that the frequency of the glottal formant is inversely proportional to the open quotient O_q [4]. It means that the glottal formant is low for a lax voice, with a high open quotient. Conversely, a tense voice has a high glottal formant, because open quotient is low.

The glottal formant amplitude is directly proportional to the amplitude of voicing. The width of the glottal formant is linked to the asymmetry of the glottal waveform. The relation is not simple, but one can assume that a symmetric waveform (a low S_q) results is a narrower and slightly lower glottal formant. Conversely, a higher asymmetry results in a broader and slightly higher glottal formant.

Around a typical value of the asymmetry coefficient (2/3) and for normal values of open quotient (between 0.5 and 1), the glottal formant is located slightly below or close to the first harmonic ($H_1 = f_0$). For O_q=0.4 and A_m=0.9, for instance, it can then reach the fourth harmonic

Up to now, we have assumed an abrupt closure of the vocal folds. A smooth closure of the vocal folds is obtained by a positive Q_a in time domain. In spectral domain, the effect of asmooth closure is to increase spectral tilt. The frequency position where this additional attenuation starts is inversely proportional to Q_a. For a low Q_a, attenuation affects only high frequencies, because the corresponding point in the spectrum is high. For a high Q_a, this attenuation changes frequencies starting at a lower point in the spectrum.

In summary, the spectral envelope of glottal flow models can be considered as the gain of a low-pass filter. The spectral envelope of the derivative can then be considered as the gain of a band-pass filter. The source spectrum can be stylized by 3 linear segments with +6dB/octave, -6dB/octave and -12dB/octave (or sometimes -18dB/oct) slopes respectively. The two breakpoints in the spectrum correspond to the glottal spectral peak and the spectral tilt cut-off frequency

3.4 Causal/Anticausal Linear Model

For synthesis in the spectral domain, it is possible to design an all-pole filter which is comparable to e.g. the LF model. This filter is a 3rd order low-pass filter, with a pair of conjugate complex poles, and a simple real pole. The simple real pole is given directly by the spectral tilt parameter. It is mainly effective in the medium and high frequencies of the spectrum. The pair of complex-conjugate poles is used for modeling the glottal formant. If one wants to preserve the glottal pulse shape, and then the glottal flow phase spectrum, it is necessary to design an anticausal filter for this poles pair. If one wants to preserve the finite duration property of the glottal pulse, it is necessary to truncate the impulse response of the filter. The spectral model is then a Causal (spectral tilt) Anti-causal (glottal formant) Linear filter Model (CALM, see [3]). This model is computed by filtering a pulse train by a causal second order system, computed according to the frequency and bandwidth of the glottal formant, whose response is reversed in time to obtain an anti-causal response. Spectral tilt is introduced by filtering this anti-causal response by the spectral tilt component of the model. The waveform is then normalized in order to control the amplitude.

An aperiodic component is added to this model, including jitter, shimmer and additive filtered white noise. The additive noise is also modulated by the glottal waveform. Then the voice source signal is passed through a vocal tract formant filter to produce various vowels.

3.4 CALM parameters and phonation dimensions

This global spectral description of the source spectrum shows that the two main effects of the source are affecting the two sides of the frequency axis. The low-frequency effect of the source, related to the lax-tense dimension is often described in terms of the first harmonic amplitudes H_1 and H_2 or in terms of the low frequency spectral envelope. A pressed voice has a higher H_2 compared to H_1, and conversely a lax voice has a higher H_1 compared to H_2. The effort dimension is often described in terms of spectral tilt. A louder voice has a lower spectral tilt, and spectral tilt increases when loudness is lowering.

Then the vocal effort dimension is mainly mapped onto the spectral tilt and glottal formant bandwidth parameters (asymmetry), although the voice pressure dimension depends mostly on the glottal formant centre frequency, associated to open quotient.

Other parameters of interest are structural aperiodicities (jitter and shimmer) and additive noise.

4. Vocal instruments

Instrument 1 MIDI controlled A-LF

The real-time augmented LF model is implemented entirely in the Pure Data environment. The implementation is based on the normalized LF model worked out in [4].

A MIDI controller (MIDI master keyboard) is driving the A-LF model along three voice dimensions. The keys from (from left to right) define the vocal effort, and the velocity of the pressed key is linked to the glottal pressure.

In order to have dynamic mapping of these two dimensions we chose to have the possibility to change the parameters driving these dimensions. So that we could easily set the mid value and the span of asymmetry, open quotient and closing phase time, these parameters are each set by two knobs.

The Pitch Bend/Modulation wheel is respectively controlling Frequency and Volume in such a way that no sound is produced the wheel is released.

In addition to this, we used the pedal board to switch between the different presets of the vocal tract formants of different predefined vowels (a,e,i,o,u).

Finally, one expression pedal of this pedal board is used to add noise to the signal generated. This instrument could serve as a real time interface for the A-LF model, but it is not particularly easy to play.

Instrument 2: CALM, keyboard and glove

The second class of instruments explores hand movements in space and hand closure/opening gestures. This application is written in the MAX environment [9]. Both types of gestures seem well suited for accurate control of voice quality dimensions like melody, effort, voice pressure, hoarseness and breathiness. For this synthesizer, a P5 data glove is used. This input device allows driving 8 continuous variable parameters at once: 3 spatial positions x, y, z associated with the movement of the glove relative to a fixed device on the table and 5 parameters associated with bending of the five fingers. Several keys on the computer keyboard are controlling vowel presets. The glove is driving the CALM. Only the two horizontal spatial dimensions (x,z) are used as follows: the x variable is linked to intensity E and the z variable is linked to fundamental frequency. All the fingers but the little finger are used to control respectively (beginning from the thumb) noise ratio, Open Quotient, Spectral Tilt and Asymmetry. This mapping is most reliable and effective (compared to the keyboard used in the first experiment). Only a short training phase seems sufficient to obtain very natural voice source variations. The computer keyboard is used for changing values of the formant filters for synthesizing different vowels, and then basic vocal tract articulations.

However this instrument seems not easily playable for "classical style" singing. The melodic transitions allowed by the keyboard do not have a typically singing quality. Hand closure/opening gestures are somewhat comparable to opening closure gesture in the vocal tract and adduction/abduction of the vocal folds. This analogy is potentially useful for synthesis and will be pursued in our future work.

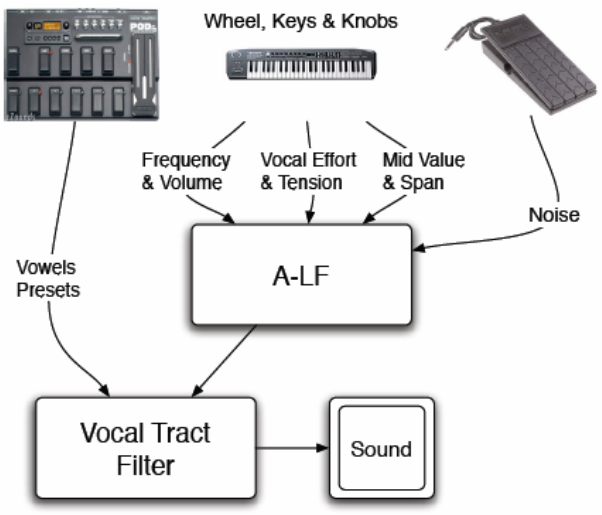


Figure 1: Instrument 1. Midi master keyboard and augmented LF model

Instrument 3: CALM calligraphic singing

The third type of instruments makes use of writing-like gestures with the help of a graphic tablet. It is also written in the MAX environment [9] The graphic tablet is organised along two spatial dimensions and a pressure dimension. The X-axis corresponds to the melodic axis. It is organized in the left to right direction from bass to treble in the same way as a musical keyboard or a guitar. The Y-axis corresponds to vocal effort. The dimensions of vocal effort and voice pressure are mapped from bottom (piano) to top (forte) along this axis. Pressure of the pen is driving the general volume of the voice. This instrument is surprisingly easy to play and expressive. Many vocal effects are possible, including *vibrato*, *portamento*, *messa di voce*, *staccato* and *legato*.

5. Discussion

Key points of this research are the number and nature of voice quality dimensions. Five main dimensions have been identified and controlled by specific gestures: vocal effort (related to spectral richness and amplitude), vocal tension (related to the glottal formant, voice open

quotient and asymmetry), fundamental frequency, breathiness and hoarseness. Several mappings between control devices and model parameters have been proposed for implementing these voice quality dimensions, depending on the underlying voice source model.

The results obtained by the two voices source models seem very close in terms of sound quality. However the CALM model appears less demanding in terms of computational load and is more intuitively controlled. This is because spectral parameters are perceptually close to voice quality dimensions. The spectral model also gives a simple framework for vocal tract modelling (using formant synthesis) and source-filter interactions.

Their sound quality and their playability render these instruments usable for musical purposes [5] [6] [7]. We plan to use these types of instruments for expressive speech synthesis and for expressive prosody research. Our future work will include implementation of a low-dimension physical model of the vocal folds (e.g. a 2-mass model) in the same real-time synthesis environment.

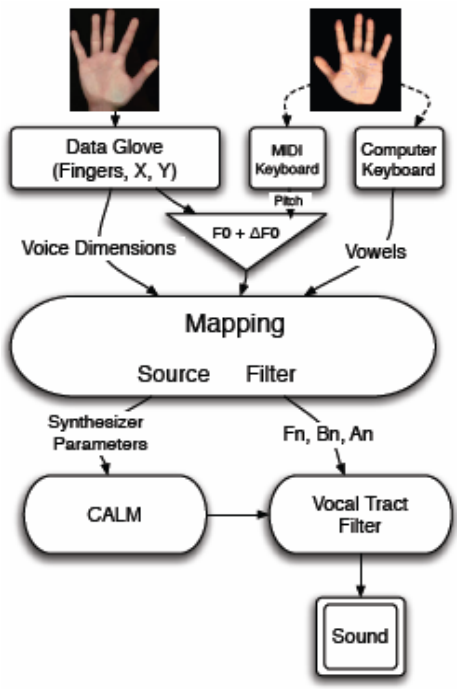


Figure 2: Instrument 2. Data glove controlled CALM synthesizer.

References

[1] C. d'Alessandro, N. D'Alessandro, S. Le Beux, J. Simko, F. Çetin, H. Pirker (2005) "The Speech Conductor: Gestural Control of Speech Synthesis", Proc. of the SIMILAR-Enterface'05 workshop, Presses univ. de Louvain, ISBN : 2-87463-003-9, p. 52-61.

[2] G. Fant, (1995) "The LF-Model Revisited. Transformations and Frequency Domain Analysis," STL–QPSR, 2–3, 119–56, 1995.

[3] B. Doval, C. d'Alessandro and N. Henrich, (2003) "The Voice Source as an Causal-Anticausal Linear Filter," Proc. ISCA ITRW VOQUAL'03, Geneva, Switzerland, pp. 15–19.

[4] B. Doval, C. d'Alessandro and N. Henrich, (2006) "The spectrum of glottal flow models" Acustica united with Acta Acustica in press.

[5] N. D'Alessandro, C. d'Alessandro, S. Le Beux, B. Doval, "Real-time CALM Synthesizer New Approaches in Hands Controlled Voice Synthesis", Proc. Int Conf. on New Interfaces for Musical Expression, NIME 2006, Paris, June 2006, p 266-271.

[6] P. Cook, (2005) "Real-Time Performance Controllers for Synthesized Singing," Proc. NIME 2005, Vancouver, Canada, May 2005, pp. 236–237

[7] L. Kessous, (2004) "Gestural Control of Singing Voice, a Musical Instrument," Proc. of Sound and Music computing 2004, Paris, October 20-22, 2004.

[8] M. Wanderley and P. Depalle, "Gestural Control of Sound Synthesis", Proc. of the IEEE, 92, 2004, p. 632-644.

[9] D. Zicarelli, G. Taylor, J. K. Clayton, jhno, and R. Dudas, Max4.3 Reference Manual, MSP4.3 Reference Manual. cycling'74/Ircam, 1994-2004.

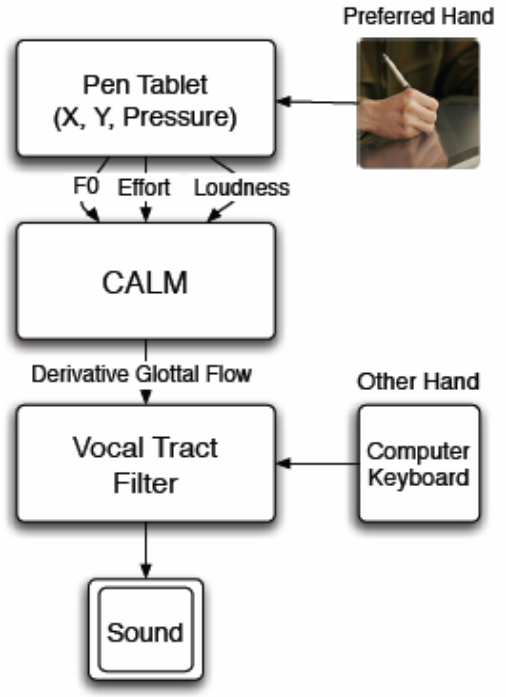


Figure 3: Instrument 3. graphic tablet controlled CALM synthesizer.

RAMCESS: Realtime and Accurate Musical Control of Expression in Singing Synthesis

N. D'Alessandro[1], B. Doval[2], S. Le Beux[2], P. Woodruff[1] and Y. Fabre[1]
[1]*TCTS Lab, Faculté Polytechnique de Mons (Belgium)*, [2]*LIMSI-CNRS, Université Paris XI (France)*

***Abstract*— The main purpose of this project is to develop a full computer-based musical instrument allowing realtime synthesis of expressive singing voice. The expression will result from the continuous action of an interpreter through a gestural control interface. That gestural parameters will influence the voice caracteristics thanks to particular mapping strategies.**

***Index Terms*— Singing voice, voice synthesis, voice quality, glottal flow models, gestural control, interfaces.**

I. Introduction

EXPRESSIVITY is nowadays one of the most challenging topics in view by the researchers in speech synthesis. Indeed, recent synthesizers provide acceptable speech in term of intelligibility and naturalness but the need to improve human/computer intercations brought researchers to develop more "human", more expressive systems. Some recent realizations have shown that an interesting option was to record multiple databases corresponding to a certain number of "labelled" expressions (e.g. happy, sad, angry, etc) [1]. At synthesis time, the expression of the virtual speaker is set by choosing the units in the corresponding database.

Last year, during eNTERFACE'05 [2], we decided to investigate the opposite option. Indeed, we postulated that "emotion" in speech was not the result of switchs between labelled expressions but a continuous evolution of voice caracteristics extremely correlated with context. Thus, we developed a set of flexible voice synthesizers "conducted" in realtime by an operator [3]. After some tests, it was clear that such a framework was particularly efficient for singing synthesis.

Remarkable achievements have been recently reached in singing voice synthesis. A review of state of the art can be found in [4]. Technology seems mature enough for replacing vocals by synthetic singing, at least for backing vocals [5] [6]. However, existing singing synthesis systems suffer from two restrictions: they are aiming at mimicking singers rather than creating new instruments, and are generally limited to MIDI controllers.

We think it worthwhile to extend vocal possibilities of voice synthesizers and design new interfaces that will open new musical possibilities. In a first attempt we decided to restrain our survey on voice quality control to the boundaries of natural voice production. As a matter of fact, it is always better trying to mimic one particular voice, as we are disposed to hear someone behind the synthesizer. This process enables to achieve analysis by synthesis : once we we are able to perceive more naturalness in the synthesized voice, this means that we understood something in voice production process. It is then easier to go astray from these limits when dealing with a musical application in a more creative way.

II. Aims of the Work

Our aims for this eNTERFACE'06 workshop can be summerized in three main axes. First, we target the implementation of intra- and inter-dimensional mappings driving low-level parameters of source models (e.g. complex interactions between vocal effort and tenseness, represented by the phonetogram). Then, we investigate the effects of the vocal tract in voice quality variations (e.g. the singer formant, lowering of the larynx). Finally, source/filter coupling effects (e.g. relations between harmonics and formants frequencies) are analysied, and several mecanisms are implemented (e.g. overtone, croatian, bulgarian, occidental singing).

III. Background in Singing Synthesis

Speech and singing both result from the same production system: the voice organ. However, the signal processing techniques developed for their synthesis evolved quite differently. One of the main reasons for this deviation is: the aim for producing voice is different for the two cases. The aim of speech production is to exchange messages. For singing, the main aim is to use the voice organ as a musical instrument. Therefore a singing synthesis system needs to include various tools to control (analyze/synthesize or modify) different dynamics of the acoustic sound produced: duration of the phonemes, vibrato, wide range modifications of the voice quality, the pitch and the intensity, etc. some of which are not needed in most of the speech synthesis systems. A pragmatic reason for that separation is that singing voice synthesizers target almost exclusively musical performances. In this case, "playability" (flexibility and real-time abilities) is much more important than intelligibility and naturalness. Discussions about various issues of singing synthesis can be found in [7] [8].

As described in [9], frequency-domain analysis/modifications methods are frequently preferred in singing synthesis research due to the need to modify some spectral characteristics of actual recorded signals. The most popular application of such a technique is the phase vocoder [10], which is a powerful tool used for many years for time compression/expansion, pitch shifting and cross-synthesis.

To increase flexibility, short-time signal frames can be modeled as sums of sinusoids (controlled in frequency, amplitude and phase) plus noise (controlled by the parameters of a filter which is excited by a white noise). HNM (Harmonic plus

Noise Model) [11] provides a flexible representation of the signal, which is particularly interesting in the context of unit concatenation. That representation of signals is thus used as a basis in many singing synthesis systems [12] [13] [14] [15].

Another approach is to use the source/filter model. Several models of glottal pulse has been proposed with different quality and flexibility. A complete study and normalisation of the main models can be found in [16]. For example, the R++ model has been used in the famous Voicer [17]. LF [18] and CALM [19] models have been used during eNTERFACE'05 [3]. Other differences appear in the method used to compute the vocal tract transfer function. Some systems [20] compute the formants from the magnitude spectrum: a series of resonant filters (controlled by formants frequencies, amplitudes and bandwidths). Some other systems compute an acoustic representation of the vocal tract, as a cascade of acoustic (variant-shape) tubes. For example, the SPASM synthesizer [21] uses digital waveguides [22] to model acoustic features of oral, nasal cavities and throat radiation (driven by a frequency-domain excitation model). The model was extended to variable length conical sections by Välimäki and Karjalainen [23].

There exist also some particular approaches like FOF (*Formes d'Ondes Formantiques*) synthesis [24], used in CHANT [25], which performs synthesis by convolving a pulse train with parallel formant wave functions (time-domain functions corresponding to individual formants resonance).

IV. Voice Production

Voice organ is usually described as a "source/filter" system. Glottal source is a non-linear volume velocity generator where sound is produced by complex movements of vocal folds (larynx) under lungs pressure. A complete study of glottal source can be found in [26]. Sounds produced by the larynx are then propagated in oral and nasal cavities which can be seen as time-varying filtering. Finally, the flow is converted into radiated pressure waves through lips and nose openings (cf. Figure 1).

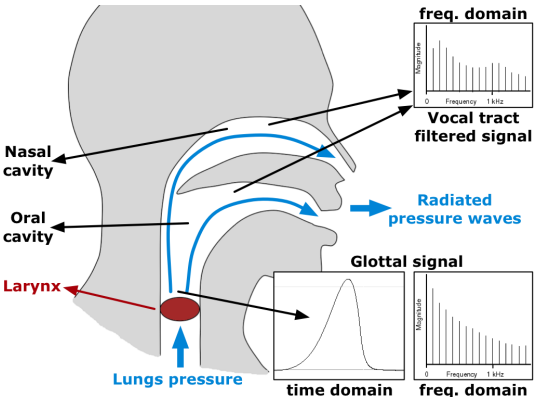


Fig. 1. Voice production mechanisms: vocal folds vibrations, vocal tract filtering and lips/nose openings radiation.

In the context of signal processing applications, and particulary in singing synthesis, some simplifications are usually accepted. First, lips and nose openings effect can be seen as derivative of the volume velocity signal. It is generally processed by a time-invariant high-pass first order linear filter [27]. Vocal tract effect can be modelized by filtering of glottal signal with multiple (usually 4 or 5) second order resonant linear filters.

Contrary to this "standard" vocal tract implementation, plenty of models have been developed for representation of glottal flow, with differences in accuracy and flexibility. Usual models are KLGLOTT88 [28], R++ [29], Rosenberg-C [30], LF [18] [31] and more recently, CALM [19].

V. The Glottal Source

In this section, we describe the work related to the realtime generation of the glottal source signal. We first explain our theoritical basics: the modelization of the glottal flow as the response of a causal/anticausal linear system (CALM). Then, we will describe two different implementations achieved during this workshop: a bufferized computation of a causal stable filter (v1.x) and a sample-by-sample computation of a causal unstable filter (v2.x).

A. The Causal/Anticausal Linear Model (CALM) [19]

Modelling vocal tract in spectral domain (with resonant filters central frequency, amplitude and bandwidth) is very powerful in term of manipulation because spectral description of sounds is close to auditory perception. Traditionally, glottal flow has been modelized in time domain. A spectral approach can be seen as equivalent only if both amplitude and phase spectra are considered in the model.

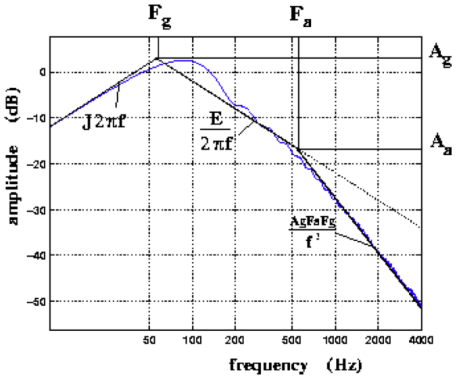


Fig. 2. Amplitude spectrum of the glottal flow derivative: illustration of glottal formant (F_g, A_g) and spectral tilt (F_a, A_a).

For amplitude spectrum, two different effects can be isolated (cf. Figure 2). On the one hand, an amount of energy is concentrated in low frequencies (i.e. below 3 kHz). This peak is usually called "glottal formant". We can see that bandwidth, amplitude and position of the glottal formant change with voice quality variations. On the other hand, a variation of spectrum slope in higher frequencies (called "spectral tilt") is also related to voice quality modifications.

Considering both "glottal formant" and "spectral tilt" effects, two cascading filters can be used. A second order resonant low-pass filter ($H_1(z)$) for glottal formant, and a first order

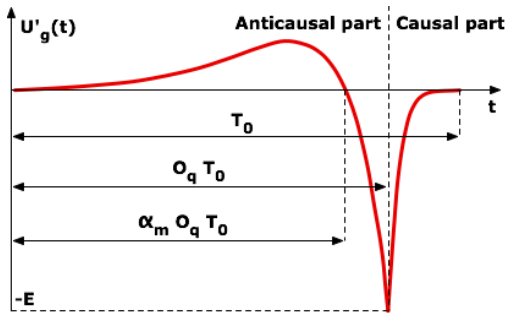


Fig. 3. Time-domain representation of derivated glottal pulse: anticausal part and causal part.

low-pass filter ($H_2(z)$) for spectral tilt. But phase information indicates us that this system is not completely causal. Indeed, as it is illustrated on Figure 3, glottal pulse is a combination of a "increasing" (or active) part and a "decreasing" (or passive) part. The decreasing part, called the return phase, mainly influences the spectral tilt and hence is causal. And we can also show that the second order low-pass filter has to be anticausal in order to provide a good phase representation. This information is sometimes refered as the mixed-phase representation of voice production [32].

A complete study of spectral features of glottal flow, detailed in [19], gives us equations linking relevant parameters of glottal pulse (F_0: fundamental frequency, O_q: open quotient, α_m: asymetry coefficient and T_l: spectral tilt, in dB at 3000Hz) to $H_1(z)$ and $H_2(z)$ coefficients. Expression of b_1 as been corrected, compared to [19] and [33].

Anticausal second order resonant filter:

$$H_1(z) = \frac{b_1 z}{1 + a_1 z + a_2 z^2}$$

$$a_1 = -2e^{-a_p T_e} cos(b_p T_e)$$

$$a_2 = e^{-2a_p T_e}, b_1 = E T_e$$

$$a_p = -\frac{\pi}{O_q T_0 tan(\pi \alpha_m)}, b_p = \frac{\pi}{O_q T_0}$$

Causal first order filter:

$$H_2(z) = \frac{b_{T_L}}{1 - a_{T_L} z^{-1}}$$

$$a_{T_L} = \nu - \sqrt{\nu^2 - 1}, b_{T_L} = 1 - a_{T_L}$$

$$\nu = 1 - \frac{1}{\eta}, \eta = \frac{\frac{1}{e^{-T_L/10ln(10)}} - 1}{cos(2\pi \frac{3000}{F_e}) - 1}$$

Full anticausal processing is only possible offline, by running algorithms backwards on data buffers. In a realtime context, anticausal response can be processed with two differents methods. On the one hand, the response of a causal version of $H_1(z)$ is stored backwards (*v1.x*). On the other hand, $H_1(z)$ is replaced by a unstable causal filter and the "divergent" impulse response is truncated (*v2.x*). We can also note that in order to be usefull our implementations have to be able to produce correct glottal flow (GF) and glottal flow derivative (GFD). Indeed, the GFD is the acoustical signal used to synthesize the voiced sounds, but the GF is important in the synthesis of turbulences, involved in unvoiced and breathy sounds.

B. RealtimeCALM v1.x Implementation

This implementation is the continuation of the development tasks of eNTERFACE'05 [3] and work presented to NIME'06 [33]. In this algorithm, we generate the impulse response by *period-synchronous anticausal processing*. It means that in order to achieve the requested waveform, the impulse response of a causal version of H_1 (glottal formant) is computed, but stored backwards in a buffer. This waveform is truncated at a length corresponding to instantaneous fundamental frequency (F_0+$Jitter$). This algorithm is now integrated in both Max/MSP [34] [35] and Pure Data [36] external objects (for Mac OS X, Windows and Linux): $almPulse_{\sim}$ $v1.x$. Then the resulting period is filtered by H_2 (spectral tilt). This algorithm is also integrated in both Max/MSP and Pure Data external objects: $stFilter_{\sim}$ $v1.x$. Coefficients of H_1 and H_2 are calculated from equations described in subsection *The Causal/Anticausal Linear Model (CALM)* and [19]. Thus, both time-domain and spectral-domain parameters can be sent.

Actually, we take advantage of physical properties of glottis to propose this real-time algorithm. Indeed, glottal pulse corresponds to opening/closing movements of vocal folds. It means that impulse responses generated by H_1 and H_2 filters can't overlap. Thus, impulse responses can be stored backwards and truncated period-synchronously without changing too much their spectral properties.

Truncation of the CALM waveform at each period gives quite good synthesis results. Nevertheless, several configurations of parameters (e.g. high value of α_m plus low value of O_q) make the impulse response oscilliating inside the period, which gives signals that are no more related to glottal source phenomena and changes voice quality perception. Thus, earlier truncation points and windowing options have been tested (e.g. first zero crossing of the GF, first zero crossing of the GFD). This study has shown us that it is not possible to set a truncation point inside the period which gives simultaneously correct synthesis results on the GF and the GFD (even with a synchronized half-Hanning windowing[1]). This modelization problem and limitations due to the use of period buffer drove us to change the architecture of this synthesis module (*v2.x*). Discontinuity in GFD due to GF truncation is illustrated at the Figure 4.

C. RealtimeCALM v2.x Implementation

This part explains another version of the anticausal filter response computation. It avoids the use of period buffer. Main

[1]This windowing method multiplies the increasing part of the glottal pulse (flow or derivative) – meaning the part between the zero crossing and the positive maximum – by the left part of a Hanning window.

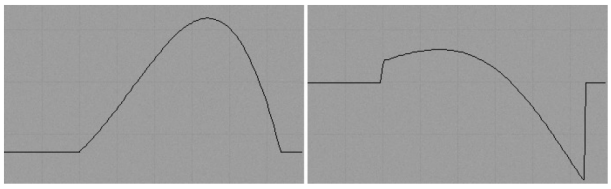


Fig. 4. Discontinuity in GFD (right) due to GF truncation at the first zero crossing of the CALM period (left).

idea behind this solution was to decrease memory allocations, in order to be able to generate simultaneously the glottal flow and the glottal flow derivative, with their own truncation points and windowings[2].

Instead of computing a causal version of the impulse response offline and then copying it backwards into a fixed buffer, the computation is here straightforward. The iterative equation corresponds indeed to the unstable anticausal filter. Anyway, the explosion of the filter is avoided by stopping the computation exactly at the Glottal Closure Instant (GCI). We can also note that glottal flow and glottal flow derivative can both be achieved with the same iterative equation, only changing the values of two first samples used as initial conditions in the iteration process.

One other main implementation problem is that the straightforward waveform generation has to be synchronized with the standard Pure Data's performing buffer size. This standard size is 64 samples which, at an audio rate of 44100Hz, corresponds to a frequency of approximately 690 Hz. Most of the time, the fundamental frequency of the glottal flow is less than 690 Hz, which means that several buffers are necessary to achieve the complete computation of one period. But whenever a buffer reaches the end, the main performing routine is called and thus the values of a_1 and a_2 have to be frozen as long as the period time has not been reached. A flag related to the opening of the glottis is then introduced, fixed to the value of the period (in samples), and the values of a_1 and a_2 are not changed until this flag is decreased to 0. Once values of T_0, T_e, γ, a_p, and b_p have been calculated at the opening instant, only a_1 and a_2 have to be frozen, as these are the only variables that are taken into account in the equations of the derivative glottal waveform.

We just tested the glottal flow/glottal flow derivative generation alone, without the addition of any vocal tract. However, strong tests have been carried out concerning this implementation and revealed that this version is more robust than the previous one. In particular, this implementation is not stuck when exotic values are send to the algorithm. Finally, we can note that this upgrade only concerns the $almPulse_{\sim}$ module. The spectral tilt filtering module ($stFilter_{\sim}$) was not modified.

[2]We can observe that our method will change the link between those two waveforms. Indeed, if two separated truncation points and windowings are applied, what we call "glottal flow derivative" is no more the derivative of the glottal flow.

D. Dimensionnal Issues

The next step in the realization of our singing tool was to define perceptual dimensions underlying the control of voice quality, and implement analytic mapping functions with low-level synthesis parameters. Dimensionnal features of voice were first collected from various research fields (signal processing, acoustics, phonetics, singing), completed, and described in a formalized set [33] [37].

- *Melody (F_0)*: short-term and long-term elements involved in the organization of temporal structure of fundamental frequency;
- *Vocal Effort (V)*: representation of the amount of "energy" involved in the creation of the vocal sound. It makes the difference between a spoken and a screamed voice [38] [39] [40] [41];
- *Tenseness (T)*: representation of the constriction of the voice source. It makes the difference between a lax and a tensed voice [26];
- *Breathiness (B)*: representation of the amount of air turbulences passing through the vocal tract, compared to the amount of voiced signal [26];
- *Hoarseness (H)*: representation of the stability of sound production parameters (especially for fundamental frequency and amplitude of the voice);
- *Mecanisms (M_i)*: voice quality modifications due to phonation type involved in the sound production [42].

E. Description of Mapping Functions

Once dimensions are defined, two main tasks can be investigated. First, the implementation of mapping functions between these dimensions and low-level parameters. Then, identification and implementation of inter-dimensionnal phenomena. In this area, many different theories have been proposed relating several intra- or inter-dimensionnal aspects of voice production [28] [41] [43] [44] [45] [46]. We decided to focus on some of them – like direct implementation of tenseness and vocal effort, realization of a phonetogram, etc. – and design our synthesis plateform in order to be able to extend it easily (e.g. correct existing relations, add new mapping functions, etc.). All current parameters are defined for a male voice.

Relations between Dimensions and Synthesis Parameters

During this workshop, we focused on several aspects of the dimensionnal process. First, we consider relations between a limited number of dimensions (F_0, V, T and M_i) and synthesis parameters (O_q, α_m and T_l). Then, we decided to achieve our data fusion scheme by considering two different "orthogonal" processes in the dimensionnal control. On the one hand, vocal effort (V) (also related to F_0 variations, cf. next paragraph: *Inter-Dimensionnal Relations*) and mecanisms (M_i) are controlling "offset" values of parameters (O_{q_0}, α_{m_0}, T_{l_0}). On the other hand, tenseness (T) controls "delta" values of O_q and α_m around their offsets (ΔO_q, $\Delta\alpha_m$). Considering this approach, effective values of synthesis parameters can be described as:

$$O_q = O_{q_0} + \Delta O_q$$

$$\alpha_m = \alpha_{m_0} + \Delta\alpha_m$$

$$T_l = T_{l_0}$$

Following equations consider V and T parameters normalized between 0 and 1 and M_i representing the i^{th} phonation mecanism.

- $O_{q_0} = f(V|M_i)$

$$O_{q_0} = 1 - 0,5 \times V|M_1$$

$$O_{q_0} = 0,8 - 0,4 \times V|M_2$$

- $\alpha_{m_0} = f(M_i)$

$$\alpha_{m_0} = 0,6|M_1$$

$$\alpha_{m_0} = 0,8|M_2$$

- $T_{l_0} = f(V)$

$$T_{l_0}(dB) = 55 - 49 \times V$$

- $\Delta O_q = f(T)$

$$\Delta O_q = (1 - 2T) \times O_q + 0,8T - 0,4|T \leq 0,5$$

$$\Delta O_q = (2T - 1) \times O_q + 2T + 1|T > 0,5$$

- $\Delta\alpha_m = f(T)$

$$\Delta\alpha_m = (0,5T - 1) \times \alpha_m - 1,2T + 0,6|T \geq 0,5$$

$$\Delta\alpha_m = (0,25 - 0,5T) \times \alpha_m + 0,4T - 0,2|T > 0,5$$

Last adaptation on parameters concerns a perceptual distorsion of O_q (square distorsion) and α_m (square root distorsion) between their ranges of variation (O_q: $0,4$ to 1; α_m: $0,6$ to $0,8$).

Inter-Dimensionnal Relations: the Phonetogram

One important characteristic of human voice production is that we are not able to produce any fundamental frequency (F_0) at any vocal effort (V). A strong relationship exists between these two perceptual features. For example, one could not produce a very low pitch (around $80Hz$) at a sound pressure level higher than $80dB$ (for a male speaker) or conversely to produce a high pitch at low intensity. Trying to do so results in a sudden stop of vocal production. This relationship is called phonetogram, and the evolution of this dependency is varying very much from one speaker to an another, considering for example that the subject is a trained singer or not, male or female, has pathological voice or not, etc. As a first approach, we decided to implement an average phonetogram, relying on the work of N. Henrich [46]. Figure 5 and Figure 6 represent two average phonetograms for male and female.

Moreover, this phenomenon involves different types of laryngeal configurations. We here dealt with mainly two configurations, first and second mechanisms of the vocal folds (M_1 and M_2). This two laryngeal mechanisms are, in the

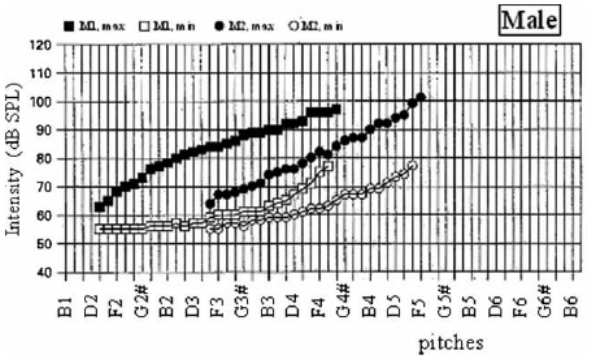


Fig. 5. Average voice range profile of male singers in mechanisms M_1 and M_2 [46].

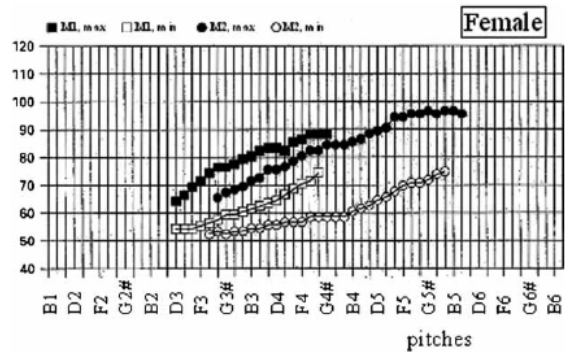


Fig. 6. Average voice range profile of female singers in mechanisms M_1 and M_2 [46].

commmon singing typology, referred as chest and falsetto registers. Hence, as shown on Figure 5 and Figure 6, it is also not possible to produce any frequency in both mechanisms, but the two configurations have an overlapping region in the middle of the phonetogram. This region enables the passing between the two mechanisms. Following the work presented in [47], the frequency range where this passing can occur is about one octave (or 12 semi-tones). The main characteristic of this passing is to provoke a break in the fundamental frequency (F_0). Thus, when producing an increasing glissando from M_1 to M_2, there is an average 8 semi-tones break, whereas it is approximately 12 semi-tones when performing a decreasing glissando. Breaking intervals probabilities are depicted on Figure 7 and Figure 8. On the first one we can actually note that the frequency breaks also depends on the fundamental frequency where it occurs.

So as to say that this phenomenon introduces an hysteresis. For most of untrained speakers or singers this break is uncontrollable whereas trained singers are able to hide more or less smoothly this break, although they cannot avoid mechanism switch.

VI. The Vocal Tract

In this section, we describe the implementation of a vocal tract model. This module is based on a physical "tubes-based" representation of vocal tract filter, which is simultaneously

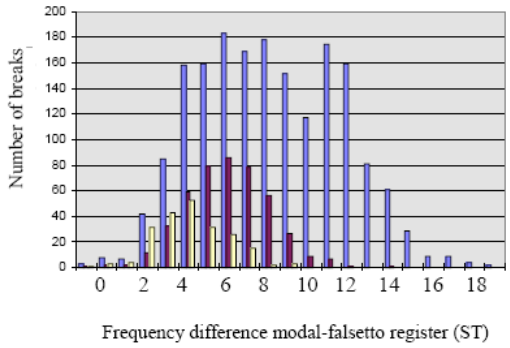


Fig. 7. Frequency drops densities in semi-tones from Chest(or Modal) to Falsetto register. In blue, when the break happens at $200Hz$, in red at $300Hz$, in yellow at $400Hz$ [47].

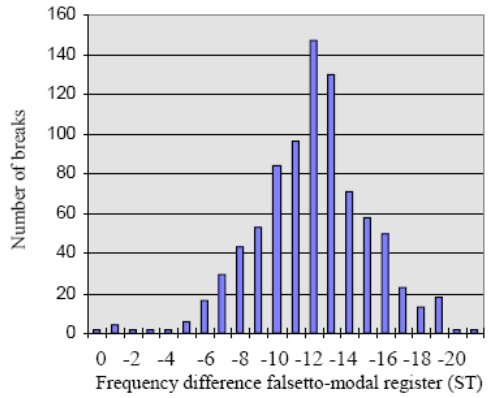


Fig. 8. Frequency drops densities in semi-tones from Falsetto to Chest(or Modal) register [47].

controllable with geometrical (areas) and spectral (formants) parameters.

A. The lattice filter

A geometrical approach of vocal tract representation

Linear Predictive Coding [48] is a method for representing the spectral envelope of a digital signal of speech in compressed form, using the information of a linear predictive model. The order of the filter is related to the complexity of the envelope, and also the number of control parameters. Thus, for representing a five-formant singing vowel, a filter containing five pairs of conjugated poles (for the resonances), and two simple poles (for the glottic wave) is needed, adding up to a total of fourteen parameters.

The LPC parameters (commonly named a_i) are non linearly interpolable. This implies that, for two configurations $[a_1 a_2 ... a_n]$ and $[b_1 b_2 ... b_n]$ corresponding to two vowels, a linear interpolation between both of these vectors will not correspond to a linear interpolation between the two spectra, and could even lead to unstable combinations. For these reasons, we will use another implementation of the LPC filter: the *lattice filter*. The control parameters of such a filter are called *reflection* coefficients (commonly named k_i). Such a filter is represented in Figure 9. It is composed of different sections, each characterized by a ki parameter.

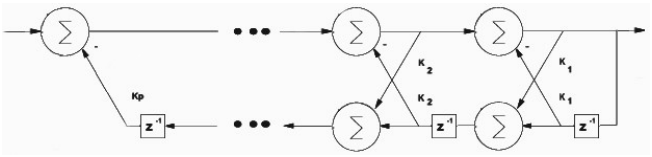


Fig. 9. Representation of k_p cells of a lattice filter.

The reflection coefficients correspond to physical characteristics of the vocal tract, which may be represented by a concatenation of cylindrical acoustic resonators, forming a lossless tube. This physical model of the lattice filter is represented in Figure 10. Each filter section represents one section of the tube; the forward wave entering the tube is partially reflected backwards, and the backward wave is partially reflected forwards. The reflection parameter k_i can then be interpreted as the ratio of acoustic reflections in the i^{th} cylindrical cavity, caused by the junction impedance with the adjacent cavity. This value varies from 1 (total reflection) to -1 (total reflection with phase inversion), and is equal to 0 when there is no reflection.

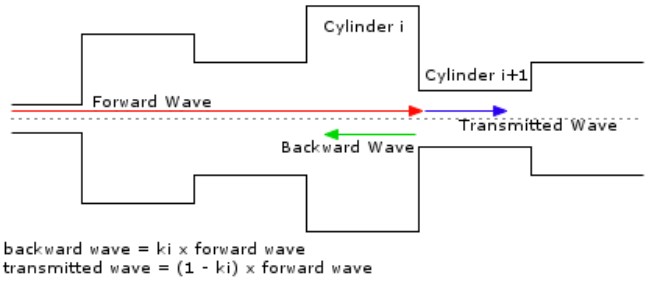


Fig. 10. Geometrical interpretation of the lattice filter: transmitted and backwards waves at each cell junction.

The filter will be stable if the k_i parameters are between -1 and 1. However, there is no direct relationship between these parameters and sound: a small modification of k_i does not imply a small modification of the spectrum. Thus, instead of using the reflection coefficients, we will be using the different cylinder areas, named A_i, which can be easily deducted from the reflection coefficients with the following expression:

$$\frac{A_i}{A_{i+1}} = \frac{1+k_i}{1-k_i}$$

By acting on these A_i parameters, the interpreter is directly connected to the physical synthesis instrument. The sound spectrum will then evolve with acoustical coherence, which makes it more natural to use. Moreover, the stability of the filter is guaranteed for all A_i values.

B. Coefficients Convertion Framework

In order to use the area parameters of the lattice filter (A_i), a Max/MSP object was created to convert them to k_i values which are used in the lattice filter. Several sets of A_i parameters corresponding to different vowels were calculated. After selecting one of these presets, certain sections of the

vocal tract can be modified by a percentage ΔA_i, which has the effect of opening or closing that section of the oral cavity.

A second approach to controlling the lattice filter was considered: a formant-based scheme was used to represent the spectral enveloppe, and the formant features, F_i, were converted to k_i parameters (after conversion to the LPC a_i coefficients), and then to A_i areas to control the lattice filter. This allowed us to easily model certain phenomena that are well known in speech processing, like overtone singing or the singer formant [49] [50], by acting on analytical parameters (the formants) rather than geometrical parameters (the areas). Similarly to the control of the areas, the formants have presets for different vowels and can be modified by a percentage ΔFi.

The parameters conversion framework described above is represented in Figure 11.

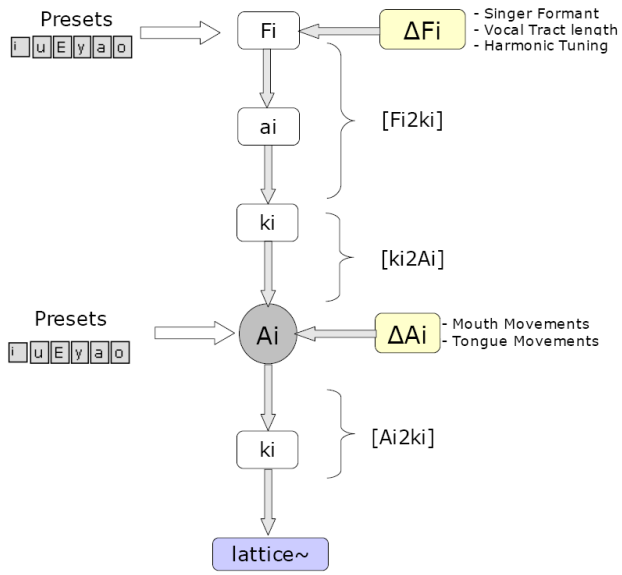


Fig. 11. Coefficients convertion and presets/modifications framework.

VII. About the Real-Time Control of Voice Synthesis

In this section, we comment some experimentations we realized in order to evaluate expressive and performing abilities of systems we developed. Modules were intergrated together inside Max/MSP and various control devices (and various combinaisons) were dynamically connected with mapping matrix. This set of tests allowed us to reach efficient configurations, considering several performing styles (classical singing, overtone singing, etc) which were demonstrated at the end of the workshop.

A. Concerning Voice Source

In order to be able to compare expressive skills of this system with the one developped before [3] [33] [37], we decided to keep the same control scheme: a graphic tablet. In that way, we were able to evaluate really clearly ameliorations achieved with this new mapping functions. Early experimentations demonstrated us that independant control of tenseness and vocal effort is really increasing performing possibilities. Anyway, current mapping equations still provide some unlikely parameters combinaisons, resulting e.g. in "ultra-tensed" perception or unwilling dynamics variations.

The implementation of the phonetogram is also a major improvement in term of naturalness. It also gives better results in terms of expressivity than without monitored control of loudness (more linear). Although we deeply investigated this phenomenon, we did not yet integrated this frequency break in the system, as we did not find a satisfying solution for controlling it. It is not straightforward to translate this frequency break in the control domain, as our hand gestures are mainly continuous and as basic switch from one configuration to another is not really satisfying from a musical point of view, as it results in a break in frequency range and thus "wrong" notes.

B. Concerning Vocal Tract

The vocal tract was controlled using a data glove (P5glove [51]) as shown in Figure 12. The glove was mapped to the area parameters of the lattice filter in four different ways:

- The folding of the fingers control the opening angle of the mouth (represented in Figure 13) (see Figure 14)
- The hand movement along the z-axis controls the position of the "tongue" in the vocal tract (towards the back or the front of the mouth)
- The hand movement along the y-axis controls the vertical position of the tongue (near or far from the palate) (see Figure 14)
- The hand movement along the x-axis changes the vowels (configurable from one preset to another, for example from an */a/* to an */o/*)

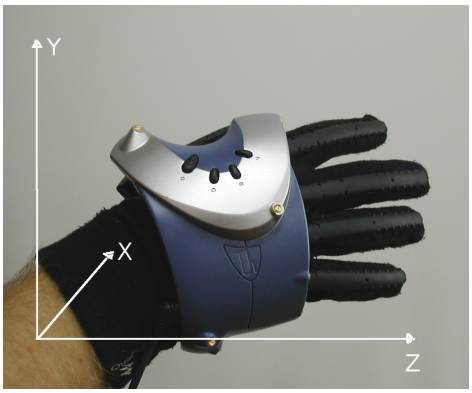


Fig. 12. Vocal tract control with a data glove: 5 finger flexion sensors and 3 dimensions (x,y,z) tracking.

This configuration allowed us to achieved typical vocal tract modification techniques – like overtone singing – quite easily. Indeed, as the spectral representation (F_i) is really efficient to configure some presets (e.g. offset vowel) or let running automatic tasks (e.g. harmonic/formant tuning), the constant access to geometrical "delta" features (ΔS_i) allows user to

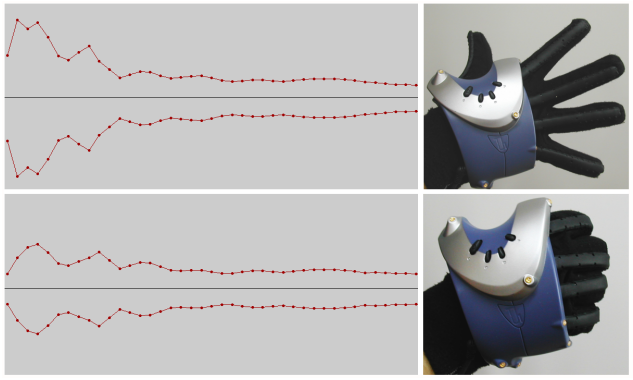


Fig. 13. Mouth opening control: finger flexion sensors mapped to variation of 9 first A_i.

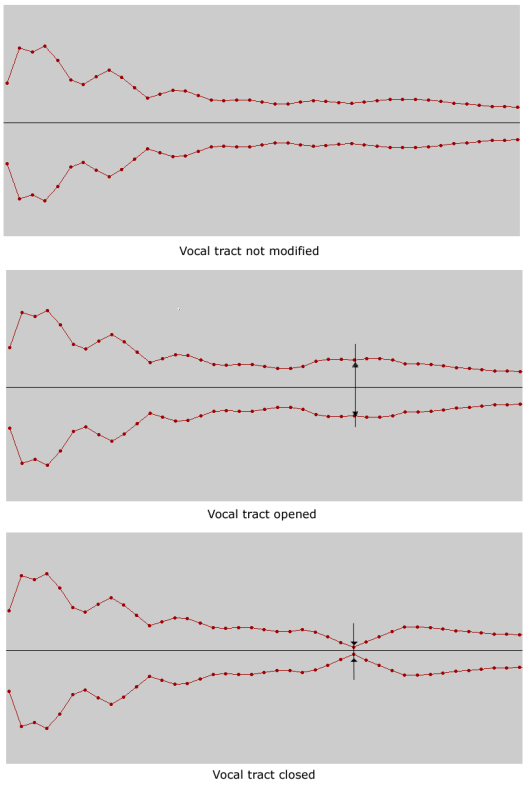


Fig. 14. Vertical tongue position control.

achieve refined tweaking techniques (e.g. lowering the vocal tract, changing tongue position, etc.) and that way increasing expressivity.

C. Transversal Remarks

In overall, at this stage of development, the synthesizer allows to control 17 parameters which are namely : pitch, vocal effort, tenseness, mechanisms, the first two formants, the singer's formant, vocal tract length, gain, transition between vowels, width of the vocal tract, position of the tongue and mouth opening (5 parameters). Considering all these parameters, only the actions on mechanisms is not a continuous parameter, so as to say that 16 parameters have to be monitored thanks to contiunuous parameters. From the controllers side, we have all in all 17 continuous parameters (out of 33), meaning that we are actually theoretically able to control all needed parameters. However, the problem is that from user's side, it is impossible to manipulate three interfaces at the same time. There are actually two solutions : one is to have multiple users (2 or 3) being in control of the interfaces, the other one is to use one-to-many mappings, allowing the performer to control several parameters with the same controller.

VIII. Conclusions

In this workshop, our main aim was to build a performant singing musical instrument allowing a wide range of expressive possibilities. Our actual work results in the implementation of new models for voice source and vocal tract, working in real-time, which are strategic tools in order to be able to work further. Improvement of expressivity in this new system really encourage us to go forward with this approach. Moreover, our modular architecture drives us to go to a widely extensible synthesis plateform which will be really usefull in order to continue to integrate other results (existing and coming) from voice production sciences.

Acknowledgment

The authors would like to thank SIMILAR Network of Excellence (and through it the European Union) which provides ressources allowing researchers from all Europe to meet, share and work together, and then achieving really exciting results. We also would like to thank croatian organization team (responsible: Prof. Igor Pandzic) which maintain local structures in order to manage the work and the life of more than 50 people. Finally, we would like to thank our respective laboratories (in our case: TCTS Lab, Mons, Belgium and LIMSI-CNRS, Paris, France) which adapt their research agendas in order to allow us to participate to those annual summer events.

REFERENCES

[1] "http://www.loquendo.com/."
[2] "http://www.enterface.net/enterface05/."
[3] C. d'Alessandro, N. D'Alessandro, S. L. Beux, J. Simko, F. Cetin, and H. Pirker, "The speech conductor: Gestural control of speech synthesis," in *Proceedings of eNTERFACE'05 Summer Workshop on Multimodal Interfaces*, 2005.
[4] M. Kob, "Singing voice modelling as we know it today," *Acta Acustica United with Acustica*, vol. 90, pp. 649–661, 2004.
[5] "http://www.virsyn.de/."
[6] "http://www.vocaloid.com/."
[7] X. Rodet and G. Bennet, "Synthesis of the singing voice," *Current Directories in Computer Music Research*, 1989.
[8] X. Rodet, "Synthesis and processing of the singing voice," in *Proceeding of the First IEEE Benelux Workshop on Model-Based Processing and Coding of Audio (MPCA-2002)*, Leuven, Belgium, 2002.
[9] P. Cook, "Identification of control parameters in an articulatory vocal tract model, with applications to the synthesis of singing," Ph.D. thesis, Standford University, 1990.
[10] J. Moorer, "The use of the phase vocoder in computer music application," *Journal of the Audio Engineering Society*, vol. 26, no. 1-2, pp. 42–45, 1978.
[11] J. Laroche, Y. Stylianou, and E. Moulines, "Hns: Speech modifications based on a harmonic plus noise model," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, vol. 2, 1993, pp. 550–553.
[12] M. Macon, L. Jensen-Link, J. Oliviero, M. Clements, and E. George, "A singing voice synthesis system based on sinusoidal modeling," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 1997, pp. 435–438.
[13] K. Lomax, "The analysis and the synthesis of the singing voice," Ph.D. thesis, Oxford University, 1997.
[14] Y. Meron, "High quality singing synthesis using the selection-based synthesis scheme," Ph.D. thesis, University of Michigan, 2001.
[15] P. Cano, A. Loscos, J. Bonada, M. de Boer, and X. Serra, "Voice morphing system for impersonating in karaoke applications," in *Proceedings of the International Computer Music Conference*, 2000.
[16] B. Doval, C. d'Alessandro, and N. Henrich, "The spectrum of glottal flow models," *Acta Acustica*, vol. In press, 2006.
[17] L. Kessous, "A two-handed controller with angular fundamental frequency control and sound color navigation," in *Proceedings of the 2002 Conference on New Interfaces for Musical Expression (NIME-02)*, 2002.
[18] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, pp. 1–13, 1985.
[19] B. Doval and C. d'Alessandro, "The voice source as a causal/anticausal linear filter," in *proc. Voqual'03, Voice Quality: Functions, analysis and synthesis, ISCA workshop*, Geneva, Switzerland, Aug. 2003.
[20] B. Larson, "Music and singing synthesis equipment (musse)," *Speech Transmission Laboratory Quarterly Progress and Statut Report (STL-QPSR)*, pp. (1/1977):38–40, 1977.
[21] P. Cook, "Spasm: a real-time vocal tract physical model editor/controller and singer: the companion software system," in *Colloque sur les Modèles Physiques dans l'Analyse, la Production et la Création Sonore*, 1990.
[22] J. O. Smith, "Waveguide filter tutorial," in *Proceedings of the International Computer Music Conference*, 1987, pp. 9–16.
[23] V. Välimäki and M. Karjalainen, "Improving the kelly-lochbaum vocal tract model using conical tubes sections and fractionnal delay filtering techniques," in *Proceedings of the International Conference on Spoken Language Processing*, 1994.
[24] X. Rodet, "Time-domain formant wave function synthesis," vol. 8, no. 3, pp. 9–14, 1984.
[25] X. Rodet and J. Barriere, "The chant project: From the synthesis of the singing voice to synthesis in general," *Computer Music Journal*, vol. 8, no. 3, pp. 15–31, 1984.
[26] N. Henrich, "Etude de la source glottique en voix parlée et chantée," Ph.D. thesis, Université Paris 6, France, 2001.
[27] G. Fant, *Acoustic theory of speech production*. Mouton, La Hague, 1960.
[28] D. Klatt and L. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acous. Soc. Am.*, vol. 87, no. 2, pp. 820–857, 1990.
[29] R. Veldhuis, "A computationally efficient alternative for the liljencrants-fant model and its perceptual evaluation," *J. Acous. Soc. Am.*, vol. 103, pp. 566–571, 1998.
[30] A. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acous. Soc. Am.*, vol. 49, pp. 583–590, 1971.
[31] G. Fant, "The lf-model revisited. transformations and frequency domain analysis," *STL–QPSR*, 1995.
[32] B. Bozkurt, "Zeros of the z-transform (zzt) representation and chirp group delay processing for the analysis of source and filter characteristics of speech signals," Ph.D. dissertation, Faculté Polytechnique de Mons, 2004.
[33] N. D'Alessandro, C. d'Alessandro, S. Le Beux, and B. Doval, "Realtime calm synthesizer, new approaches in hands-controlled voice synthesis," in *NIME'06, 6th international conference on New Interfaces for Musical Expression*, IRCAM, Paris, France, 2006, pp. 266–271.
[34] D. Zicarelli, G. Taylor, J. Clayton, jhno, and R. Dudas, *Max 4.3 Reference Manual*. Cycling'74 / Ircam, 1993-2004.
[35] ——, *MSP 4.3 Reference Manual*. Cycling'74 / Ircam, 1997-2004.
[36] M. Puckette, *Pd Documentation*. http://puredata.info, 2006.
[37] C. d'Alessandro, N. D'Alessandro, S. L. Beux, and B. Doval, "Comparing time-domain and spectral-domain voice source models for gesture controlled vocal instruments," in *Proc. of the 5th International Conference on Voice Physiology and Biomechanics*, 2006.
[38] R. Schulman, "Articulatory dynamics of loud and normal speech," *J. Acous. Soc. Am.*, vol. 85, no. 1, pp. 295–312, 1989.
[39] H. M. Hanson, "Glottal characteristics of female speakers," Ph.D. thesis, Harvard University, 1995.
[40] ——, "Glottal characteristics of female speakers : Acoustic correlates," *J. Acous. Soc. Am.*, vol. 101, pp. 466–481, 1997.
[41] H. M. Hanson and E. S. Chuang, "Glottal characteristics of male speakers : Acoustic correlates and comparison with female data," *J. Acous. Soc. Am.*, vol. 106, no. 2, pp. 1064–1077, 1999.
[42] M. Castellengo, B. Roubeau, and C. Valette, "Study of the acoustical phenomena characteristic of the transition between chest voice and falsetto," in *Proc. SMAC 83, vol. 1*, Stockholm, Sweden, July 1983, pp. 113–23.
[43] P. Alku and E. Vilkman, "A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers," *Folia Phoniatr.*, vol. 48, pp. 240–54, 1996.
[44] H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," *J. Acous. Soc. Am.*, vol. 107, no. 6, pp. 3438–51, 2000.
[45] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo, "On the use of the derivative of electroglottographic signals for characterization of non-pathological phonation," *J. Acous. Soc. Am.*, vol. 115, no. 3, pp. 1321–1332, Mar. 2004.
[46] N. Henrich, C. d'Alessandro, M. Castellengo, and B. Doval, "Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency," *J. Acous. Soc. Am.*, vol. 117, no. 3, pp. 1417–1430, Mar. 2005.
[47] G. Bloothooft, M. van Wijck, and P. Pabon, "Relations between vocal registers in voice breaks," in *Proccedings of Eurospeech*, 2001.
[48] J. D. Markel and A. H. Gray, *Linear prediction of speech*. Springer-Verlag, Berlin, 1976.
[49] B. STORY, "Physical modeling of voice and voice quality," in *proc. Voqual'03, Voice Quality: Functions, analysis and synthesis, ISCA workshop*, Geneva, Switzerland, Aug. 2003.
[50] G. Carlsson and J. Sundberg, "Formant frequency tuning in singing," *J. Voice*, vol. 6, no. 3, pp. 256–60, 1992.
[51] "http://www.vrealities.com/p5.html."

Nicolas D'Alessandro holds an Electrical Engineering degree from the Faculté Polytechnique de Mons (FPMs) since 2004. I did the master's thesis in the Faculty of Music of the University de Montréal (UdeM) (supervisor: Prof. Caroline Traube). That work gathered the development of applications based on perceptual analogies between guitar sounds and voice sounds, and a study of mapping strategies between (hand-based) gestures and speech production models (source/filters and concatenative approaches). He started a PhD thesis in September 2004 in the TCTS Lab of the FPMs (supervisor: Prof. Thierry Dutoit) related to the real-time control of unit-based synthesizers.

Boris Doval holds an Engineering degree from the "Ecole Centrale de Paris" since 1987. He did his master thesis at IRCAM and his PhD thesis at "Université Paris VI" on fundamental frequency estimation of sound signals. He then joined LIMSI-CNRS as an associate professor where he concentrated until now on voice source analysis, synthesis and modelisation. In particular, he coorganized the ISCA workshop VOQUAL'03 on voice quality in 2003.

Sylvain Le Beux graduated from Master in Electronics, Telecommunications and Informatics from CPE Lyon engineer school in 2004. During his training he did an internship at Infineon Technology A.G. in Munich for one year, and achieved his master's thesis at IRCAM which topic was about speech recognition. So he actually helped IRCAM wreck on a nice beach using calm insense. He then graduated from a Master thesis on embedded systems and data processing from Orsay University in 2005, and then integrated LIMSI Laboratory where he is currently achieving his PhD focused on the gestural control of speech synthesis and relationship between intention and expressivity.

Pascale Woodruff holds an Electrical Engineering degree from FPMs since june 2004. She is currently working on a project which aims to improve the workflow in industrial maintenance by equiping technicians with a mutimodal wearable system allowing them to access maintenance information using speech and/or other modalities.

Yohann Fabre is ending a master thesis in audiovisual technologies at the ISIS (Ingénierie des Systèmes, Image et Son) of Valenciennes. He is currently working as a trainee on voice synthesis at the TCTS Lab of FPMs.

Journal on Multimodal User Interfaces, Vol. 1, No. 1, March 2007

REALTIME AND ACCURATE MUSICAL CONTROL OF EXPRESSION IN SINGING SYNTHESIS

Nicolas D'Alessandro, Pascale Woodruff, Yohann Fabre, Thierry Dutoit

TCTS Lab, Faculté Polytechnique de Mons, Belgium
{nicolas.dalessandro;pascale.woodruff;yohann.fabre;thierry.dutoit}@fpms.ac.be

Sylvain Le Beux, Boris Doval, Christophe d'Alessandro

LIMSI-CNRS, Université Paris XI, Orsay, France
{sylvain.le.beux;boris.doval;christophe.dalessandro}@limsi.fr

ABSTRACT

In this paper, we describe a full computer-based musical instrument allowing realtime synthesis of expressive singing voice. The expression results from the continuous action of an interpreter through a gestural control interface. In this context, expressive features of voice are discussed. New real-time implementations of a spectral model of glottal flow (CALM) are described. These interactive modules are then used to identify and quantify voice quality dimensions. Experiments are conducted in order to develop a first framework for voice quality control. The representation of vocal tract and the control of several vocal tract movements are explained and a solution is proposed and integrated. Finally, some typical controllers are connected to the system and expressivity is evaluated.

KEYWORDS

Singing voice – Voice synthesis – Voice quality – Glottal flow models – Gestural control – Interfaces.

1. INTRODUCTION

Expressivity is nowadays one of the most challenging topics studied by researchers in speech synthesis. Indeed, recent synthesisers provide acceptable speech in term of intelligibility and naturalness but the need to improve human/computer interactions has brought researchers to develop more human, expressive systems. Some recent realisations have shown that an interesting option was to record multiple databases corresponding to a certain number of labelled expressions (e.g. happy, sad, angry, etc.) [1]. At synthesis time, the expression of the virtual speaker is then set by choosing the units in the corresponding database.

We decided to investigate the opposite option. Indeed, we postulated that emotion in speech was not the result of switches between labelled expressions but a continuous evolution of voice characteristics highly correlated with the context. Thus, we developed a set of flexible voice synthesisers conducted in realtime by an operator [2]. After some tests, it was clear that such a framework was particularly efficient for singing synthesis.

Remarkable achievements have been recently reached in singing voice synthesis. A review of state of the art can be found in [3]. The technology seems mature enough now to allow for the replacement of human vocals with synthetic singing, at least for backing vocals [4] [5]. However, existing singing synthesis systems suffer from two restrictions: they are aiming at mimicking singers rather than creating new instruments, and are generally limited to MIDI controllers.

We think it worthwhile to extend vocal possibilities of voice synthesisers and design new interfaces that will open new musical possibilities. In a first attempt we decided to restrain our survey on voice quality control to the boundaries of natural voice production - in fact, it is always better trying to mimic one particular voice. This process enables us to achieve analysis by synthesis : once we are able to perceive more naturalness in the synthesised voice, then we understood something about the voice production process. It is then easier to diverge from these limits when dealing with a musical application in a more creative way.

2. AIMS OF THIS WORK

Our aims can be summarised in three main axes. First, we target the implementation of intra and inter-dimensional mappings driving low-level parameters of source models (e.g. complex interactions between vocal effort and tenseness, represented by the phonetogram). Then, we investigate the effects of the vocal tract in voice quality variations (e.g. the singer formant, lowering of the larynx). Finally, source/filter coupling effects (e.g. relations between harmonics and formant frequencies) are analysed, and several mechanisms are implemented (e.g. overtone, bulgarian, occidental singing).

3. BACKGROUND IN SINGING SYNTHESIS

Speech and singing both result from the same production system: the vocal apparatus. However, the signal processing techniques developed for their synthesis evolved quite differently. One of the main reasons for this deviation is that the aim for producing voice is different in the two cases. The aim of speech production is to exchange messages. For singing, the main aim is to use the voice organ as a musical instrument. Therefore a singing synthesis system needs to include various tools to control (analyse/synthesise or modify) different dynamics of the acoustic sound produced: duration of the phonemes, vibrato, wide range modifications of the voice quality, the pitch and the intensity, etc., some of which are not needed in most of the speech synthesis systems. A pragmatic reason for that separation is that singing voice synthesisers target almost exclusively musical performances. In this case, playability (flexibility and real-time abilities) is much more important than intelligibility. Discussions about various issues of singing synthesis can be found in [6, 7].

As described in [8], frequency-domain analysis/modification methods are frequently preferred in singing synthesis research due to the need to modify some spectral characteristics of actual

recorded signals. The most popular application of such a technique is the phase vocoder [9], which is a powerful tool used for many years for time compression/expansion, pitch shifting and cross-synthesis.

To increase flexibility, short-time signal frames can be modelled as sums of sinusoids (controlled in frequency, amplitude and phase) plus noise (controlled by the parameters of a filter which is excited by a white noise). HNM (Harmonic plus Noise Model) [10] provides a flexible representation of the signal, which is particularly interesting in the context of unit concatenation. That representation of signals is thus used as a basis in many singing synthesis systems [11, 12, 13, 14].

Another approach is to use the source/filter model. Several models of glottal pulse has been proposed with different quality and flexibility. A complete study and normalisation of the main models can be found in [15]. For example, the R++ model has been used in the famous Voicer [16]. LF [17] and CALM [18] models have been used during eNTERFACE workshops [2]. Other differences appear in the method used to compute the vocal tract transfer function. Some systems [19] compute the formants from the magnitude spectrum: a series of resonant filters (controlled by formants frequencies, amplitudes and bandwidths). Some other systems compute an acoustic representation of the vocal tract, as a cascade of acoustic (variant-shape) tubes. For example, the SPASM synthesiser [20] uses digital waveguides [21] to model acoustic features of oral, nasal cavities and throat radiation (driven by a frequency-domain excitation model). The model was extended to variable length conical sections by Välimäki and Karjalainen [22].

There exist also some particular approaches like FOF (*Formes d'Ondes Formantiques*) synthesis [23], used in CHANT [24], which performs synthesis by convolving a pulse train with parallel formant wave functions (time-domain functions corresponding to individual formants resonance).

4. VOICE PRODUCTION

The vocal apparatus is usually described as a "source/filter" system. Glottal source is a non-linear volume velocity generator where sound is produced by complex movements of vocal folds (larynx) under lung pressure. A complete study of glottal source can be found in [25]. Sounds produced by the larynx are then propagated in oral and nasal cavities which can be seen as time-varying filters. Finally, the flow is converted into radiated pressure waves through the lips and nose openings (cf. Figure 1).

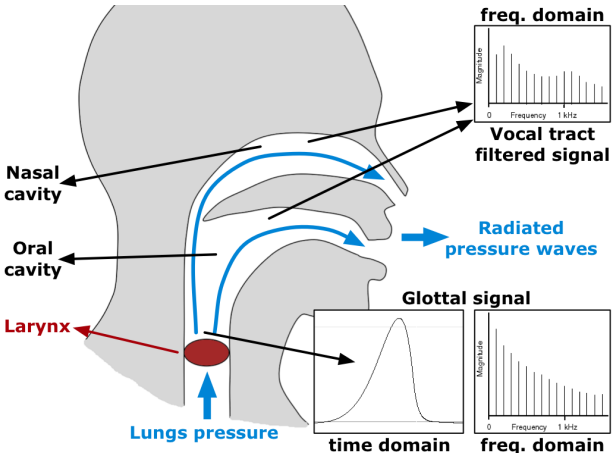


Figure 1: Voice production mechanisms: vocal folds vibrations, vocal tract filtering and lips/nose openings radiation.

In the context of signal processing applications, and particularly in singing synthesis, some simplifications are usually accepted. Firstly, the effect of lip and nose openings can be seen as derivative of the volume velocity signal. It is generally processed by a time-invariant high-pass first order linear filter [26]. Vocal tract effect can be modelled by filtering the glottal signal with multiple (usually 4 or 5) second order resonant linear filters.

Contrary to this "standard" vocal tract implementation, plenty of models have been developed for representation of glottal flow, with differences in accuracy and flexibility. Usual models are KLGLOTT88 [27], R++ [28], Rosenberg-C [29], LF [17, 30] and more recently, CALM [18].

5. THE GLOTTAL SOURCE

In this section, we describe the work related to the realtime generation of the glottal source signal. We first explain our theoretical basics: the modelling of the glottal flow as the response of a causal/anticausal linear system (CALM). Then, we will describe two different implementations achieved: a buffered computation of a causal stable filter (v1.x) and a sample-by-sample computation of a causal unstable filter (v2.x).

5.1. The Causal/Anticausal Linear Model (CALM) [18]

Modelling the human vocal tract in the spectral domain (with resonant filters central frequency, amplitude and bandwidth) is very powerful in term of manipulation because spectral description of sounds is close to auditory perception. Traditionally, glottal flow has been modelled in the time domain. A spectral approach can be seen as equivalent only if both amplitude and phase spectra are considered in the model.

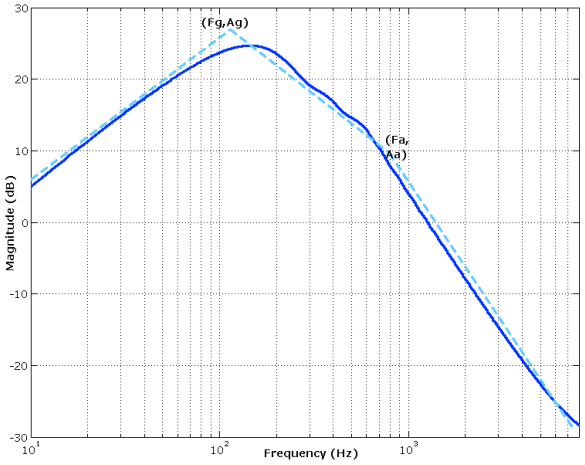


Figure 2: Amplitude spectrum of the glottal flow derivative: illustration of glottal formant (F_g, A_g) and spectral tilt (F_a, A_a).

For amplitude spectrum, two different effects can be isolated (cf. Figure 2). On the one hand, an amount of energy is concentrated in low frequencies (i.e. below 3 kHz). This peak is usually called the "glottal formant". We can see that bandwidth, amplitude and position of the glottal formant can change with voice quality variations. On the other hand, a variation of spectrum slope in higher frequencies (called "spectral tilt") is also related to voice quality modifications.

Considering both "glottal formant" and "spectral tilt" effects, two cascading filters can be used. A second order resonant low-pass filter ($H_1(z)$) for glottal formant, and a first order

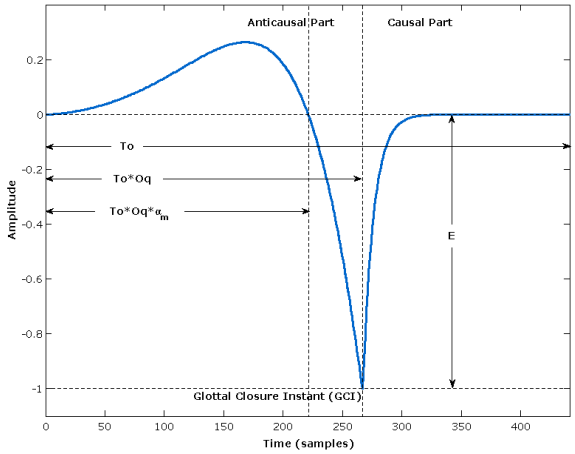


Figure 3: Time-domain representation of derivated glottal pulse: anticausal and causal parts, respectively on the left and right of the glottal closure instant.

low-pass filter ($H_2(z)$) for spectral tilt. But phase information indicates to us that this system is not completely causal. Indeed, as it is illustrated on Figure 3, glottal pulse is a combination of a "increasing" (or active) part and a "decreasing" (or passive) part. The decreasing part, called the return phase, mainly influences the spectral tilt and is causal. And we can also show that the second order low-pass filter has to be anticausal in order to provide a good phase representation. This information is sometimes referred as the mixed-phase representation of voice production [31].

A complete study of spectral features of glottal flow, detailed in [18], gives us equations linking relevant parameters of glottal pulse (F_0: fundamental frequency, O_q: open quotient, α_m: asymetry coefficient and T_l: spectral tilt, in dB at 3000Hz) to $H_1(z)$ and $H_2(z)$ coefficients. Expression of b_1 as been corrected, compared to [18] and [32].

Anticausal second order resonant filter:

$$H_1(z) = \frac{b_1 z}{1 + a_1 z + a_2 z^2}$$
$$a_1 = -2e^{-a_p T_e} cos(b_p T_e)$$
$$a_2 = e^{-2a_p T_e}, b_1 = ET_e$$
$$a_p = -\frac{\pi}{O_q T_0 tan(\pi \alpha_m)}, b_p = \frac{\pi}{O_q T_0}$$

Causal first order filter:

$$H_2(z) = \frac{b_{T_L}}{1 - a_{T_L} z^{-1}}$$
$$a_{T_L} = \nu - \sqrt{\nu^2 - 1}, b_{T_L} = 1 - a_{T_L}$$
$$\nu = 1 - \frac{1}{\eta}, \eta = \frac{\frac{1}{e^{-T_L/10ln(10)}} - 1}{cos(2\pi \frac{3000}{F_e}) - 1}$$

Full anticausal processing is only possible offline, by running algorithms backwards on data buffers. In a realtime context, anticausal response can be processed with two different methods. On the one hand, the response of a causal version of $H_1(z)$ is stored backwards (*v1.x*). On the other hand, $H_1(z)$ is replaced by a unstable causal filter and the "divergent" impulse response is truncated (*v2.x*). We can also note that in order to be useful our implementations have to be able to produce correct glottal flow (GF) and glottal flow derivative (GFD). Indeed, the GFD is the acoustical signal used to synthesise the voiced sounds, but the GF is important in the synthesis of turbulence, involved in unvoiced and breathy sounds.

5.2. RealtimeCALM v1.x Implementation

This implementation is the continuation of the development tasks of eNTERFACE'05 [2] and work presented to NIME'06 [32]. In this algorithm, we generate the impulse response by *period-synchronous anticausal processing*. It means that in order to achieve the requested waveform, the impulse response of a causal version of H_1 (glottal formant) is computed, but stored backwards in a buffer. This waveform is truncated at a length corresponding to instantaneous fundamental frequency (F_0+$Jitter$). This algorithm is now integrated in both Max/MSP [33, 34] and Pure Data [35] external objects (for Mac OS X, Windows and Linux): $almPulse_{\sim}$ $v1.x$. Then the resulting period is filtered by H_2 (spectral tilt). This algorithm is also integrated in both Max/MSP and Pure Data external objects: $stFilter_{\sim}$ $v1.x$. Coefficients of H_1 and H_2 are calculated from equations described in subsection *The Causal/Anticausal Linear Model (CALM)* and [18]. Thus, both time-domain and spectral-domain parameters can be sent.

Actually, we take advantage of physical properties of glottis to propose this real-time algorithm. Indeed, glottal pulse corresponds to opening/closing movements of vocal folds. It means that impulse responses generated by H_1 and H_2 filters can't overlap. Thus, impulse responses can be stored backwards and truncated period-synchronously without excessively changing their spectral properties.

Truncation of the CALM waveform at each period gives quite good synthesis results. Nevertheless, several configurations of parameters (e.g. high value of α_m plus low value of O_q) make the impulse response oscillating inside the period, which gives signals that are no more related to glottal source phenomena and changes voice quality perception. Thus, earlier truncation points and windowing options have been tested (e.g. first zero crossing of the GF, first zero crossing of the GFD). This study has shown us that it is not possible to set a truncation point inside the period which gives simultaneously correct synthesis results on the GF and the GFD (even with a synchronized half-Hanning windowing[1]). This modelization problem and limitations due to the use of period buffer drove us to change the architecture of this synthesis module (*v2.x*). Discontinuity in GFD due to GF truncation is illustrated at the Figure 4.

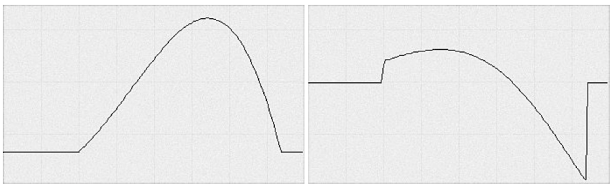


Figure 4: Discontinuity in GFD (right) due to GF truncation at the first zero crossing of the CALM period (left).

5.3. RealtimeCALM v2.x Implementation

This part explains another version of the anticausal filter response computation. It avoids the use of period buffer. The main idea behind this solution was to decrease memory allocations, in

[1] This windowing method multiplies the increasing part of the glottal pulse (flow or derivative) – meaning the part between the zero crossing and the positive maximum – by the left part of a Hanning window.

order to be able to generate simultaneously the glottal flow and the glottal flow derivative, each with their own truncation points and windowings[2].

Instead of computing a causal version of the impulse response off-line and then copying it backwards into a fixed buffer, the computation is here straightforward. The iterative equation corresponds indeed to the unstable anticausal filter. At any rate, the explosion of the filter is avoided by stopping the computation exactly at the Glottal Closure Instant (GCI). We can also note that glottal flow and glottal flow derivative can both be achieved with the same iterative equation, only changing the values of two first samples used as initial conditions in the iteration process.

One other main implementation problem is that the straightforward waveform generation has to be synchronised with the standard Pure Data performing buffer size. This standard size is 64 samples which, at an audio rate of 44100Hz, corresponds to a frequency of approximately 690 Hz. Most of the time, the fundamental frequency of the glottal flow is less than 690 Hz, which means that several buffers are necessary to achieve the complete computation of one period. But whenever a buffer reaches the end, the main performing routine is called and thus the values of a_1 and a_2 have to be frozen as long as the period time has not been reached. A flag related to the opening of the glottis is then introduced, fixed to the value of the period (in samples), and the values of a_1 and a_2 are not changed until this flag is decreased to 0. Once values of T_0, T_e, γ, a_p, and b_p have been calculated at the opening instant, only a_1 and a_2 have to be frozen, as these are the only variables that are taken into account in the equations of the derivative glottal waveform.

We just tested the glottal flow/glottal flow derivative generation alone, without the addition of any vocal tract information. However, extensive tests have been carried out concerning this implementation and revealed that this version is more robust than the previous one. In particular, this implementation is not stuck when exotic values are sent to the algorithm. Finally, we can note that this upgrade only concerns the $almPulse_{\sim}$ module. The spectral tilt filtering module ($stFilter_{\sim}$) was not modified.

5.4. Dimensionnal Issues

The next step in the realisation of our singing tool was to define perceptual dimensions underlying the control of voice quality, and to implement analytic mapping functions with low-level synthesis parameters. Dimensional features of voice were first collected from various research fields (signal processing, acoustics, phonetics, singing), completed, and described in a formalised set [32, 36].

- *Melody (F_0)*: short-term and long-term elements involved in the organisation of temporal structure of fundamental frequency;
- *Vocal Effort (V)*: a representation of the amount of "energy" involved in the creation of the vocal sound. It makes the clear difference between a spoken and a screamed voice for example [37, 38, 39, 40];
- *Tenseness (T)*: a representation of the constriction of the voice source. It makes the difference between a lax and a tensed voice [25];
- *Breathiness (B)*: a representation of the amount of air turbulence passing through the vocal tract, compared to the amount of voiced signal [25, 27];
- *Hoarseness (H)*: a representation of the stability of sound production parameters (especially for fundamental frequency and amplitude of the voice);
- *Mecanisms (M_i)*: voice quality modifications due to type of phonation involved in sound production [41].

[2] We can observe that our method will change the link between those two waveforms. Indeed, if two separated truncation points and windowings are applied, what we call "glottal flow derivative" is no more the derivative of the glottal flow.

5.5. Description of Mapping Functions

Once dimensions are defined, two main tasks can be investigated. First, the implementation of mapping functions between these dimensions and low-level parameters. Then, identification and implementation of inter-dimensional phenomena. In this area, many different theories have been proposed relating to several intra or inter-dimensional aspects of voice production [27, 40, 42, 43, 44, 45]. We decided to focus on some of them, like direct implementation of tenseness and vocal effort, realisation of a phonetogram, etc. and design our synthesis platform in order to be easily extensible (e.g. to correct existing relations and add new mapping functions etc.). All current parameters are defined for a male voice.

Relations between Dimensions and Synthesis Parameters

We focused on several aspects of the dimensionnal process. First, we consider relations between a limited number of dimensions (F_0, V, T and M_i) and synthesis parameters (O_q, α_m and T_l). Then, we decided to achieve our data fusion scheme by considering two different "orthogonal" processes in the dimensionnal control. On the one hand, vocal effort (V) (also related to F_0 variations, cf. next paragraph: *Inter-Dimensionnal Relations*) and mecanisms (M_i) are controlling "offset" values of parameters (O_{q_0}, α_{m_0}, T_{l_0}). On the other hand, tenseness (T) controls "delta" values of O_q and α_m around their offsets (ΔO_q, $\Delta\alpha_m$). Considering this approach, effective values of synthesis parameters can be described as:

$$O_q = O_{q_0} + \Delta O_q$$
$$\alpha_m = \alpha_{m_0} + \Delta\alpha_m$$
$$T_l = T_{l_0}$$

Following equations consider V and T parameters normalized between 0 and 1 and M_i representing the i^{th} phonation mecanism.

- $O_{q_0} = f(V|M_i)$

$$O_{q_0} = 0,8 - 0,4 \times V|M_1$$
$$O_{q_0} = 1 - 0,5 \times V|M_2$$

- $\alpha_{m_0} = f(M_i)$

$$\alpha_{m_0} = 0,8|M_1$$
$$\alpha_{m_0} = 0,6|M_2$$

- $T_{l_0} = f(V)$

$$T_{l_0}(dB) = 55 - 49 \times V$$

- $\Delta O_q = f(T)$

$$\Delta O_q = (1 - 2T)O_{q_0} + 0,8T - 0,4|T \leq 0,5$$
$$\Delta O_q = (2T - 1)O_{q_0} + 2T + 1|T > 0,5$$

- $\Delta\alpha_m = f(T)$

$$\Delta\alpha_m = (0,5T - 1)\alpha_{m_0} - 1,2T + 0,6 | T \geq 0,5$$

$$\Delta\alpha_m = (0,25 - 0,5T)\alpha_{m_0} + 0,4T - 0,2 | T < 0,5$$

Last adaptation on parameters concerns a perceptual distorsion of O_q (square distorsion) and α_m (square root distorsion) between their ranges of variation (O_q: $0,4$ to 1; α_m: $0,6$ to $0,8$) [46].

Inter-Dimensionnal Relations: the Phonetogram

One important characteristic of human voice production is that we are not able to produce any fundamental frequency (F_0) at any vocal effort (V). A strong relationship exists between these two production features. For example, one could not produce a very low pitch (around $80Hz$) at a sound pressure level higher than $80dB$ (for a male speaker) or conversely to produce a high pitch at low intensity. Trying to do so results in a sudden stop of vocal production. This relationship is called the phonetogram, and the evolution of this dependency varies very much from one speaker to an another. Consider, for example, whether the subject is a trained singer or not, male or female, has a pathological voice or not, etc. As a first approach, we decided to implement an average phonetogram, relying on the work of N. Henrich [47]. Figure 5 and Figure 6 represent two average phonetograms for male and female.

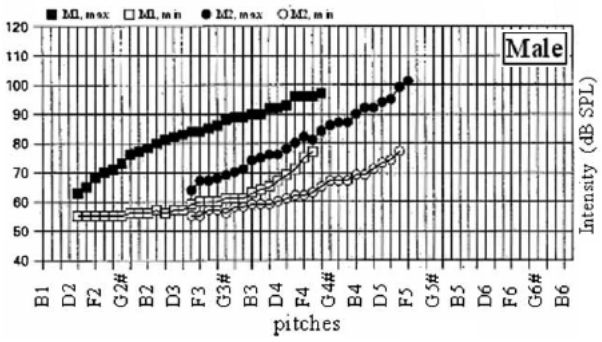


Figure 5: Average voice amplitude range profile (phonetogram) of male singers in mechanisms M_1 and M_2 [47].

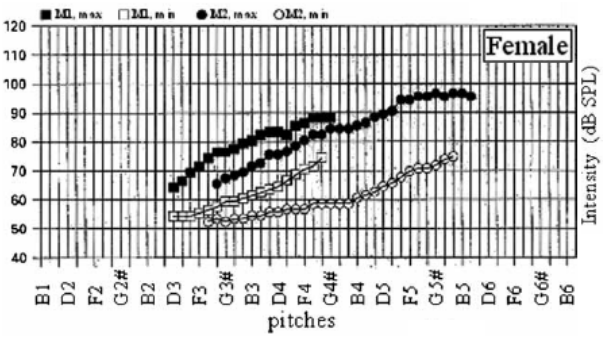


Figure 6: Average voice amplitude range profile (phonetogram) of female singers in mechanisms M_1 and M_2 [47].

Moreover, this phenomenon involves different types of laryngeal configurations. We here dealt with mainly two configurations, first and second mechanisms of the vocal folds (M_1 and M_2). This two laryngeal mechanisms are, in the common singing typology, referred as chest and falsetto registers. Hence, as shown on Figure 5 and Figure 6, it is not possible to produce any frequency in both mechanisms, but the two configurations have an overlapping region in the middle of the phonetogram. This region enables the passing between the two mechanisms. Following the work presented in [48], the frequency range where this passing can occur is about one octave (or 12 semi-tones). The main characteristic of this passing is to provoke a break in the fundamental frequency (F_0). Thus, when producing an increasing glissando from M_1 to M_2, there is an average 8 semi-tones break, whereas it is approximately 12 semi-tones when performing a decreasing glissando. Breaking intervals probabilities are depicted on Figure 7 and Figure 8. In the first one we can actually see that the frequency breaks also depends on the fundamental frequency where it occurs.

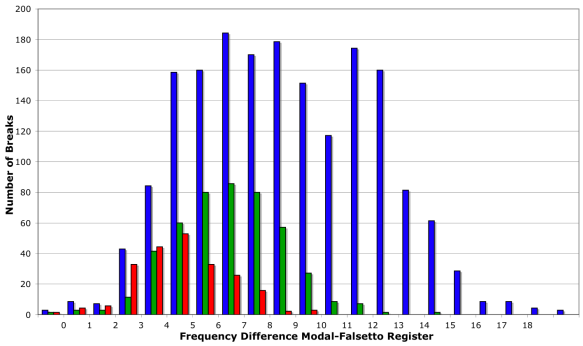


Figure 7: Frequency drops densities in semi-tones from Chest (or Modal) to Falsetto register. In blue, when the break happens at $200Hz$, in green at $300Hz$, in red at $400Hz$ [48].

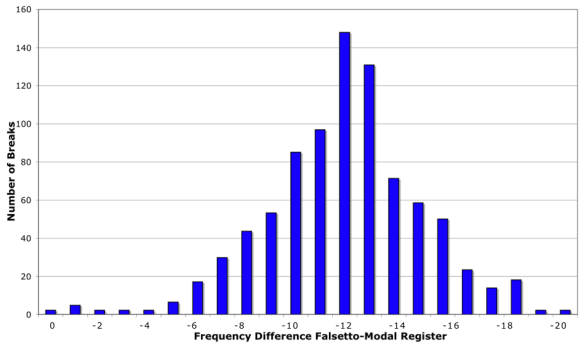


Figure 8: Frequency drops densities in semi-tones from Falsetto to Chest (or Modal) register [48].

In other words, this phenomenon introduces an hysteresis. For most untrained speakers or singers this break is uncontrollable whereas trained singers are able to hide more or less smoothly this break, although they cannot avoid the mechanism switch altogether.

6. THE VOCAL TRACT

In this section, we describe the implementation of a vocal tract model. This module is based on a physical "tube-based" representation of a vocal tract filter, which is simultaneously controllable using geometrical (area) and spectral (formant) parameters.

6.1. The lattice filter, a geometrical approach of vocal tract representation

Linear Predictive Coding [49] is a method for representing the spectral envelope of a digital signal of speech in compressed form, with the information given by a linear predictive model. The order of the filter is related to the complexity of the envelope, and also the number of control parameters. Thus, to represent a five-formant singing vowel, a filter containing five pairs of conjugated poles (for the resonances) is needed, adding up to a total of ten parameters for the vocal tract.

The LPC parameters (commonly named a_i) are non linearly interpolable. This implies that, for two configurations $[a_1 a_2 ... a_n]$ and $[b_1 b_2 ... b_n]$ corresponding to two vowels, a linear interpolation between both of these vectors will not correspond to a linear interpolation between the two spectra, and could even lead to unstable combinations. For these reasons, we will use another implementation of the LPC filter: the *lattice filter*. The control parameters of such a filter are called *reflection* coefficients (commonly named k_i). Such a filter is represented in Figure 9. It is composed of different sections, each characterized by a ki parameter.

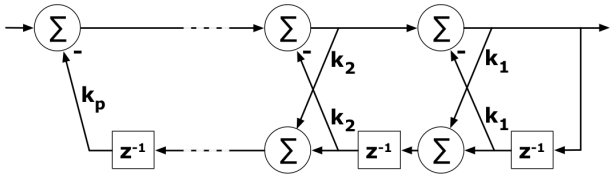


Figure 9: Representation of k_p cells of a lattice filter.

The reflection coefficients correspond to physical characteristics of the vocal tract, which may be represented by a concatenation of cylindrical acoustic resonators, forming a lossless tube. This physical model of the lattice filter is represented in Figure 10. Each filter section represents one section of the tube; the forward wave entering the tube is partially reflected backwards, and the backward wave is partially reflected forwards. The reflection parameter k_i can then be interpreted as the ratio of acoustic reflections in the i^{th} cylindrical cavity, caused by the junction impedance with the adjacent cavity. This value varies from 1 (total reflection) to -1 (total reflection with phase inversion), and is equal to 0 when there is no reflection.

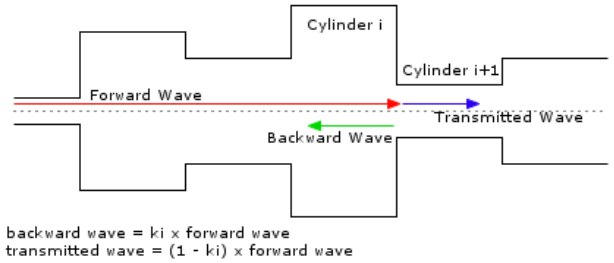


Figure 10: Geometrical interpretation of the lattice filter: transmitted and backwards waves at each cell junction.

The filter will be stable if the k_i parameters are between -1 and 1. However, there is no direct relationship between these parameters and sound: a small modification of k_i does not imply a small modification of the spectrum. Thus, instead of using the reflection coefficients, we will be using the different cylinder areas, named A_i, which can be easily deducted from the reflection coefficients with the following expression:

$$\frac{A_i}{A_{i+1}} = \frac{1+k_i}{1-k_i}$$

By acting on these A_i parameters, the interpreter is directly connected to the physical synthesis instrument. The sound spectrum will then evolve with acoustical coherence, which makes it more natural to use. Moreover, the stability of the filter is guaranteed for all A_i values.

6.2. Coefficients Convertion Framework

In order to use the area parameters of the lattice filter (A_i), a Max/MSP object was created to convert them to k_i values which are used in the lattice filter. Several sets of A_i parameters corresponding to different vowels were calculated. After selecting one of these presets, certain sections of the vocal tract can be modified by a percentage ΔA_i, which has the effect of opening or closing that section of the oral cavity.

A second approach to controlling the lattice filter was considered: a formant-based scheme was used to represent the spectral enveloppe, and the formant features, F_i, were converted to k_i parameters (after conversion to the LPC a_i coefficients), and then to A_i areas to control the lattice filter. This allowed us to easily model certain phenomena that are well known in speech processing, like overtone singing or the singer formant [50, 51], by acting on analytical parameters (the formants) rather than geometrical parameters (the areas). Similarly to the control of the areas, the formants have presets for different vowels and can be modified by a percentage ΔFi.

The parameters conversion framework described above is represented in Figure 11.

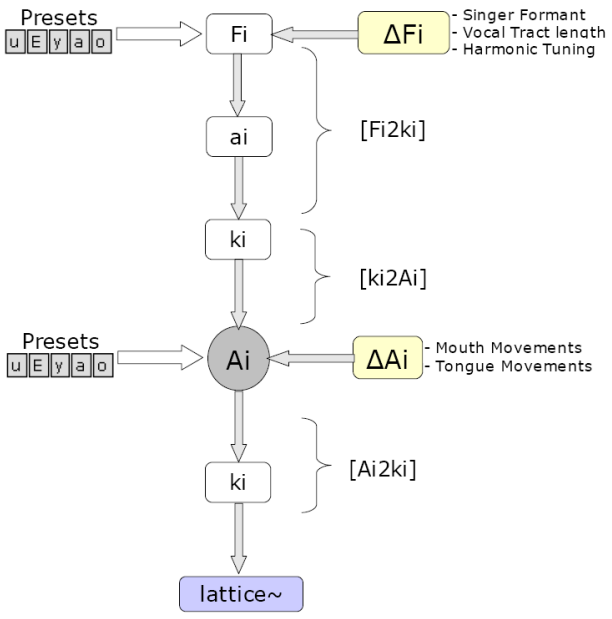


Figure 11: Coefficients convertion and presets/modifications framework, allowing user to modify spectrum-related and shape-related features at the same time.

7. ABOUT THE REAL-TIME CONTROL OF VOICE SYNTHESIS

In this section, we comment some experiments we conducted in order to evaluate expressive and performing abilities of systems

we developed. Modules which were integrated together inside Max/MSP and various control devices (and in various combinations) were dynamically connected using a mapping matrix. This set of tests allowed us to acheive efficient configurations employing several different performing styles (classical singing, overtone singing, etc.).

7.1. Concerning Voice Source

In order to be able to compare expressive skills of this system with the one developed before [2, 32, 36], we decided to keep the same control scheme: a graphic tablet. In that way, we were able to evaluate clearly the improvements achieved using the new mapping functions. Early experimentation demonstrated to us that independent control of tenseness and vocal effort significantly increased performance possibilities. Current mapping equations still provide some unlikely parameters combinations, resulting, for example, in "ultra-tensed" perception or unexpected variations of dynamics.

The implementation of the phonetogram is also a major improvement in term of naturalness. It gives better results in terms of expressivity rather than that without monitored control of loudness which is more linear. Although we have significantly investigated the frequency break phenomenon, we have not yet integrated it into the system, as we did not find a satisfying solution for controlling it. It is not straightforward to translate this frequency break in the control domain, and as our hand gestures are mainly continuous or used as basic switch from one configuration to another is not really satisfying from a musical point of view, and can result in breaks in frequency range and "wrong" notes.

7.2. Concerning Vocal Tract

The vocal tract was controlled using a data glove [52] as shown in Figure 12. The glove was mapped to the area parameters of the lattice filter in four different ways:

- The folding of the fingers control the opening angle of the mouth (represented in Figure 13) (see Figure 14)
- The hand movement along the z-axis controls the position of the "tongue" in the vocal tract (towards the back or the front of the mouth)
- The hand movement along the y-axis controls the vertical position of the tongue (near or far from the palate) (see Figure 14)
- The hand movement along the x-axis changes the vowels (configurable from one preset to another, for example from an */a/* to an */o/*)

This configuration allowed us to achieve vocal tract modification techniques such as overtone singing quite easily. Indeed, as the spectral representation (F_i) is very efficient to configure for some presets (e.g. offset vowel) or to leave running on automatic tasks (e.g. harmonic/formant tuning), the constant access to geometrical "delta" features (ΔS_i) allows the user to refine techniques (e.g. lowering the vocal tract, changing tongue position, etc.) and thus increase expressivity.

7.3. Incidental Remarks

Overall, at this stage of development, the synthesiser allows control of 17 parameters, namely : pitch, vocal effort, tenseness, mechanisms, the first two formants, the singers formant, vocal tract length, gain, transition between vowels, width of the vocal tract, position of the tongue and mouth opening (5 parameters).

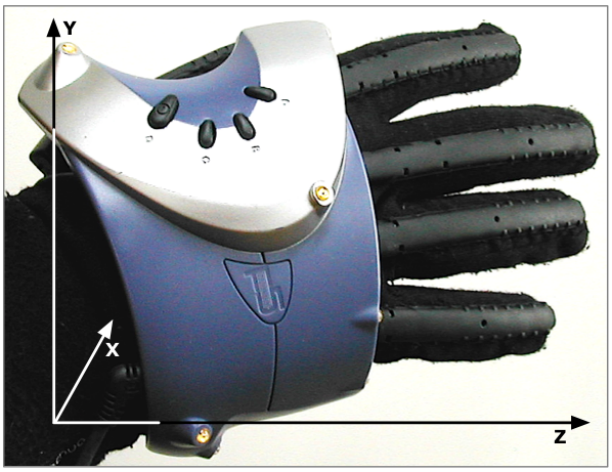


Figure 12: Vocal tract control with a data glove: 5 finger flexion sensors and 3 dimensions (x,y,z) tracking.

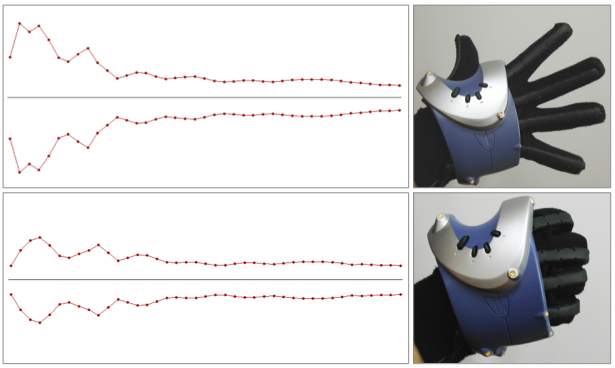


Figure 13: Mouth opening control: finger flexion sensors mapped to variation of 9 first A_i.

Considering all of these parameters, only the actions on mechanisms is not a continuous parameter, thus 16 parameters must be monitored using continuous parameters. From the controller side, we have 17 continuous parameters (out of 33), meaning that we are actually theoretically able to control all needed parameters. However, the problem is that from user's perpective, it is impossible to manipulate three interfaces at the same time. There are actually two solutions: one is to have multiple users (2 or 3) being in control of the interfaces, the other is to use one-to-many mappings, allowing the performer to control several parameters with the same controller.

8. CONCLUSIONS

In this work, our main aim was to build a high performance musical instrument allowing a wide range of expressive singing possibilities. Our actual work resulted in the implementation of new models for voice source and vocal tract, in real-time, which will be strategic tools in order to further this work. Improvements in expressivity of this new system have encouraged us to go forward with this approach. Moreover, our modular architecture inspires us to move towards a highly extensible synthesis platform which will be useful in the integration of other results from existing and forthcoming vocal production techniques.

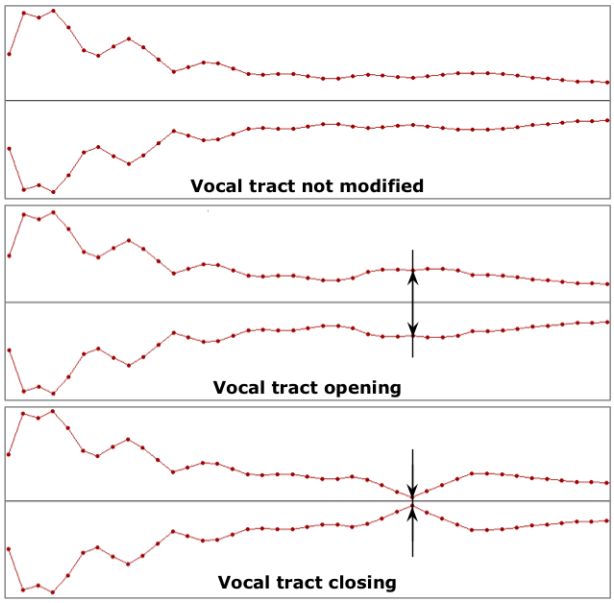


Figure 14: Control of the vertical position of the tongue.

9. ACKNOWLEDGMENTS

The authors would like to thank SIMILAR Network of Excellence (and thus the European Union) which has provided resources to allow researchers from all over Europe to meet, share and work together, thus achieving exciting results. We also would like to thank the Croatian organisation team of eNTERFACE'06, led by Prof. Igor Pandzic, where most of this work were done. Finally, we would like to thank our respective laboratories (TCTS Lab, Mons, Belgium and LIMSI-CNRS, Paris, France) who have adapted their research agendas in order to allow us to collaborate in this project.

10. REFERENCES

[1] http://www.loquendo.com/. 31

[2] C. d'Alessandro, N. D'Alessandro, S. L. Beux, J. Simko, F. Cetin, and H. Pirker, "The Speech Conductor: Gestural Control of Speech Synthesis", in *Proceedings of eNTERFACE'05 Summer Workshop on Multimodal Interfaces*, 2005. 31, 32, 33, 37

[3] M. Kob, "Singing Voice Modelling As We Know It Today", *Acta Acustica United with Acustica*, vol. 90, pp. 649–661, 2004. 31

[4] http://www.virsyn.de/. 31

[5] http://www.vocaloid.com/. 31

[6] X. Rodet and G. Bennet, "Synthesis of the Singing Voice", *Current Directories in Computer Music Research*, 1989. 31

[7] X. Rodet, "Synthesis and Processing of the Singing Voice", in *Proceeding of the First IEEE Benelux Workshop on Model-Based Processing and Coding of Audio (MPCA-2002)*, (Leuven, Belgium), 2002. 31

[8] P. Cook, *Identification of Control Parameters in an Articulatory Vocal Tract Model, with Applications to the Synthesis of Singing*. Ph.d. thesis, Standford University, 1990. 31

[9] J. Moorer, "The Use of the Phase Vocoder in Computer Music Application", *Journal of the Audio Engineering Society*, vol. 26, no. 1-2, pp. 42–45, 1978. 32

[10] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech Modifications Based on a Harmonic plus Noise Model", in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 550–553, 1993. 32

[11] M. Macon, L. Jensen-Link, J. Oliviero, M. Clements, and E. George, "A Singing Voice Synthesis System Based on Sinusoidal Modeling", in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 435–438, 1997. 32

[12] K. Lomax, *The Analysis and the Synthesis of the Singing Voice*. Ph.d. thesis, Oxford University, 1997. 32

[13] Y. Meron, *High Quality Singing Synthesis Using the Selection-Based Synthesis Scheme*. Ph.d. thesis, University of Michigan, 2001. 32

[14] P. Cano, A. Loscos, J. Bonada, M. de Boer, and X. Serra, "Voice Morphing System for Impersonating in Karaoke Applications", in *Proceedings of the International Computer Music Conference*, 2000. 32

[15] B. Doval, C. d'Alessandro, and N. Henrich, "The spectrum of glottal flow models", *Acta Acustica*, vol. 92, pp. 1026–1046, 2006. 32

[16] L. Kessous, "A two-handed controller with angular fundamental frequency control and sound color navigation", in *Proceedings of the 2002 Conference on New Interfaces for Musical Expression (NIME-02)*, 2002. 32

[17] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow", *STL-QPSR*, vol. 4, pp. 1–13, 1985. 32

[18] B. Doval and C. d'Alessandro, "The voice source as a causal/anticausal linear filter", in *proc. Voqual'03, Voice Quality: Functions, analysis and synthesis, ISCA workshop*, (Geneva, Switzerland), Aug. 2003. 32, 33

[19] B. Larson, "Music and Singing Synthesis Equipment (MUSSE)", *Speech Transmission Laboratory Quarterly Progress and Statut Report (STL-QPSR)*, pp. (1/1977):38–40, 1977. 32

[20] P. Cook, "SPASM: a Real-Time Vocal Tract Physical Model Editor/Controller and Singer: the Companion Software System", in *Colloque sur les Modèles Physiques dans l'Analyse, la Production et la Création Sonore*, 1990. 32

[21] J. O. Smith, "Waveguide Filter Tutorial", in *Proceedings of the International Computer Music Conference*, pp. 9–16, 1987. 32

[22] V. Välimäki and M. Karjalainen, "Improving the Kelly-Lochbaum Vocal Tract Model Using Conical Tubes Sections and Fractionnal Delay Filtering Techniques", in *Proceedings of the International Conference on Spoken Language Processing*, 1994. 32

[23] X. Rodet, "Time-Domain Formant Wave Function Synthesis", vol. 8, no. 3, pp. 9–14, 1984. 32

[24] X. Rodet and J. Barriere, "The CHANT Project: From the Synthesis of the Singing Voice to Synthesis in General", *Computer Music Journal*, vol. 8, no. 3, pp. 15–31, 1984. 32

[25] N. Henrich, *Etude de la source glottique en voix parlée et chantée*. Ph.d. thesis, Université Paris 6, France, 2001. 32, 34

[26] G. Fant, *Acoustic theory of speech production*. Mouton, La Hague, 1960. 32

[27] D. Klatt and L. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *J. Acous. Soc. Am.*, vol. 87, no. 2, pp. 820–857, 1990. 32, 34

[28] R. Veldhuis, "A Computationally Efficient Alternative for the Liljencrants-Fant Model and its Perceptual Evaluation", *J. Acous. Soc. Am.*, vol. 103, pp. 566–571, 1998. 32

[29] A. Rosenberg, "Effect of Glottal Pulse Shape on the Quality of Natural Vowels", *J. Acous. Soc. Am.*, vol. 49, pp. 583–590, 1971. 32

[30] G. Fant, "The LF-Model Revisited. Transformations and Frequency Domain Analysis", *STL–QPSR*, 1995. 32

[31] B. Bozkurt, *Zeros of the Z-Transform (ZZT) Representation and Chirp Group Delay Processing for the Analysis of Source and Filter Characteristics of Speech Signals*. PhD thesis, Faculté Polytechnique de Mons, 2004. 33

[32] N. D'Alessandro, C. d'Alessandro, S. Le Beux, and B. Doval, "Realtime CALM Synthesizer, New Approaches in Hands-Controlled Voice Synthesis", in *NIME'06, 6th international conference on New Interfaces for Musical Expression*, (IRCAM, Paris, France), pp. 266–271, 2006. 33, 34, 37

[33] D. Zicarelli, G. Taylor, J. Clayton, jhno, and R. Dudas, *Max 4.3 Reference Manual*. Cycling'74 / Ircam, 1993-2004. 33

[34] D. Zicarelli, G. Taylor, J. Clayton, jhno, and R. Dudas, *MSP 4.3 Reference Manual*. Cycling'74 / Ircam, 1997-2004. 33

[35] M. Puckette, *Pd Documentation*. 2006. http://puredata.info. 33

[36] C. d'Alessandro, N. D'Alessandro, S. L. Beux, and B. Doval, "Comparing Time-Domain and Spectral-Domain Voice Source Models for Gesture Controlled Vocal Instruments", in *Proc. of the 5th International Conference on Voice Physiology and Biomechanics*, 2006. 34, 37

[37] R. Schulman, "Articulatory dynamics of loud and normal speech", *J. Acous. Soc. Am.*, vol. 85, no. 1, pp. 295–312, 1989. 34

[38] H. M. Hanson, *Glottal characteristics of female speakers*. Ph.d. thesis, Harvard University, 1995. 34

[39] H. M. Hanson, "Glottal characteristics of female speakers : Acoustic correlates", *J. Acous. Soc. Am.*, vol. 101, pp. 466–481, 1997. 34

[40] H. M. Hanson and E. S. Chuang, "Glottal characteristics of male speakers : Acoustic correlates and comparison with female data", *J. Acous. Soc. Am.*, vol. 106, no. 2, pp. 1064–1077, 1999. 34

[41] M. Castellengo, B. Roubeau, and C. Valette, "Study of the acoustical phenomena characteristic of the transition between chest voice and falsetto", in *Proc. SMAC 83, vol. 1*, (Stockholm, Sweden), pp. 113–23, July 1983. 34

[42] P. Alku and E. Vilkman, "A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers", *Folia Phoniatr.*, vol. 48, pp. 240–54, 1996. 34

[43] H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children", *J. Acous. Soc. Am.*, vol. 107, no. 6, pp. 3438–51, 2000. 34

[44] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo, "On the use of the derivative of electroglottographic signals for characterization of non-pathological phonation", *J. Acous. Soc. Am.*, vol. 115, pp. 1321–1332, Mar. 2004. 34

[45] N. Henrich, C. d'Alessandro, M. Castellengo, and B. Doval, "Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency", *J. Acous. Soc. Am.*, vol. 117, pp. 1417–1430, Mar. 2005. 34

[46] N. Henrich, G. Sundin, D. Ambroise, C. d'Alessandro, M. Castellengo, and B. Doval, "Just noticeable differences of open quotient and asymmetry coefficient in singing voice", *Journal of Voice*, vol. 17, no. 4, pp. 481–494, 2003. 35

[47] N. Henrich, "Mirroring the voice from garcia to the present day: Some insights into singing voice registers", *Logopedics Phoniatrics Vocology*, vol. 31, pp. 3–14, 2006. 35

[48] G. Bloothooft, M. van Wijck, and P. Pabon, "Relations between Vocal Registers in Voice Breaks", in *Proccedings of Eurospeech*, 2001. 35

[49] J. D. Markel and A. H. Gray, *Linear prediction of speech*. Springer-Verlag, Berlin, 1976. 36

[50] B. Story, "Physical modeling of voice and voice quality", in *proc. Voqual'03, Voice Quality: Functions, analysis and synthesis, ISCA workshop*, (Geneva, Switzerland), Aug. 2003. 36

[51] G. Carlsson and J. Sundberg, "Formant frequency tuning in singing", *J. Voice*, vol. 6, no. 3, pp. 256–60, 1992. 36

[52] http://www.vrealities.com/P5.html. 37

Le projet ORA

Le projet ORA (Orgue et Réalité Augmentée) est rapporté ici à titre informatif, car il ne fait pas partie de l'étude menée au cours de mon doctorat à proprement parler. Ce projet est issu d'une collaboration entre une dizaine de personnes du LIMSI, dans le cadre d'un festival de vulgarisation scientifique, le festival Sciences sur Seine 2008. Ce travail a consisté à l'élaboration d'une installation audio-visuelle autour de l'orgue de l'église Ste-Elisabeth à Paris, qui a pu être présentée au cours de deux concerts nocturnes les 15 et 17 Mai 2008.

Comme décrit en détails dans les deux articles suivants, cette installation avait pour but de capter les données des différentes parties de l'orgue à l'aide de microphones, pour d'une part réaliser des effets audio-numériques diffusés sur un ensemble de 8 haut-parleurs disposés autour du public et d'autre part effectuer une analyse spectrale du son provenant des différentes microphones pour afficher une visualisation "transposée" de ces données sur les tuyaux visibles de l'orgue.

Au cours de ce projet, j'étais en charge de l'analyse spectrale des microphones et du mapping des données afin d'envoyer aux ordinateurs en charge de la partie graphique les données analysées en temps réel. Une description simple de l'élaboration du projet, ainsi que des photos et des vidéos des concerts sont disponibles à l'adresse suivante : http://vida.limsi.fr/index.php/Orgue_Augmentee

Le premier article, présenté à la conférence ICMC en 2009 est centré sur les aspects acoustiques, audio-numériques et musicaux du projet, tandis que le second article, présenté à Smart graphics 2009 se focalise plus sur les aspects graphiques. A la suite de cette seconde conférence, nous avons été invité à publier une version longue de cet article, pour un numéro spécial de l'IJCICG (International Journal of Creative Interfaces and Computer Graphics) à paraître au cours de l'année 2010.

THE ORA PROJECT: AUDIO-VISUAL LIVE ELECTRONICS AND THE PIPE ORGAN

Christophe d'Alessando, Markus Noisternig (*), Sylvain Le Beux, Lorenzo Picinali, Brian FG Katz, Christian Jacquemin, Rami Ajaj, Bertrand Planes, Nicolas Strurmel, Nathalie Delprat

LIMSI-CNRS, BP 133 – F91403, Orsay, France
* IRCAM, place Igor Stravinsky, F75004, Paris, France

ABSTRACT

This paper presents musical and technological aspects of real-time digital audio processing and visual rendering applied to a grand nineteen-century pipe organ. The organ is "augmented" in both its musical range and visual dimensions, thus increasing its potential for expression. The discussed project was presented to a public audience in the form of concerts. First, a brief project description is given, followed by in-depth discussions of the signal processing strategies and general musical considerations. Digital audio effects allow for the addition of new electronic registers to the organ stops. The "direct" sound is captured inside the organ case close to the pipes in order to provide "dry" audio signals for further processing. The room acoustic strongly affects the pipe organ sound perceived by the listener; hence, to combine the processed sound with the organ sound both room simulation and spatial audio rendering are applied. Consequently, the transformed sound is played back via a multitude of loudspeakers surrounding the audience. Finally, considerations of musical aspects are discussed, comprising reflections on virtuosity and technique in the musical play and how the new possibilities could affect composition practice and the use of the organ in contemporary music.

1. INTRODUCTION

The use of live-electronics and augmented instruments is common practice in contemporary music performance and media art. In general practice, the instrument's direct sound is captured with microphones or other kinds of sound pick-up, processed in real-time, and played back over loudspeakers. Applying audio signal processing and spatial rendering to grand pipe organs in the same manner as with orchestra instruments raises many interesting questions, which will be discussed in this article. Previous pieces using live-electronic processing applied to the pipe organ are described in [1]. The pipe organ timbre depends not only on the instrument itself, but also on the interaction with the room acoustic; the instrument's 'inner sound' significantly differs from the sound perceived by the listeners. It is well known in organ music interpretation and composition that the organist plays with the organ and the corresponding room acoustic response in a unique way. This inner / outer sound dialogue is envisaged from a technical and musical point of view in Section 3.

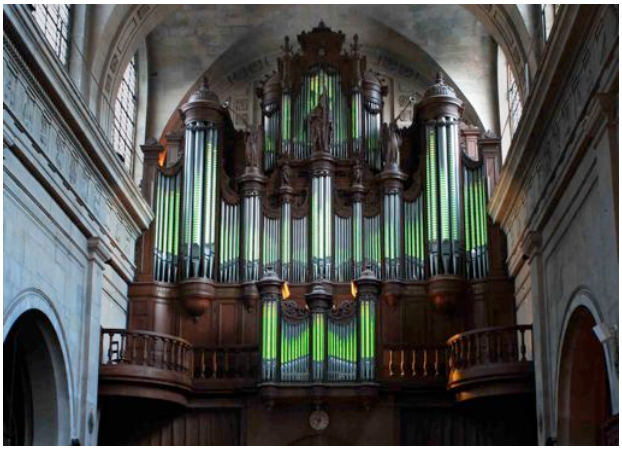


Figure 1. Snapshot of the visual animation during an ORA concert; the instrument's façade is lively animated by the played music.

Another contribution of this project is to extend registration possibilities by mixing the direct acoustic sound of the instrument with virtual registers computed in real-time using digital audio effects on the acoustic pipe sound (Section 4).
New sound possibilities, new registration possibilities, and new spatial sound restitution can be used for the classical repertoire, but call also for new music (section 5).

2. THE ORA PROJECT

The ORA (*fr. Orgue et Réalité Augmentée*) project was realized in the Sainte Elisabeth church in Paris, spring 2008. The primary goal was to immerse the audience in a unique multimodal audio-visual experience. For visualization the spectral content of the played music was projected as digital VU-meters on the organ's façade pipes (see Figure 1; and [4] for details on visual aspects).
The instrument is a large romantic organ (1853) comprising 39 stops, 42 registers, 3 manual keyboards, and

a pedal-board. Its base footprint is about 10x10 m, with three main levels for the four divisions (2322 pipes, 141 visible at the organ façade).

3. INNER/OUTER SOUND: CAPTURE AND SPATIAL AUDIO

3.1. Inner sound capture and audio setup

To decouple the sound from the influence of the room acoustic, it is captured inside the organ case. The achievable sound quality strongly depends on the number and positions of microphones used. Five omnidirectional microphones were placed in the *Positif* (1), the *Grand Orgue* and *Pédale* (3), and the *Récit* (1). Typically, omnidirectional microphones have a more extended low frequency response and lower distortion than directional microphones. To avoid distortions caused by the very high sound pressure level close to the pipes, microphones with a very high dynamic range have been used. The divisions of the organ case are acoustically well separated, so that the captured sound is well decoupled from neighboring divisions (at least for mid and high frequencies).

The sound captured as described above contains significantly less reverberation, higher low frequency energy, and provides better articulation, higher precision, and clearer transients than the sound outside the organ. As the attacks of the inner sound are very precisely defined, it offers much more variation to the organist for playing with articulation.

The microphone signals are processed in real-time using Pure Data and are re-distributed to 8 Haliaetus Blackbird loudspeakers surrounding the audience. An active Genelec 7071A subwoofer is fed with a low-pass filtered sum of the surround loudspeaker feed signals. In addition, the power spectral density of each microphone is estimated in third-octave bands and each sub-band level is sent to the graphical rendering units via Ethernet (UDP). Figure 3 illustrates the audio setup.

3.2. Spatial audio reproduction and effects

Combining the processed sounds with the organ sounds at the listener's position requires spatial redistribution and additional room simulation. The proposed environment uses third-order Ambisonics for sound field reproduction in the horizontal plane [3] [7]. The inherent free-field assumption of the Ambisonics approach, i.e. the basic assumption of reproducing the sound field in a source free medium, usually limits its use in real situations. Especially in churches, strong early reflections and long reverberation times – necessary for the traditional organ sound – deteriorate the accuracy of the sound field reproduction and degrade subjective localization. The limitation to a finite number of playback channels further reduces the area of accurate sound field reproduction (sweet spot). Applying weighting functions before decoding the Ambisonics signals into loudspeaker feed signals broadens the sweet spot area but also increases the perceived source width. To overcome the problems linked to a real reproduction environment as mentioned above, the audio effect design for extended registration concentrates on principally non-localizable / non-focused sounds and spatial granular synthesis.

The use of multichannel sound field reproduction for spatialization and movement of sound through space adds a new compositional feature to organ music. It allows projecting the organ's "inner" sound into the "outer" space, thus changing its relation to the acoustic environment. Room acoustic and psychoacoustic effects determine the perceived dialogue between the acoustic and electroacoustic sound. For example, as the loudspeakers are closer to the audience than the organ pipes, the processed sound might - due to the shorter acoustic propagation path delay - arrive earlier at the listener than the direct organ sound. This creates the impression that the sound comes from the location of the loudspeaker not from the organ itself (Precedence Effect); the organ sound then is perceived as early reflection or reverberant energy, rather than as the direct sound. Many different spatiotemporal interrelations can be observed in this context, which also depend on the listener's position.

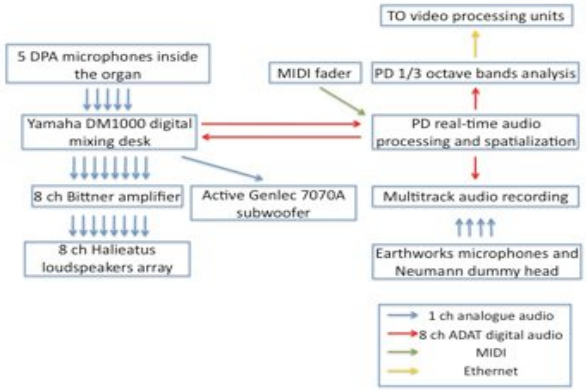


Figure 2. Audio setup

The following spatial effects have been used as a compositional element:

- **Virtual organ divisions**: the sound captured inside the different divisions of the organ case can be relocated in space; for instance, the Récit on top of the organ can be virtually placed behind the audience. The spatialized "inner" sound is reverberated by the room acoustics and can merge, interfere, or dominates the real organ sound.
- **Sound motion**: due to real-time spatial control the "inner" sound of a division can be moved around the audience in a time-varying manner, for instance to mimic a cortege of marching singers.
- **Reverberation effects**: adding directionally encoded early reflections and late reverberation to the "inner"

organ sound allows one to virtually enlarge the perceived spaciousness.

- **Spatial granular synthesis**: the "inner" sound is chopped into short sound snippets, which are randomly redistributed in space.

4. ACOUSTIC MIXTURES AND ELECTRONIC EFFECTS

From the very beginning, the organ sound is based on the combination of pipe ranks or registers. Augmenting the pipe organ by means of digital audio signal processing – i.e. by adding "electronic registers" - pushes the boundary of traditional organ sounds. Various effects can be added, which strongly depend on the acoustics of the different pipes. Acoustic documentation of this instrument is reported in [2].

4.1. Some acoustic features of pipe families

Reeds are characterized by a spectrum that is very rich in harmonics and a relatively loud tone, but are somehow unbalanced in their loudness profile; the bass is normally much louder than the high notes. In general transients are short (short attack time) and not very prominent. Reeds saturate the spectrum and are therefore often used for powerful, strong acoustic effects in traditional organ music. Some ranks, like the *Voix Humaine*, are built with short pipes yielding spectral resonances similar to the vocal formants of the human voice, close to the vowel /a/.
Due to the rich spectrum of reed pipes, audio effects derived from subtractive synthesis are best applicable; by adding further harmonics the resulting dense spectrum creates noise like sounds.
In contrast to reed pipes, flue pipe sounds are mainly limited to a few harmonics for some stopped ranks (*e.g.* the *Bourdon*). The transient sound is markedly different from the sustained sound, allowing for a large variety in articulation. Historically, flue pipes were combined to create sounds in an additive synthesis like manner: the so-called "mixture" or "mutation" registers. Flue pipes are well suited to additive synthesis resulting in harmonically rich sounds.

1.1 Additive and subtractive digital audio effects

In addition to spatial rendering two effect categories have been applied: "additive" effects that enrich the original sound, and "subtractive" effects that spectrally shape the original sound.
Within this project real-time harmonizers and ring modulators have been applied to the pipe sound captured inside the organ case [8] [9]. Harmonizing relates to mixing a sound with several pitch-shifted versions of itself. In practice the microphones capture the global sound of each organ division, rather than the sound of individual pipes. Applying a harmonizer to this polyphonic input signal produces many inharmonic partials, which add to the original spectrum of the signal, creating a dense and inharmonic sound. Various shifting ratios are used in order to produce different degrees of inharmonicity.
Reed pipe sounds have a dense frequency spectrum, which makes additive synthesis algorithms non-applicable. On the contrary, subtractive synthesis techniques allow to spectrally shaping the rich pipe organ sound. The Karplus-Strong synthesis technique provides a computational efficient and simple approach to subtractive synthesis [5]. The algorithm consists of a delay line and low-pass filter arranged in a closed loop simulating the reflected waves of a string. Using variable delays allows dynamic control of the resonance effects.
The effect of the application of the Karplus-Strong algorithm to *e.g.* reed pipes is variable spectral shaping. When playing in a fast tempo, the resulting sound has a sparkling quality, with fast formant motions like a human voice.

5. SOME MUSICAL CONSEQUENCES

It is difficult to apply audio effects to classical music repertoire without destroying its subtle musical content. Then only spatial audio and reverberation effects were used in conjunction with classical music.

The first type of effects is sound relocation in the church, i.e. the captured inner sound captured is played-back from different places in the church. The second type of effects is sound motion, which typically works very well with music accompanied by solo voices: like a singer moving in the church. A stronger effect is given by the slow extension and retraction of the sound of a division in the acoustic space, like a tide rising and falling. A third type of effects is the virtual acoustical enlargement of the room augmentation by adding artificial reverberation and early reflections. The relatively small church was acoustically transformed into a grand cathedral. The use of digital effects in classical music is somewhat paradoxical, as very often these effects are considered as euphonic as long as they do not sound "electronic", and therefore remain primarily unnoticeable.
A cycle of pieces in 12 parts was especially composed for this project. The main argument is to play with inner space and outer space, capturing inside and playing outside the instrument. This argument is also a metaphor for the music itself, based on a short text by Dorothée Quoniam: "les 12 degrés du silence" (" the 12 degrees of silence"). Quoniam, a 19th century Carmelite, explained to a young sister the teachings of her inner voice. Then the cycle is about speech, silence, inner and outer voices. It is played in alternation with classical repertoire music. This piece makes use of the unusual sound possibilities offered by the system.
The technical system used gives a successful fusion of direct sound and live electronics. Depending on the

balance between the direct sound and processed sound, the result can have an "electronic" quality or an "acoustic" quality, even with electronic modifications that are not perceived as such.

Different digital audio effects give different results depending on the type of pipes they are used with. "Additive" effects continue the tradition of "mutation" ranks in historical instruments; electronics allow for dynamic inharmonic ranks addition/suppression.

The inharmonicity provided by the harmonizer and reverberations transforms the pipe sounds in percussion-like sounds (see Figure 3). The harmonizer associated to all the foundation stops in the bass and medium bass register gives a very inharmonic sound, like a plate of metal played with a bow.

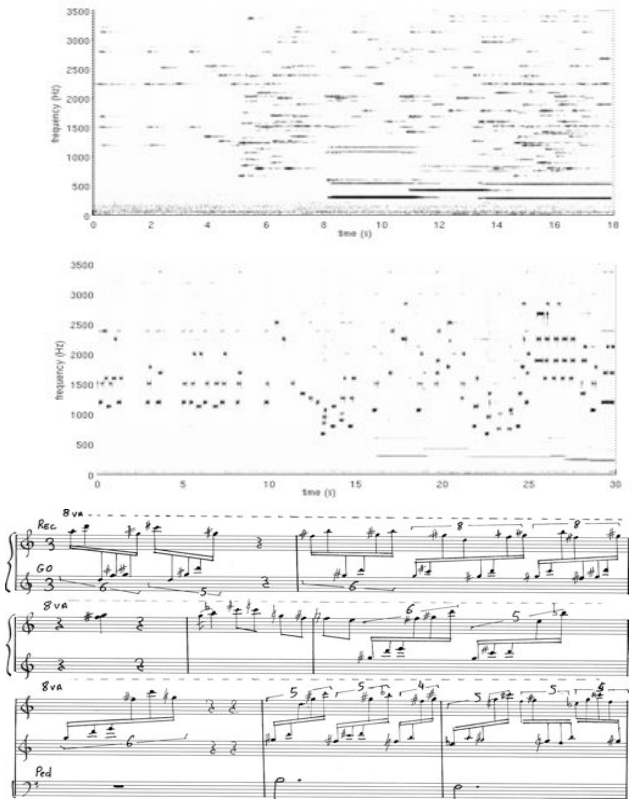


Figure 3. Effect of the harmonizer: Spectrogram of the outer sound in the church, with visible inharmonic partials. Middle: spectrogram of the inner sound in the Récit division, and bottom, corresponding score (played on 4' and 2' flutes, corresponding to 0-20'').

"Spectrum shaping" effects give formant like qualities to reed pipes (pipes with rich spectra). This would correspond to dynamic modifications of the pipe shape and dimension, and effect which is not possible with acoustic pipes. The Karplus-strong effect is used in conjunction with reeds. The variable filtering effect of the algorithm associated to the rich reed spectra gives a vocalic quality of the organ sound. The instrument speaks like a giant voice, but in an unknown and unintelligible language.

Spatial audio effects are an integral part of the instrument augmentation, and contribute much to the electronic/acoustic perceptual fusion. For instance, when impulse-like chords or clusters are exciting the virtual acoustics of a very large room, the short articulation silence between impulses is carefully controlled, in order to play with the real and virtual room as an acoustics filter.

In summary, this project demonstrated that live electronics and the pipe organ can extent the instrument's musical possibilities and repertoire, while maintaining its historical character. It is then possible to mix classical and contemporary music harmoniously. The augmented instrument is offering performers and composers new expression means.

Acknowledgments

This research was supported by the « Science sur Seine » program of City of Paris. Special thanks to the church Sainte Elisabeth, Paris and particularly F. Xavier Snoëk.

References

[1] S. Everett (Vanitas, 2005), and R. Uijlenhoet (Dialogo Sopra I due Sistemi, for organ, eight microphones, laptop and four speakers, 2003) http://ecmc.rochester.edu/ecmc25/concert8.pdf

[2] C. d'Alessandro « Voicing documentation of a pipe organ » *Acoustics'08, Acoustics 08, 155th ASA meeting, JASA, Vol 123, 1307, Paris, France*

[3] M. A. Gerzon, "Periphony: With-height sound reproduction", J. Audio Eng. Soc. 21(1), 2–10 (1973).

[4] C. Jacquemin, R. Ajaj, S. Le Beux, C. d'Alessandro, M. Noisternig, B. Katz, and B. Planes. The Glass Organ: Musical Instrument Augmentation for Enhanced Transparency. *SmartGraphics 2009, Salamanca. Spain*.

[5] K. Karplus, A. Strong "Digital Synthesis of Plucked String and Drum Timbres", Computer Music Journal (MIT Press). 7(2), 43–55 (1983).

[6] M. Noisternig, M., A. Sontacchi, T., Musil, R., Höldrich: A 3D ambisonic based binaural sound reproduction system. In: Proc. AES 24th International Conference, Banff, Canada (2003)

[7] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics", J. Audio Eng. Soc. 53(11), 1004–1025 (2005). .

[8] V. Verfaille, "Adaptive Digital Audio Effects (A-DAFx) : A new class of sound transformations", IEEE Trans. Audio, Speech and Language Proc. 14(5), 1817–1831 (2006).

[9] U. Zölzer, "DAFx – Digital Audio Effects," John Wiley and Sons, (2002).

The Glass Organ: Musical Instrument Augmentation for Enhanced Transparency

Christian Jacquemin[1], Rami Ajaj[1], Sylvain Le Beux[1], Christophe d'Alessandro[1], Markus Noisternig[2], Brian F.G. Katz[1], and Bertrand Planes[3]

1. LIMSI-CNRS, BP 133, 91400 Orsay, France
2. IRCAM, 1 pl. I. Stravinsky, 75004 Paris, France
3. Artist, 41 bis quai de la Loire, 75019 Paris, France

Abstract. The Organ and Augmented Reality (ORA) project has been presented to public audiences at two immersive concerts, with both visual and audio augmentations of an historic church organ. On the visual side, the organ pipes displayed a spectral analysis of the music using visuals inspired by LED-bar VU-meters. On the audio side, the audience was immersed in a periphonic sound field, acoustically placing listeners inside the instrument. The architecture of the graphical side of the installation is made of acoustic analysis and calibration, mapping from sound levels to animation, visual calibration, real-time multi-layer graphical composition and animation. It opens new perspectives to musical instrument augmentation where the purpose is to make the instrument more legible while offering the audience enhanced artistic content.

Key words: Augmented musical instrument, Augmented reality, Sound to graphics mapping, Real-time visualization

1 Introduction

Augmented musical instruments are traditional instruments modified by adding controls and mono- or cross-modal outputs (e.g. animated graphics) [1, 2]. Augmentation generally results in a more complex instrument (on the player's side) and a more complex spectacle (on the spectator's side). The increased functionality and controllability of the instrument might eventually distort the perceived link between the performer's gestures and the produced music and graphics. The augmentation might confuse the audience because of its lack of transparency. In contrast, the increased functionality could enhance the perceived link.

We agree on the interest of musical instrument augmentation that extends a traditional instrument, preserves and enriches its performance and composition practices. This article focuses on a rarely stressed use of augmentation that increases the comprehension and the legibility of the instrument instead of increasing its complexity and its opacity. Our research on output augmentation follows the work on ReacTable [3], an augmented input for the control of electronic musical instruments. The ReacTable is a legible, graspable, and tangible control interface, which facilitates the use of an electronic instrument so as to

be accessible to novices. Its professional use confirms that transparency does not entail boredom and is compatible with long term use of the instrument.

This paper presents the principles and implementation of the Glass Organ, the augmentation of an historical church organ to enhance the understanding and perception of the instrument through intuitive and familiar mappings and outputs. It relies on the following main features:

- the visual augmentation is directly projected on the facade of the instrument (and not on peripheral screens),
- the visual augmentation is aligned in time and space: the visual rendering is cross-modally synchronized with an audio capture and the graphical projection is accurately aligned with the organ geometry,
- the augmentation does not affect the musician's play. It can adapt to traditional compositions and deserves the creation of new artworks,
- the augmentation is designed to gain a better understanding of the instrument's principle by visualizing hidden data such as spectral sound analysis and the spatial information distribution within the large instrument.

The aim of the Organ and Augmented Reality (ORA) project was to visually and acoustically augment the grand organ at Ste Elisabeth church in Paris. This project was funded by the city of Paris program for popular science "Science sur Seine". The aim was to explain in general sound and sound in space, with specifics relating to the context of live performances. Concert were accompanied by a series of scientific posters explaining the background and technical aspects of the project. This project involved researchers in live computer graphics and computer music, a digital visual artist, an organ player and composer, and technicians.[1] ORA has been presented to public audiences through two immersive concerts, with both visual and audio augmentation. On the visual part, the organ pipes displayed a visual spectral analysis of the music, inspired by LED-bar VU-meters. On the audio side, the audience was immersed in a periphonic sound field, acoustically placing listeners inside the instrument. This article focusses on visual augmentation, the audio side is more detailed in [4].

2 Visual Augmentation of Instruments

The augmentation of a musical instrument can either concern the interface (the capture of the performer's gestures, postures, and actions), the output (the music, the sound, or non-audio rendering) or the intermediate layer that relates the incoming stimuli with the output signals (the mapping layer). Since our approach minimizes the modification of the instrument's playing techniques, we focus on the augmentation of the mapping and output layers that can be used to enhance composition, performance, or experience.

[1] In addition to the authors, participants included Nathalie Delprat, Lorenzo Picinali, and Nicolas Sturmel. Videos of the event can be found on `http://www.youtube.com/watch?gl=FR&hl=fr&v=JlYVUtJsQRk` or from the project site `http://vida.limsi.fr/index.php/Orgue_Augmentee`.

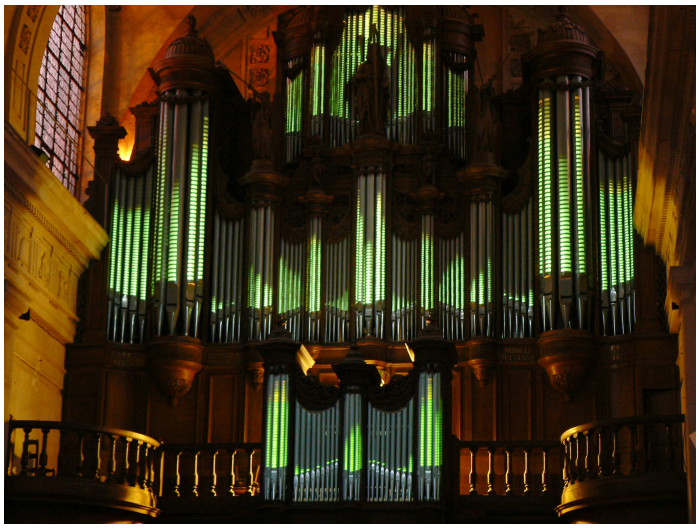


Fig. 1. ORA Concerts, May 15th and 17th, 2008

On the composer's side, Sonofusion [2] is a programing environment and a physically-augmented violin that can be used for performing multimedia compositions. Such compositions are "written" through programing, and are controlled in real-time by the performer through the additional knobs, sliders, and joystick. Whereas this work interestingly addresses the question of multi- and cross-modal composition and performance, it results in a quite complex control system. The variety of control devices yields a multiplicity of possible mappings; therefore, the correlation between the performer's gesture and his multimedia performance might seem arbitrary to the audience at times. Musikalscope [5], a cross-modal digital instrument, is designed with a similar purpose, and is criticized by some users for the lack of transparency between its visual output and the user's input.

On the audience side, the Synesthetic Music Experience Communicator [6] (SMEC) focuses on synesthetic cross-modal compositions that attempt to reproduce some of the visual illusions experienced by synesthetes. When compared with Sonofusion, SMEC has a better motivation for the graphic renderings because they are based on reports of visual illusions by synesthetes. This work however raises the question whether we can display and share deeply personal and intimate perceptions. Is it by displaying visual illusions that we are most likely to enhance the spectators' experience?

Visual augmentation can also address the human voice. Messa di Vocce [7] analyzes the human voice in real-time in order to generate a visual representation and interface used to control audio processing algorithms. The autonomous behavior of the graphical augmentation of voice almost creates an alter ego of the performer. Such an augmentation is less arbitrary than the preceding examples, because it is governed by an "intelligent" program. Within these environments,

the spectators are however immersed by a complex story which they would not normally expect when attending a musical event.

3 Artistic Design

Visual Design. The visual artwork was designed to transform the instrument so that it would appear both as classical and contemporary, and so that visual augmentation would contrast with the baroque architecture of the instrument. Church organs are generally located high on the rear wall of the building. The audience faces the altar and listens to the music without seeing the instrument. Even if one looks at the organ, it is rare to see the actual organist playing, resulting in a very static visual performance experience. During the ORA concerts the seating was reversed, with the audience facing towards the organ at the gallery.

The acoustics of the church is an integral part of the organ's sound as perceived by the listeners. Through the use of close microphone capture, rapid signal processing, and a multichannel reproduction system, the audience is virtually placed inside the "organ" acoustic providing a unique sound experience. To outline the digital transformation of the organ music, VU-meters are projected on the visible pipes, making a reference to many audio amplifiers. These VU-meters dynamically follow the music and build a subtle visual landscape that has been reported as "hypnotic" by some members of the audience. The static and monumental instrument becomes fluid, mobile, and transparent.

Sound Effects & Spatial Audio Rendering. The organ is one of the oldest musical instruments in Western musical tradition. It provides a large pitch range, high dynamics, and can produce a great richness of different timbres; hence, it is able to imitate orchestral voices. Due to the complexity of pipe organ instruments, advances in organ building have been closely connected to the application of new technologies. In the twentieth century electronics have been applied to the organ (a) to control the key and stop mechanism of the pipes – the action is electro-pneumatic – and (b) to set the registration, *i.e.* the combination of pipe ranks. However, very little has been achieved so far for modifying the organ's sound itself. The ORA project directly processes the captured pipe organ sound in real-time using a multitude of digital audio effects and renders it via loudspeakers surrounding the audience. Therefore, the sound perceived by the listener is a common product of the pipe organ's natural sound, the processed sound, and the room acoustics. Spatial audio rendering is capable of placing inner sounds of the organ in the outer space, interacting differently with the natural room acoustic, which adds a new musical dimension.

Augmented Organ. Miranda and Wanderley [8] refer to augmented instruments[2] as "*the original instrument maintaining all its default features in the sense that it continues to make the same sounds it would normally make, but with the addition of extra features that may tremendously increase its functionality*". With this in mind, the ORA project aims to enrich the natural sound of

[2] In scientific articles augmented instruments are often called hybrid instruments, hyperinstruments, or extended instruments.

pipe organs through real-time audio signal processing and multi-channel sound reinforcement, to meet the requirements of contemporary and experimental music.

Audio signal processing algorithms require the capture of the direct sound of the instrument. To minimize the effects of the room acoustics, multiple microphones have been placed inside the organ's case (section 4.1). The algorithms consider that separate divisions of the organ have different tonal properties (timbre, dynamics, and pitch range) and often contrast other divisions. The microphone signals are digitally converted via multi-channel audio cards with low-latency drivers; the real-time audio processing is implemented in Pure Data [9]. Selected algorithms include ring modulation, harmonizer, phaser, and granular synthesis. Audio rendering used an 8-channel full-bandwidth speaker configuration along the perimeter of the audience area with the addition of a high-powered subwoofer at the front of the church, at the location opposite from the organ.

In recent years, a variety of multi-channel spatial audio systems have been developed, e.g. quadrophony, vector base amplitude panning (VBAP), wave field synthesis, and Ambisonics. The proposed environment uses third-order Ambisonics for 2D sound projection in space. Ambisonics was invented by Gerzon [10]. While the room acoustic provides reverberation, essential to the sound of the church organ, the presence of early reflections and late reverberation (inherent to the acoustics of churches) deteriorates sound localization accuracy. Different weighting functions, as described in [11], have been applied before decoding to widen or narrow the directional response pattern. However, the sound design mainly deals with the reduced localization accuracy by focussing on non-focused sounds and spatial granular synthesis

Musical Program. The event was designed as an organ concert, with a bit of *strangeness* added by a lively animation of the organ facade and live electronics transformations of the organ sound. The musical program followed the path of a "classical" program mixed with somewhat unusual digital augmentation. Pieces of the great classical organ repertoire (Bach, Couperin Franck, Messiaen) were alternated with a piece in 12 parts especially written by C. d'Alessandro for the event (exploiting the various musical possibilities offered by the sound capture, transformation, and diffusion system).

4 Architecture & Implementation

4.1 The Instrument

The instrument used for these performance is a large nineteen-century organ (listed as historical monument), with three manual keyboards (54 keys), a pedal board (30 keys), 41 stops, with mechanical action. It contains approximately 2500 pipes, of which only 141 are visible in the organ facade. The organ case is inscribed in a square of about 10x10 m. Pipes are organized in 4 main divisions: the "positif", a small case on the floor of the organ loft, corresponding to the first manual keyboard; the "grand orgue" and "pédale" divisions at the main

level, corresponding to the second manual keyboard and to the pedal board; and the "récit" division, a case of about the same size as the "positif", crowning the instrument, and corresponding to the third manual keyboard. The "récit" is enclosed in a swell-box.

Five microphones were placed in the instrument divisions according to figure 2 left. These divisions are relatively separate and somewhat sound isolated, so that the near-field sound captured in one region was significantly louder than that received from other regions. Therefore each region could be considered as acoustically "isolated" for sound effects purposes.

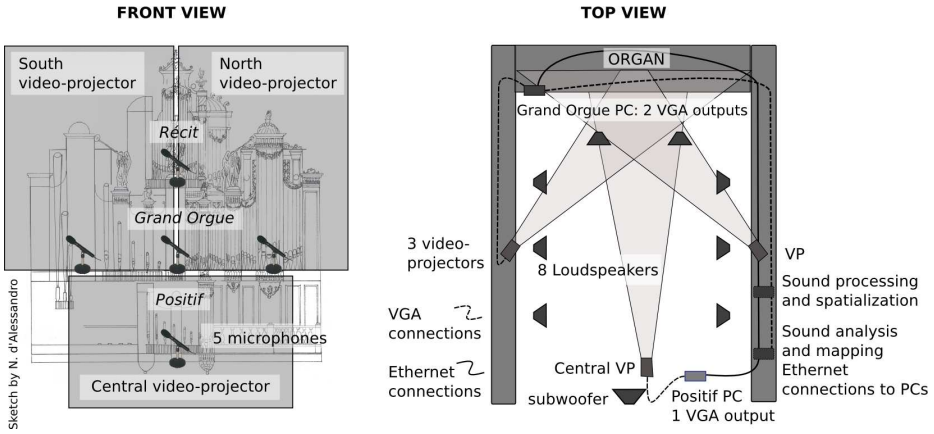


Fig. 2. Architecture of the installation: sound capture and video-projection

4.2 Architecture

Preliminary tests of video projection on the organ pipes showed that the pipes, despite their gray color and their specular reflections, were an appropriate surface for video-projection. The visual graphic ornamentation of the organ was using 3 video-projectors: 2 for the upper part of the instrument and 1 for the lower part (see figure 2 left). This figure also shows the rough locations where the microphones were placed to capture the direct sound of the instrument.

The sound captured inside the instrument was routed to a digital signal processing unit. Within this stage the captured sounds are processed and diffused back into the church as described in section 3. Traditional and contemporary organ music and improvisations were presented during the concerts. Classical organ music was spatialized and the reverberation time of the church was altered, resulting in the organ sound becoming independent from the organ location and the room sounding much larger than the actual Ste Elisabeth church. During the improvisation parts, the captured sounds were transformed and distorted applying real-time signal processing algorithms. As explained in section 4.4,

sound spectral analysis and sampling were used to compute the levels of the projected virtual VU-meters. These values were sent to the 3D renderer, and used as parameters of the vertex programs to animate the textures projected on the pipes and give the illusion of LED-bars. The right part of figure 2 shows the location of the video-projectors and the main data connections.

4.3 Graphic Rendering

Graphic rendering relies on Virtual Choreographer (VirChor)[3], a 3D graphic engine offering communication facilities with audio applications. The implementation of graphic rendering in VirChor involved the development of a calibration procedure and dedicated shaders for blending, masking, and animation. The architecture is divided into three layers: initial calibration, real-time compositing, and animation.

Calibration. The VU-meters are rendered graphically as quads that are covered by two samples of the same texture (white and colored LED-bars), depending on the desired rendering style. These quads must be registered spatially with the organ pipes. Due to the complexity of the instrument and its immobility, registration of the quads with the pipes was performed manually. Before the concert began, a still image digital photograph of the projection of a white image was taken with a camera placed on each video projector, near the projection lens. Each photo was loaded as a background image in Inkscape[4] and as many quads as visible pipes were manually aligned with the pipes in the editor. The amount of effort for this work was only significant the first time. Successive registrations (for each re-installation) amounted mostly to a slight translation of the previous ones, since attempts were made to locate the video-projectors in similar positions for each concert. The resulting SVG vector image was then converted into an XML scene graph and loaded into VirChor.

During a concert, the VU-meter levels are received from the audio analysis component (section 4.4) and are transmitted to the GPU which in turn handles the VU-meter rendering. GPU programming has offered us a flexible and concise framework for layer compositing and masking through multi-texture fragment shaders, and for interactive animation of the VU-meters through vertex shader parameterization. Moreover, the use of one quad for VU-meter per visual pipe handeled by shaders has facilitated the calibration process. Frame rate for graphic rendering was above 70 FPS and no lag could be noticed between the perceived sound and the rendered graphics.

Compositing. The graphical rendering is made of 4 layers: (1) the background, (2) the VU-meters, (3) the masks, and (4) the keystone (see left part of figure 3). The VU-meter layer is a multi-textured quad, and the background and mask layers are quads parallel to the projection plane that fill the entire display. Real-time compositing, homography, and control of background color are made through fragment shaders applied on these layers. The keystone layer (4)

[3] http://virchor.sf.net
[4] Inkscape is a vector graphic editor: http://www.inkscape.org

is a quad textured by the image generated by layers (1) to (3). The keystone layer is not necessarily parallel to the projection plane. The modification of the quad orientation is equivalent to applying a homography to the final image. This transformation enables slight adjustments in order to align the numerical rendering with the organ and compensate for any inaccuracies in the calibration phase. This transformation could be automatically computed from views of a calibration pattern [12]. Elaborate testing has shown that the background, VU-meter, and mask layers were perfectly registered with the physical organ, and thus made the keystone layer unnecessary. The mask layer is used to avoid any projection of the VU-meters onto the wooden parts of the organ, and to apply specific renderings through the background layer seen by transparency.

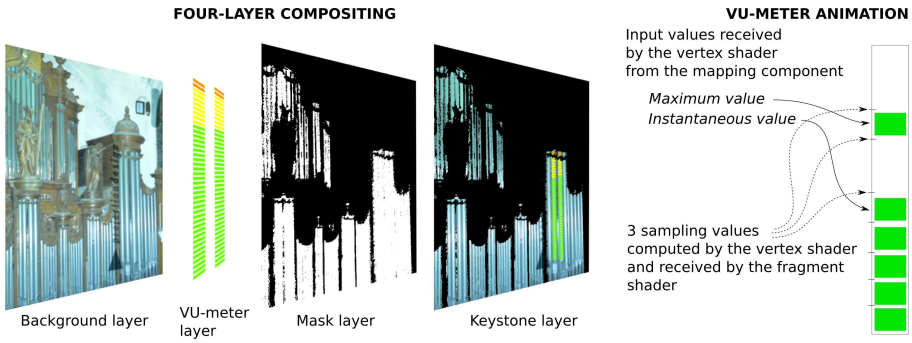


Fig. 3. Multi-layer composition and VU-meter animation through sampling

Animation. The VU-meter layer is made of previously calibrated quads exactly registered with all the visible pipes of the organ. The texture for VU-meter display is made of horizontal colored stripes on a transparent background (42 stripes for each pipe of the Grand Orgue and Récit and 32 stripes for each pipe of the Positif). The purpose of the animation is to mimic real LED-bar VU-meters that are controlled by the energy of their associated spectral band (see next section). The VU-meter levels received from the sound analysis and mapping components are sampled to activate or deactivate the bars. The level sampling performed in the vertex shader and applied to each quad was based on a list of *sampling values*. Considering that the height of a VU-meter texture is between 0 and 1, a sampling value is the height of a transparent interval between two stripes that represent 2 bars. A texture for 42 LED-bars has 43 sampling values. The sampling values are then transmitted to the fragment shader that only displays the bars below the sampled instantaneous value and the bar associated with the sampled maximal value. The resulting perception by the audience is that LED-bars are activated and deactivated.

Each virtual VU-meter receives the instantaneous value and the maximum value for the past 500ms (typical peak-hold function). They are sampled into 3 values by the vertex shader: the instantaneous sampled value, and the sampled

values below and above the maximal value. These samples are sent to the fragment shader that displays the texture between 0 and the first sampled value and between the second and third sampled values (see right part of figure 3).

4.4 Analysis & Mapping

This section describes the real-time audio analysis and mapping for VU-meter visualization. Most of the approximately 2500 organ pipes are covered by organ case, while only the 141 of the facade are visible to the audience. As such, a direct mapping of frequency played to visual pipe is not relevant, due to the large number of hidden pipes. In the context of ORA, the main purpose of the correspondence between audio data and graphical visualization was:

1. to metaphorically display the energy levels of the lowest spectral bands on the largest pipes (resp. display the highest bands on the smallest pipes)[5],
2. to maintain the spatial distribution of the played pipes by separating the projected spectral bands in zones, corresponding to the microphone capture regions and thereby retaining the notion of played pipe location,
3. to visualize the energy of each spectral band in the shape of a classical audio VU-meter (as in many audio hardware amplifiers and equalizers). The display was based on the instantaneous value of the energy and its last maximal value with a slower refreshing rate.

Analysis In order to estimate the mapping of sound level values to VU-meter projection, pre-recordings have been analyzed (figure 4). This analysis allowed us to roughly estimate the overall range of the various sections and to cut these spectral ranges into different frequency bands, according to the evolution of the harmonic amplitudes over frequency. The analysis resulted in a maximum spectral range of 16 kHz for the Positif and Récit sections of the organ, and 12 kHz and 10 kHz for the central and lateral parts of the Grand Orgue.

Each spectral band is further divided into subbands corresponding to the number of visually augmented pipes, *i.e.* 33 for Positif and Récit, 20 for the lateral and 35 for the central Grand Orgue. The subbands were not equally distributed over frequency range (warping) in order to gain a better energy balance between low and high frequencies. For re-calibration the energy of the lowest subband (the largest pipe) was used as reference signal.

Mapping The real-time spectral analysis consists of three stages: estimation of the power spectral density for each subband, mapping, and broadcasting over IP. The concert mapping process is described on figure 5.

Power spectral density (PSD). The PSD is estimated via periodograms as proposed by Welch [13]. The buffered and windowed input signal is Fourier transformed (Fast Fourier Transform, FFT) and averaged over consecutive frames.

[5] Since only few pipes are visible, the exact note of a pipe was not necessarily falling into the frequency band displayed on its surface.

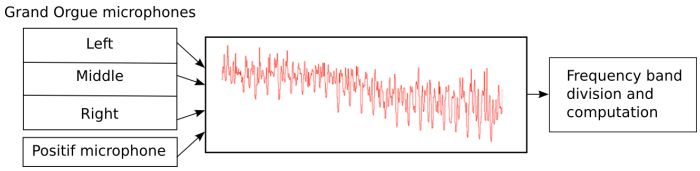


Fig. 4. Spectral sound analysis

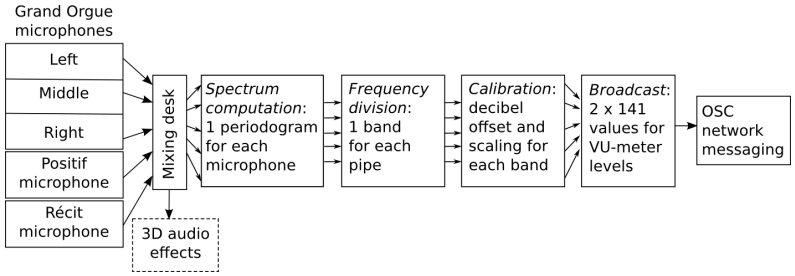


Fig. 5. Mapping between sound analysis and graphics

Assuming ergodicity the time average gives a good estimation of the PSD. One has to note, that through long term averaging the estimated subband levels are not sensitive to short peaks, they represent the root mean square (RMS) value. The decay of the recursive averaging has been set such that the VU-meter values changed smoothly for visual representation. The actual averaging was such that every incoming frame was added to the last three buffered frames.

Frequency band division. These periodograms are then transmitted as five 512 point spectra in order to correct the spectral tilt. The second part of the processing performs the division of the Welch periodograms into 141 frequency bands. Since the number of visible pipes in each section of the organ was inferior to 512 (resp. 33, 20, 35, 20, and 33) an additional averaging must be made in order to map the whole frequency range to the pipes of the organ. According to the spectral tilt, lower frequency bands (app. below 1.5 kHz) had more energy, thus only 3 frequency bands were added for the biggest pipes, whereas up to 30-40 bands were added for the highest frequency range (app. above 8 kHz).

Calibration. The third and most critical part is the calibration of the VU-meter activations through a scaling of the frequency band dynamics to values ranging from 0 to 1. The null value corresponds to an empty VU-meter (no sound energy in this frequency band), and 1 to a full VU-meter (maximum overall amplitude for this frequency band). To calibrate the output, so that 0 would correspond to no sound, we applied the pre-calculated decibel shifts computed from frequency band analysis of the initial recordings. The 1 value corresponds approximately to a 30 dB amplitude dynamic in each frequency band. After the shift, a division by 30 is made so that every VU-meter would vary from the lowest to the highest position on the associated organ pipe during the concert.

This method raised the following difficulties:

1. for each session, the microphones were located in a position slightly different from the preceding one, thus slightly changing the various amplitude levels due to the close proximity to different pipes,
2. the mixing desk dealt with both audio effects and mapping, and due to the slight presence of feedback between microphones and loudspeakers, maximum levels were configurated according to the 3D audio composition part for every concert. This turned out to change the offsets of the VU-meter calibration for each concert,
3. the dynamics of the pipes depended on the loudness of the concert pieces. These variations resulted either in a saturation or in a lack of reaction of the corresponding VU-meters,
4. the electric bellows system for pressurized air supply for the organ generated a low-frequency noise,
5. even though the microphones were inside the organ, there was some interference between the sound of the instrument and the sounds inside the church (audience applause and loudspeakers). Some of the spectators noticed the action of their hand claps on the visualization, eventually using this unintended mapping to transform the instrument into an applause meter.

To cope with these problems, before the beginning of the concert, approximately half an hour was devoted to the manual correction of the different shifts for each pipe, with the air pressurizer switched on. In order to deal with the variations of dynamics between the concert pieces, we decided to monitor the dynamics of each organ section with a slider. Last, the applause effects were cancelled through a downward shift of all the section sliders after each piece.

Broadcast. The third and last part of the process was the concatenation of all frequency band values into a list. Values were scaled between 0 and 1 and doubled, in order to give the spectator the impression of a real VU-meter with an instantaneous value and a peak-hold maximum value. Hence two lists of 141 values were sent to the visual module through ethernet (current frequency bands amplitudes and corresponding last maxima).

5 Perspectives

Technically, the geometrical and musical calibrations could be improved and automatized. By equipping the instrument with fiducials, the quads could be automatically re-aligned with the organ pipes if the video-projectors are slightly displaced. On the mapping side, the background noise detection could be improved by automatically detecting the decibel amplitudes of the different frequency bands and calibrating the lowest values of the VU-meters. Automatic individual maximum detection would allow for amplitude calibration so that the VU-meters take the full range of graphical animation during the whole concert.

The ORA project has shown that the audience is receptive to a new mode of instrument augmentation that does not burden the artistic expression with

additional complexity, but instead subtly reveals hidden data, and makes the performance both more appealing and more understandable. This work could be extended in several directions. First, graphical ornamentation could be applied to smaller and non-static musical instruments by tracking their spatial location. Second, graphical visualization could concern other physical data such as air pressure, keystrokes, or valve closings and openings that would require more sensors in the instrument than just microphones. Visualization could also concern the fine capture of ambient sounds such as audience noise, acoustic reflections, or even external sound sources such as street noise.

References

1. Bouillot, N., Wozniewski, M., Settel, Z., Cooperstock, J.R.: A mobile wireless augmented guitar. In: NIME '07: Proc. of the 7th international conference on New interfaces for musical expression, Genova, Italy (June 2007)
2. Thompson, J., Overholt, D.: Sonofusion: Development of a multimedia composition for the overtone violin. In: Proc. of the ICMC 2007 International Computer Music Conference. Volume 2., Copenhagen, Denmark (August 2007)
3. Jordà, S., Geiger, G., Alonso, M., Kaltenbrunner, M.: The reactable: exploring the synergy between live music performance and tabletop tangible interfaces. In: TEI '07: Proc. of the 1st international conference on Tangible and embedded interaction, New York, NY, USA, ACM (2007) 139–146
4. d'Alessando, C., Noisternig, M., Le Beux, S., Katz, B., Picinali, L., Jacquemin, C., Ajaj, R., Planes, B., Strurmel, N., Delprat, N.: The ORA project: Audio-visual live electronics and the pipe organ. In: Submitted to ICMC 2009. (2009)
5. Fels, S., Nishimoto, K., Mase, K.: Musikalscope: A graphical musical instrument. IEEE MultiMedia **5**(3) (1998) 26–35
6. Lewis Charles Hill, I.: Synesthetic Music Experience Communicator. PhD thesis, Iowa State University, Ames, IA, USA (2006)
7. Levin, G., Lieberman, Z.: In-situ speech visualization in real-time interactive installation and performance. In: NPAR '04: Proc. of the 3rd international symposium on Non-photorealistic animation and rendering, New York, ACM (2004) 7–14
8. Miranda, E.R., Wanderley, M.: New Digital Musical Instruments: Control And Interaction Beyond the Keyboard (Computer Music and Digital Audio Series). A-R Editions, Inc., Madison, WI, USA (2006)
9. Puckette, M.S.: Pure data: Another integrated computer music environment. In: Proc., International Computer Music Conference. (1996) 37–41
10. Gerzon, M.A.: Periphony: With-height sound reproduction. J. Audio Eng. Soc. **21**(1) (1973) 2–10
11. Noisternig, M., Sontacchi, A., Musil, T., Höldrich, R.: A 3D ambisonic based binaural sound reproduction system. In: Proc. AES 24th International Conference, Banff, Canada (2003)
12. Raskar, R., Beardsley, P.: A self-correcting projector. In: Proc. of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), Kauai, HI, USA, IEEE Computer Society (2001) 504–508
13. Welch, P.: The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. IEEE Trans. Audio Electroacoust. **AU-15** (June 1967) 70–73