HAL open science

Analyse de la qualité vocale appliquée à la parole expressive

Nicolas Sturmel

To cite this version:

Nicolas Sturmel. Analyse de la qualité vocale appliquée à la parole expressive. Musique, musicologie et arts de la scène. Université Paris Sud - Paris XI, 2011. Français. NNT : 2011PA112021 . tel-00591638

HAL Id: tel-00591638

https://theses.hal.science/tel-00591638v1

Submitted on 9 May 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N° D'ORDRE

THÈSE DE DOCTORAT

SPECIALITE : PHYSIQUE

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

Présentée par : Nicolas Sturmel

Sujet :

Analyse de la qualité vocale appliquée à la parole expressive

Soutenue le 2 Mars 2011 devant les membres du jury :

M. Thierry Dutoit (président)

M. Christophe d'Alessandro (directeur de thèse)

M. Yves Laprie (rapporteur)

M. Gaël Richard (rapporteur)

M. Boris Doval (examinateur)

M. Olivier Rosec (examinateur)

Remerciements

Une thèse est un travail de longue haleine, et sans l'aide de mes proches, de mes collègues et de mes mentors, elle n'aurait pas lieu d'être. Il suffit parfois d'un conseil, d'une main tendue, d'une porte ouverte pour faire germer une idée, une ambition. Pendant cette thèse, et plus que jamais, j'ai apprécié les vertus de l'approfondissement, de la rigueur, du travail d'équipe et du partage scientifique.

J'aimerais tout d'abord remercier Christophe d'Alessandro pour son encadrement sans faille. Il aura su me guider tout au long de cette thèse afin qu'elle ait la forme que vous trouverez aujourd'hui en lisant ce document. Merci à Boris Doval qui m'aura encadré au LIMSI pendant les premières années de ce travail ; sa rigueur aura toujours été d'une grande aide pour préciser ma pensée scientifique.

Merci ensuite à Gaël Richard et Yves Laprie qui ont accepté de rapporter sur ce travail, leurs remarques et conseils ont permis de pousser plus loin la qualité du document final. Merci à Thierry Dutoit et Olivier Rosec d'avoir accepté d'assister à cette soutenance et pour leurs remarques a posteriori. C'est un grand honneur et un grand plaisir de présenter son travail devant les personnes qui font activement partie de la communauté scientifique de l'analyse des signaux vocaux.

Merci à mes collègues du LIMSI pour cette agréable ambiance de travail, et en particulier au groupe Audio et Acoustique. Merci à Albert Rilliard et Sylvain Le Beux pour leur investissement dans la relecture et la critique du manuscrit. Merci à Tifanie, Lionel, Marc, Gaëtan et David pour leurs remarques qui m'ont aidé à perfectionner ma soutenance.

Merci aux personnes qui m'ont accompagné pendant les années de monitorat : à Guy Demoment, Patrick Gonord, Thomas Rodet, Frédérique Giorgiutti, Alexandre Renaux, Clarisse Hamadache et Delphine Monnier qui m'ont guidés dans mes premiers pas d'enseignement et mont donné les outils pour transmettre et partager mes connaissances.

Merci au personnel technique et administratif du LIMSI, de l'Université Paris Sud et de l'école doctorale STITS.

Merci à Thomas Helie et Bertrand David qui m'ont donné lors de mes tous premiers stages, dans une période charnière, le goût du traitement du signal audio et merci pour leur soutien qui a contribué au commencement de cette thèse. Merci à Cécile Durieu qui, durant mon cursus à l'ENS de Cachan, m'aura donné les bases et la rigueur que j'utilise quotidiennement en traitement des signaux.

Et comment ne pas remercier ma famille toute entière mais spécialement mon père, ma mère, mes soeurs et ma compagne, Charlotte et nos amis qui m'ont tous permis de traverser ces années de thèse sans faillir de leurs soutiens et de leurs encouragements. Une pensée particulière pour André qui, j'en suis sûr, aurait été fier de me voir réussir cette entreprise.

"La physique ressemble à la plus exigeante et parfois à la plus destructive des maîtresses. Nuit et jour, été, hiver, matin et soir, elle vous poursuit, vous envahit, vous comble ou vous désespère" - Georges Charpak

Table des matières

Introduction **9**

I Modélisation et État de l'art **15**

1 Modèle de la production vocale **17**
1.1 Le signal vocal : production et modélisation 19
1.2 Les différentes échelles du signal vocal 21
1.3 Le cycle glottique et sa caractérisation 23
1.4 Le modèle du point de vue signal 24
1.5 Qualités vocales 32
1.6 Conclusion 35

2 État de l'art de l'estimation des paramètres de la source **39**
2.1 La détection des instants de fermeture glottique 41
2.2 Filtrage inverse et caractérisation de la source 46
2.3 Périodicités, Apériodicités 62
2.4 Conclusion 68

II Outils pour l'analyse de la qualité vocale **71**

3 Ondelettes pour l'analyse des signaux vocaux **73**
3.1 Méthode multi-échelles et application aux signaux vocaux. 75
3.2 Etude prospective : ondelettes appliquées aux signaux de parole 77
3.3 Méthode LoMA pour la détection de GCI 82
3.4 LOMA pour la mesure de l'énergie relative 90
3.5 Shimmer et jitter par les ondelettes 94
3.6 Quotient ouvert et ondelettes 99
3.7 Parallèle avec Mean Square Phase 105
3.8 Conclusion 106

4 Décomposition Périodique/Apériodique **109**
4.1 Amélioration de l'algorithme PAP 110
4.2 Application à des signaux de tests 114
4.3 Application à des signaux réels 120
4.4 Impact de la décomposition sur l'estimation des LoMA 123
4.5 Conclusion 127

5 Estimation des paramètres de la source glottique **129**
5.1 Validation des Zéros de la Transformée en Z comme technique de séparation source/filtre . . . 131
5.2 Précision nécessaire pour l'estimation de O_q et α_m . . . 143
5.3 Formalisation du modèle pour l'extraction des paramètres . . . 143
5.4 Mesures préliminaires . . . 149
5.5 Protocole d'analyse sur signaux naturels . . . 152
5.6 Méthode hybride combinant ZZT et LoMA pour l'estimation du quotient ouvert . 155
5.7 Conclusion . . . 160

III Application à de la parole expressive **163**

6 Analyse d'un grand corpus **165**
6.1 Constitution de la base . . . 167
6.2 Analyse et Protocole . . . 168
6.3 Résultats . . . 169
6.4 Confirmation des tendances par analyse statistique . . . 175
6.5 Interactions source-filtre . . . 177
6.6 Corrélation entre les estimations . . . 177
6.7 Caractérisation des styles . . . 181
6.8 Conclusion . . . 184

7 Conclusion **189**

Références **195**

IV Annexes **203**

A Analyses complémentaires du grand corpus de parole naturelle et expressive **205**

Table des figures

1.1 Vue des éléments du larynx. Issu de *Gray's Anatomy*, 20ème édition (1918). . . . 19
1.2 Modélisation de la production vocale comme une succession de filtres linéaires. En parallèle est donné un modèle source/filtre comme celui utilisé en prédiction linéaire. 20
1.3 Décomposition d'un signal de parole en temps au niveau de la phrase et de la période, et en fréquence. 21
1.4 Représentation du cycle glottique par vidéo ultra rapide, électroglottographie (EGG) et dérivée de l'EGG, extrait de [Henrich *et al.*, 2004]. 23
1.5 A) Un masque de Rothenberg équipé d'une mesure de pression intraorale (source : site web de l'INPG). B) Une visualisation de la glotte par endoscopie. C) Une locutrice instrumentée pour une acquisition EGG (source : site web du LIMSI/CNRS). 24
1.6 Passage du modèle acoustique au modèle signal de la production vocale. Bien souvent, on associe la dérivation due au rayonnement du débit à l'onde de débit glottique (ODG). On considère donc généralement le modèle de la dérivée de l'onde de débit glottique (DODG) directement. 25
1.7 Le modèle LF [Fant *et al.*, 1985] et ses paramètres. En haut les paramètres normalisés, en bas les paramètres temporels. Forme du haut : DODG, forme du bas : ODG. 28
1.8 Effet de la variation du quotient ouvert sur la forme d'onde dérivée du débit glottique. (inspiré de [Doval *et al.*, 2006]). 28
1.9 Effet de la variation du quotient d'asymétrie de la forme d'onde dérivée du débit glottique (inspiré de [Doval *et al.*, 2006]). 29
1.10 Effet de la variation du quotient de retour sur la forme d'onde dérivée du débit glottique (inspiré de [Doval *et al.*, 2006]). 29
1.11 Position de la langue et des résonateurs du filtre pour 3 voyelles différentes. . . . 30
1.12 Placement de voyelles par leur lieu d'articulation. (alphabet phonétique international) 31
1.13 Modélisation du conduit vocal, décrire le conduit sous la forme de résonateurs permet de modéliser le conduit comme un filtre en treillis, pour plus tard utiliser une modélisation autoregresive. 32
1.14 Différents flux glottiques (en cm de mercure) obtenus avec un masque de Rothenberg pour différentes qualités vocales (figure 1 de [Sundberg, 1994]). 34

2.1 Utilisation du filtrage inverse pour la détection des GCI (fenêtre glissante de 20ms toutes les 10ms, signal échantillonné à 16kHz, 18 pôles estimés pour le filtre AR). Signal original en vert, résidu de la prédiction linéaire en bleu. 43
2.2 Illustration de l'utilité de la pente de la phase pour déterminer l'emplacement des impulsions dans un signal. 43

2.3 La méthode de Smits et al. [Smits et Yegnanarayana, 1995] appliquée à un signal réel, des oscillations peuvent causer des fausses détections. 43
2.4 Exemple problématique d'estimation des GCI sur un signal de parole avec la méthode DYPSA (DYnamic programming Phase SLope Algorithm) [Naylor *et al.*, 2007] avec le signal en vert, et les GCI détectés en mauve. 44
2.5 Le produit multi-échelles face à un produit de la partie basse fréquence du signal de parole. On constate que les résultats obtenus sont similaires, mais plus contrastés dans le cas du MSP 45
2.6 Exemple d'estimation LPC sur un signal de parole. 18 pôles, fenêtre de 20ms (pondération Hanning), superposition de 10ms. Temps en secondes, amplitudes arbitraires. 48
2.7 Illustration du positionnement des pôles pour les deux parties du modèle CALM. La forme d'onde de la partie ouverte ressemble bien à la réponse impulsionnelle d'un filtre du deuxième ordre anticausal stable. De même, la forme d'onde de la partie causale est la réponse impulsionnelle d'un filtre du premier ordre. 50
2.8 Illustration du défaut d'estimation des phases anti-causales par la prédiction linéaire. Le signal résiduel n'est pas un train d'impulsion synchrone, mais est déphasé. 50
2.9 Analyse du signal résiduel de la figure 2.8 par les ondelettes (sur la gauche) et son spectre de phase (sur la droite). Un maximum du déphasage est observé vers 2000Hz. 51
2.10 Comparaison des étapes de filtrage inverse par LPC et ZZT. La ZZT utilise moins d'*a priori* avant et après l'estimation. 52
2.11 Illustration du paradigme de la ZZT : Les signaux temporels (première ligne) de la source (première colonne) et de la réponse du filtre (deuxième colonne) sont convolués (troisième colonne) selon le modèle de production linéaire de la parole. En deuxième rang on peut visualiser leurs spectres d'énergie et enfin le troisième rang représente comment les zéros sont combinés par convolution. 53
2.12 La forme, position et taille de la fenêtre d'analyse utilisée en ZZT. 54
2.13 Algorithme de la décomposition source filtre par ZZT, point par point. 56
2.14 Exemple de décomposition par ZZT sur un signal synthétique (/a/, 160 Hz, voix modale - $0_q = 0.5; \alpha_m = 0.8$). 57
2.15 Exemple de décomposition par ZZT sur un signal réel (/a/, 120 Hz, voix modale - $0_q \approx 0.5; \alpha_m \approx 0.8$). 58
2.16 Calcul du quotient d'amplitude normalisé sur l'onde de débit glottique. Le calcul se fait traditionnellement sur l'ODG et la DODG obtenues par filtrage inverse LPC sur le signal vocal. 60
2.17 Variation de la valeur des paramètres NAQ et H1H2 pour différentes valeurs de O_q et α_m. Tests sur signaux synthétiques par le modèle LF. 61
2.18 Transformée de Fourier discrète de quelques fenêtres d'analyse, fenêtre de 1024 points, transformée sur 4096 points par complément de zéros. Seuls les bins de 0 à 40 sont représentés. 64
2.19 Séparation du spectre d'un signal de parole en une partie voisée et une partie non voisée. 65
2.20 Illustration de la décomposition périodique-apériodique sur le spectre du signal. Le trait continu représente le spectre du signal, le trait pointillé représente le filtre en peigne. Les itérations successives de la méthode PAP sont désignées par le trait discontinu. 67

3.1 Différentes réponses impulsionnelles et TF d'un banc de filtres en ondelettes dyadiques. 76
3.2 Détection de singularité sur un signal (en bas) par la TF (à droite) puis par la sortie d'un banc de filtres en ondelettes (à gauche). 77
3.3 Analyse par ondelettes de deux segments de signaux vocaux. La richesse spectrale visible sur le signal se traduit par l'observation des différentes échelles de décomposition. 78
3.4 Analyse par ondelettes d'un train d'impulsions, on retrouve bien l'alignement des maxima à travers les échelles. 79
3.5 Analyse par ondelettes d'une DODG, le filtrage du train d'impusions à l'origine de l'onde décale les maxima à travers les échelles. 79
3.6 Analyse par ondelettes d'un signal synthétique - signal de la figure 3.5 filtré par la fonction de transfert d'une voyelle /a/. Le filtre vocalique modifie d'avantage l'alignement des maxima. 80
3.7 Localisation du GCI par la position haute fréquence de la ligne (en rouge). 81
3.8 Le premier harmonique présente un décalage par rapport au GCI. 81
3.9 Ondelette retenue, une gaussienne modulée en fréquence pour a=16 et Fe=8000. . 83
3.10 Illustration de la détection des GCI par méthode multi-échelles : Lines Of Maximum Amplitudes - LoMA. 84
3.11 Protocole de validation de l'estimation des GCI par la méthode des LoMA sur un cas très défavorable. Les GCI détectés sur l'EGG sont considérés comme une référence. 86
3.12 Distribution de l'erreur de détection des GCI sur la base de données des signaux réels en micro-secondes. 87
3.13 Zoom sur un GCI estimé par les LoMA sur deux signaux produits par un locuteur masculin. Pour chaque figure, de bas en haut : ondelettes et signal acoustique (première case), DEGG puis Images de la vidéo, l'abscisse correspond aux échantillons du signal ($Fs = 44,1kHz$). Les épingles donnent les GCI estimés par LoMA (en rouge) et EGG (en vert). Les traits verts sur la DEGG donnent les instants d'échantillonnage de la vidéo, le trait rouge indique le GCI estimé par les LoMA. 91
3.14 Exemple d'un signal normalisé dans la base de données. De haut en bas : signal original et son spectrogramme, signal normalisé et son spectrogramme. Malgré la normalisation, il persiste des différences entre voyelles. Locuteur féminin, voyelle /a/ - fichier F8.wav. 93
3.15 Barycentre de la LoMA mesuré sur des signaux normalisés (en haut). Les signaux originaux (en bas) sont donné pour visualiser l'évolution de l'effort. Le barycentre de la LoMA est donné en Hz selon l'équation 3.7. 94
3.16 Estimation du jitter par la méthode LoMA sur la base synthétique de signaux. . 96
3.17 Estimation du shimmer par la méthode LoMA sur la base synthétique de signaux. 97
3.18 Effet d'une variation locale de la période de voisement - jitter - sur l'amplitude de voisement. Le repliement de la réponse des filtres peut causer des variations de l'amplitude de voisement. 98
3.19 Prédiction de la forme de la LoMA pour un débit glottique non filtré. Haut : variation du quotient ouvert, Bas : variation de l'asymétrie. De gauche à droite : forme de la DODG, phase du spectre en degrés, représentation temps fréquence du délai dû à la phase (prédiction de la forme de la LoMA) et zoom sur cette représentation au niveau de la fréquence fondamentale. Les temps et fréquences sont normalisés. 101

3.20 Prévision de la forme de la LoMA sur un signal synthétique complet (DODG filtrée). La phase du spectre n'est pas montrée ici (illisible), mais on constate que les prévisions restent identiques, à quelques oscillations prêt autour des formants (traits noirs sur la figure du milieu). L'amplitude du déplacement ne change pas non plus. Variation du quotient ouvert de 0.3 à 0.9. 102
3.21 Retard de groupe du signal filtré par rapport à l'instant de fermeture glottique autour de la fréquence fondamentale (unité). Variation du quotient ouvert de 0.3 à 0.9. 103
3.22 Analyse d'un signal synthétique (voyelle /a/, 133Hz - 120 échantillons) avec O_q variant de 0.3 à 0.9 selon une loi quadratique. Deux zooms sur l'analyse en ondelettes sont donnés au début et à la fin du signal. Le décalage varie linéairement avec O_q. 104

4.1 Algorithme de la décomposition PAP (inspiré de [d'Alessandro *et al.*, 1998]). . . . 111
4.2 Adaptation dynamique de la fenêtre d'observation en amont de la décomposition afin d'avoir toujours n périodes au minimum. S_L est le spectre de la fenêtre de longueur L observant le signal. 113
4.3 Adaptation de la durée d'observation L en fonction de la dispersion en fréquence fondamentale du signal. 114
4.4 Résultats d'estimation du RSB pour les 5 méthodes sur des signaux de tests à fréquence fondamentale fixe, en fonction de la quantité de jitter, de shimmer. . . 118
4.5 Erreur d'estimation de la partie apériodique pour 5 méthodes sur des signaux de tests à fréquence fondamentale fixe, en fonction de la quantité de jitter, de shimmer.119
4.6 Résultats d'estimation du RSB pour les 5 méthodes sur des signaux de tests à fréquence fondamentale variable, en fonction de la quantité de jitter, de shimmer. 120
4.7 Erreur d'estimation de la partie apériodique pour 5 méthodes sur des signaux de tests à fréquence fondamentale variable, en fonction de la quantité de jitter, de shimmer. 121
4.8 Décomposition PAP-A d'une voyelle /a/ produite par une locutrice féminine. $F_0 \approx 250Hz$. En haut, phonation très douce, O_q élevé, bruit important. En bas, phonation modale et claire. Sur la gauche, les formes d'onde ; sur la droite les spectres du signal, de la partie périodique et de la partie apériodique. 121
4.9 Exemple de décomposition sur le son [ʒə] : représentation des signaux temporels et de leurs spectres. 122
4.10 Spectrogrammes en bande étroite de décomposition périodique/apériodique par PAP et PAP-A du fichier C202. 124
4.11 Spectrogrammes en bande étroite de décomposition périodique/apériodique par PAP et PAP-A du fichier C203. 124
4.12 Spectrogrammes en bande étroite de décomposition périodique/apériodique par PAP et PAP-A du fichier M202. 125
4.13 Estimation du jitter et du shimmer avant et après décomposition périodique apériodique par méthode PAP (rond pointillés) et PAP-A (croix trait continu), l'estimation sur la source non bruitée est donnée par le trait mixte et l'estimation sur le signal bruité par le trait discontinu. 126

5.1 Boîtes à moustaches (voir texte) représentant la distance spectrale pour chaque méthode d'estimation pour le sous-corpus de voisement sans bruit. 134

5.2 Boîtes à moustaches (voir texte) représentant la distance spectrale pour chaque méthode d'estimation pour le sous-corpus de voisement standard. 134
5.3 Boîtes à moustaches représentant la distance spectrale pour chaque méthode d'estimation pour le corpus complet. 135
5.4 Représentation détaillée des distances spectrales pour le sous-corpus standard et le corpus complet. Moyenne des distances pour les conditions correspondant à la valeur du paramètre fixe. 135
5.5 Analyse d'un signal de voix produit par un locuteur, voyelle /a/ de fréquence fondamentale proche de 120Hz, voix modale. De haut en bas : signal original, signal EGG, analyse par autocorrélation, analyse itérative (IAIF) et analyse ZZT. Les 3 analyses donnent des résultats similaires. 138
5.6 Analyse d'un signal de voix produit par un locuteur, voyelle /u/ de fréquence fondamentale proche de 120Hz, voix modale. De haut en bas : signal original, signal EGG, analyse par autocorrélation, analyse itérative (IAIF) et analyse ZZT. 139
5.7 Analyse d'un signal de voix produit par un locuteur, voyelle /i/ de fréquence fondamentale proche de 120Hz, voix serrée. De haut en bas : signal original, signal EGG, analyse par autocorrélation, analyse itérative (IAIF) et analyse ZZT. . . . 140
5.8 Analyse d'un signal de voix produit par une locutrice, voyelle /a/ de fréquence fondamentale proche de 240Hz, voix modale. De haut en bas : signal original, signal EGG, analyse par autocorrélation, analyse itérative (IAIF) et analyse ZZT. 141
5.9 Abaques montrant les variations, respectivement de haut en bas, de α_m en fonction du rapport $\frac{F_g}{F_0}$, de α_m en fonction de la grandeur A (normalisée par T_0) et finalement du rapport $\frac{F_g}{F_0}$ en fonction de A (normalisée par T_0) pour différentes valeurs de O_q allant de 0.3 à 0.9. 145
5.10 Trois cas possibles pour l'estimation de l'ODG par ZZT, en fonction de la forme, il faut dériver ou intégrer le signal pour retrouver une ODG conforme au modèle LF. 147
5.11 Algorithme schématique de la méthode proposée pour l'estimation conjointe O_q-α_m. 148
5.12 Estimation combinée O_q-α_m sur un échantillon à qualité vocale variable. Si la mesure de O_q sur le signal suit bien les données EGG, il est important de sélectionner les valeurs de l'asymétrie couplées avec une erreur minimum de O_q sous peine de mal interpréter la mesure. 151
5.13 Programmation dynamique pour sélectionner la valeur de O_q la plus cohérente en fonction des valeurs précédentes. Le chemin continu est le chemin choisi par l'algorithme. Certains chemin en pointillés sont coupés, car les valeurs estimées ne sont pas toujours réalistes (dans ce cas, elles sont mises à NaN). 152
5.14 Algorithme modifié pour la méthode proposée. La différence principale réside dans la mesure de 3 fréquences différentes de F_g (une pour chaque valeur de ρ, c.f. texte) qui donnent chacune un couple O_q-α_m. Les valeurs sont ensuite sélectionnées par programmation dynamique puis moyennées. 153
5.15 Résultat de l'estimation du quotient ouvert sur la base de données A de voix parlée. Les résultats sont présentés sous forme d'histogrammes superposés : valeurs mesurées sur l'EGG (blanc), du nombre de détections pour chaque valeur dans chaque plage d'erreur précédemment décidée : le JND (noir) et 25% (gris). 154
5.16 Analyse de la base de données B par la méthode LoMA, mesure du quotient ouvert et distribution des détections pour les deux seuils. 157
5.17 Analyse de la base de données B par la méthode ZZT + PAP-A : mesure du quotient ouvert et distribution des détections pour les deux seuils. 157

5.18 Pondération arbitraire des deux méthodes pour favoriser les résultats obtenus par ZZT sur les O_q élevés. 157
5.19 Analyse de la base de données B par la méthode hybride LoMA + ZZT. 158
5.20 Estimation du quotient ouvert par LoMA, par ZZT (+PAP-A) et par hybridation des deux méthodes sur la base de donnée "brian" de VOCQUAL'03. 159
5.21 Estimation du quotient ouvert par LoMA, par ZZT (+PAP-A) et par hybridation des deux méthodes sur la base de donnée "singing" (fichiers JF-mem-6-A et LP-mem-6-a) de VOCQUAL'03. 159

6.1 Enchainement des méthodes d'analyse. 168
6.2 Analyse de la phrase numéro 20 : "vous êtes professeur de sciences politiques ?", pour trois styles (voix basse, vieux, didactique). De bas en haut : le signal, son spectrogramme en bande étroite, les valeurs mesurées pour O_q, le barycentre (en Hz), et le rapport Harmonique sur Bruit en dB. 170
6.3 Tendance générale par style sur les 4 paramètres estimés. 171
6.4 Histogramme des distributions statistiques par style du quotient ouvert. 172
6.5 Histogramme des distributions statistiques par style du jitter exprimé en %. . . . 172
6.6 Histogramme des distributions statistiques par style du rapport harmonique sur bruit exprimé en dB. 173
6.7 Histogramme des distributions statistiques par style du barycentre de la LoMA. . 173
6.8 Résultats de l'analyse par composantes principales des quatre paramètres sur tout le corpus 176
6.9 Dendogrammes sur toutes les analyses, par styles, pour le quotient ouvert et le barycentre de la LoMA. Distance euclidienne standardisée moyennée et non pondérée. Les groupements de couleurs indiquent des groupes aux éléments très proches les uns des autres. 178
6.10 Dendogrammes sur toutes les analyses, par styles, pour le jitter et le rapport harmonique sur bruit. Distance euclidienne standardisée moyennée et non pondérée. Les groupements de couleurs indiquent des groupes aux éléments très proches les uns des autres. 179
6.11 Représentation des moyennes des analyses pour chaque style en deux dimensions. Certains styles se démarquent plus clairement que d'autres. 185

A.1 Tendance par phonèmes, pour les valeurs mesurées du barycentre de la LoMA. . 206
A.2 Tendance par phonèmes, pour les valeurs mesurées du quotient ouvert. 207
A.3 Tendance par phonèmes, pour les valeurs mesurées du jitter. 208
A.4 Tendance par phonèmes, pour les valeurs mesurées du rapport harmonique sur bruit. 209
A.5 Tendance par styles, pour les valeurs mesurées du barycentre de la LoMA. 210
A.6 Tendance par styles, pour les valeurs mesurées du quotient ouvert. 211
A.7 Tendance par styles, pour les valeurs mesurées du jitter. 212
A.8 Tendance par styles, pour les valeurs mesurées du rapport harmonique sur bruit. 213
A.9 Hiérarchie des voyelles pour le style *contraste bas* 214
A.10 Hiérarchie des voyelles pour le style *dialogue* 215
A.11 Hiérarchie des voyelles pour le style *dictée* 215
A.12 Hiérarchie des voyelles pour le style *didactique* 216
A.13 Hiérarchie des voyelles pour le style *enjoué* 216
A.14 Hiérarchie des voyelles pour le style *fort* 217

A.15 Hiérarchie des voyelles pour le style *grave* 217
A.16 Hiérarchie des voyelles pour le style *insiste* 218
A.17 Hiérarchie des voyelles pour le style *narratif calme* 218
A.18 Hiérarchie des voyelles pour le style *narratif tendu* 219
A.19 Hiérarchie des voyelles pour le style *point* 219
A.20 Hiérarchie des voyelles pour le style *suspensif* 220
A.21 Hiérarchie des voyelles pour le style *narratif* 220
A.22 Hiérarchie des voyelles pour le style *vieux* 221
A.23 Hiérarchie des voyelles pour le style *virgule* 221
A.24 Hiérarchie des voyelles pour le style *voix basse* 222

Liste des tableaux

1.1 Tableau récapitulatif des liens entre qualité vocale et paramètre du modèle de production . . . 35

3.1 Sous-détections (MR) et fausses alarmes (FA) pour le locuteur masculin (M), féminin (F) et la totalité de la base (T). EGG comparé à DYPSA (DYP), LoMA (LOM) et LPC-LoMA (LPC). . . . 88
3.2 Résultat de la détection des GCI par LOMA . . . 90
3.3 Récapitulatif des paramètres utilisés pour la génération de la base de données de signaux synthétiques testant l'estimation du Jitter et du Shimmer par la méthode des LoMA. . . . 95

4.1 Ensemble des paramètres utilisés pour la base de données de signaux synthétiques pour le test des cinq méthodes de décomposition. . . . 116

5.1 Table des valeurs choisies pour la variation des paramètres lors de la création de la base de données de signaux synthétiques pour un total de 54880 conditions de tests. . . . 133
5.2 Résultats de l'analyse sur la base de données de voyelles expressives. Deux locuteurs pour trois voyelles et trois expressions. O_q : estimé par la méthode, $\hat{O}_q$ mesuré sur l'EGG, α_m est donné comme la moyenne des estimations appariées avec un O_q dans le JND. . . . 150

6.1 Énoncé et description des 16 consignes (appelés plus tard *styles*) de la base de données. . . . 167
6.2 Table donnant la correspondance entre l'appellation des voyelles dans ce chapitre et leur représentation phonétique. . . . 168
6.3 Corrélation des résultats par style. . . . 180
6.4 Corrélation des résultats par phonème. . . . 180

Introduction

Contexte scientifique

Dans le contexte actuel où les moyens classiques de communication homme-machine (clavier, pointeur) sont de plus en plus remplacés par des communications multi-modales, la parole joue un rôle très particulier par sa simplicité d'utilisation. En effet, tout un chacun sait communiquer avec la parole en transmettant bien plus que des informations linguistiques : la manière dont nous nous exprimons traduit aussi notre état affectif, notre âge, notre sexe ou notre origine. Tant en synthèse qu'en reconnaissance de la parole, cette dimension expressive est une composante essentielle pour la qualité de la communication homme-machine, qui permet de rendre l'interaction plus "naturelle".

L'analyse des signaux vocaux est un processus essentiel pour comprendre et modéliser les relations entre perception et production du signal vocal. La production de parole peut être analysée tant au niveau des mécanismes articulatoires qu'au niveau de sa forme acoustique. Pour ce faire, l'analyse des signaux de parole se fait classiquement sur des échelles de temps différentes.

Dans une échelle macroscopique le signal vocal est analysé pour en extraire le contenu phonétique et prosodique. Ce niveau d'analyse s'étend de quelques centièmes de secondes à plusieurs secondes : du trait phonétique à la phrase. A partir de cette échelle on cherche à segmenter le signal pour l'analyser à une échelle plus petite.

Cette deuxième échelle, de l'ordre de quelques millisecondes à quelques dixièmes de secondes, permet d'analyser les formes du signal vocal afin d'en extraire des informations sur la configuration de l'appareil de production du locuteur. Cette analyse peut se faire avec ou sans connaissances ou informations *a priori*. A cette échelle, on définit le *modèle source-filtre* qui sous entend la parole comme le résultat de la modification d'un signal source par un filtre. En première approximation et dans la langue française notamment, on associe la qualité vocale à la source, et l'articulation au filtre.

Pour comprendre les mécanismes articulatoires à l'origine des différentes formes de signaux vocaux, l'analyse des signaux acoustiques doit être complétée par des mesures plus précises et plus proches de l'appareil de production vocale. Des modélisations de production vocale ont été réalisées à l'aide de vidéos ultra rapides du larynx obtenues par endoscopie[Kendall, 2009], de radiographies de la partie supérieure du thorax et de la tête ou encore de mesures électroglottographiques [Fabre, 1957]. Certains paramètres de ce modèle de production vocale ont ainsi été reliés aux propriétés acoustiques du signal. Des études ont montré la relation directe entre les caractéristiques spectrales du signal vocal et la configuration du larynx [Hanson, 1994]. Il en ressort notamment que la configuration du larynx et des plis vocaux est prédominante dans la production et la perception de la parole expressive [Rosenberg, 1971].

Il s'agit ensuite de lier toutes les informations acoustico-articulatoires à leur perception par l'interlocuteur. Ces relations permettent une meilleure modélisation de la communication homme-machine par la prise en compte d'un modèle de parole plus réaliste. Une des solutions préconisées pour concevoir un tel système est un algorithme d'apprentissage sur de grandes bases de données étiquetées. Le système construit alors lui même des relations entre les caractéristiques du signal et leur valeur perceptive. Une autre approche consiste à construire un modèle analytique de l'appareil de production pour la synthèse et l'analyse de la parole, et à établir un lien direct entre la valeur des paramètres de ce modèle et la valeur perceptive du son généré par ces paramètres : on cherche alors à expliquer les relations trouvées entre perception et forme du signal.

Problématique

Dans tous les cas, il est nécessaire d'extraire un maximum d'informations des signaux de parole afin de comprendre au mieux leur comportement. Or, les méthodes existantes pour l'extraction de ces informations sont perfectibles, et quelquefois inadaptées à la grande variation des formes acoustiques de la parole expressive.

Premièrement, le signal vocal doit être segmenté pour permettre son analyse. Une nouvelle approche sur cette segmentation [Tuan et d'Alessandro, 1999] propose d'utiliser une analyse temps-fréquence pour détecter des instants d'importance sur le signal acoustique de la parole : les instants de fermeture glottique. Il s'agit donc d'explorer cette approche pour en tirer un maximum d'informations sur le signal et de comparer la localisation des instants de fermeture avec celle des méthodes courantes [Naylor *et al.*, 2007].

Deuxièmement les micro-perturbations du signal, leurs évolutions en temps (jitter) ou en amplitude (shimmer), jouent aussi un rôle perceptif prépondérant [Meike *et al.*, 2010]. Ces perturbations sont associées aux voix rauques ou rugueuses. Deux approches sont alors généralement utilisées pour mesurer ces variations : la première est la modélisation de ces perturbations au niveau spectral [Vasilakis et Stylianou, 2009], permettant d'approcher la valeur moyenne des dispersions sur le laps de temps d'observation. La seconde est basée sur la segmentation du signal vocal pour l'étude des variations d'un segment à l'autre. Dans tous les cas, si les micro-perturbations sur l'échelle temporelle sont bien estimées, celles propres à l'amplitude restent encore mal définies. Il est donc nécessaire de traiter ces questions à nouveau, et de les appliquer plus précisément au cadre de la parole expressive.

Troisièmement, il s'agit d'explorer le concept de la paramétrisation du signal vocal. Cette idée est apparue très tôt dans la problématique de l'analyse vocale. Déjà dans les années 70 [Markel et Gray, 1976, Makhoul, 1975], la prédiction linéaire a été utilisée pour estimer les paramètres d'un modèle linéaire de la production vocale. Elle est aujourd'hui à la base du codage, de la compression et de l'analyse/synthèse des signaux vocaux. Elle propose l'estimation de paramètres dans des versions plus ou moins complexes qui seront vues au chapitre 2. En particulier, ces méthodes de prédiction linéaires permettent de retrouver la forme du débit d'air issu des plis vocaux, appelé *débit glottique*. La prédiction linéaire est cependant basée sur un modèle contraignant du conduit vocal. Une approche innovante [Doval *et al.*, 2003] a permis de proposer de nouvelles méthodes d'analyse [Bozkurt *et al.*, 2005], il s'agira donc d'étudier dans quelle mesure ce nouveau modèle permet des analyses plus précises.

Quatrièmement, le signal vocal contient aussi des phénomènes stochastiques, composants aléatoires s'ajoutant au signal dit harmonique. Une forte présence de cette composante aléatoire est associée aux voix soufflées ou chuchotées. Il est donc requis de connaitre à la fois l'énergie et la forme de cette composante pour analyser les propriétés acoustiques d'un signal vocal [de Krom, 1993]. Il s'agira donc d'explorer cet aspect de la qualité vocale, notamment en proposant une méthode plus fidèle de séparation des deux composantes.

Enfin, aucun système ne permet l'estimation automatique de tous les paramètres précédemment abordés, et aucun travail ne s'est penché sur le lien entre ces différents paramètres (dans leur globalité) en regard de la qualité vocale considérée. Une fois ces quatre problèmes d'estimation abordés, il sera bon de les utiliser ensemble sur un seul et même corpus de parole expressive.

Contribution Scientifique

Cette thèse porte donc sur chacune des quatre problématiques dégagées précédemment.

1. Une méthode d'estimation des instants d'excitation glottique par analyse multi-échelles a été reprise et améliorée [Sturmel *et al.*, 2009]. Elle a été testée sur une base de données de parole naturelle. Le paradigme d'analyse utilisé a aussi permis l'observation de phénomènes liés à l'effort vocal et à la qualité vocale [d'Alessandro et Sturmel,].

2. Une nouvelle approche de la décomposition périodique/apériodique des signaux vocaux a été développée afin d'aboutir à une décomposition de meilleure qualité, notamment pendant les passages non stationnaires du signal vocal.

3. Une méthode récente de déconvolution source/filtre a été comparée aux méthodes classiques de filtrage inverse [Sturmel *et al.*, 2007]. Ses propriétés ont motivé son utilisation pour l'extraction des paramètres de source. Un algorithme d'estimation des paramètres du modèle de l'onde de débit glottique a ainsi été développé et testé [Sturmel et d'Alessandro, 2010]. Une méthode hybride a aussi été proposée pour combiner les avantages de la déconvolution et de l'analyse multi-échelles.

4. Finalement, ces méthodes ont été réunies pour être utilisées dans le cas de l'analyse d'un corpus de parole naturelle présentant différents styles d'élocution.

Méthode scientifique

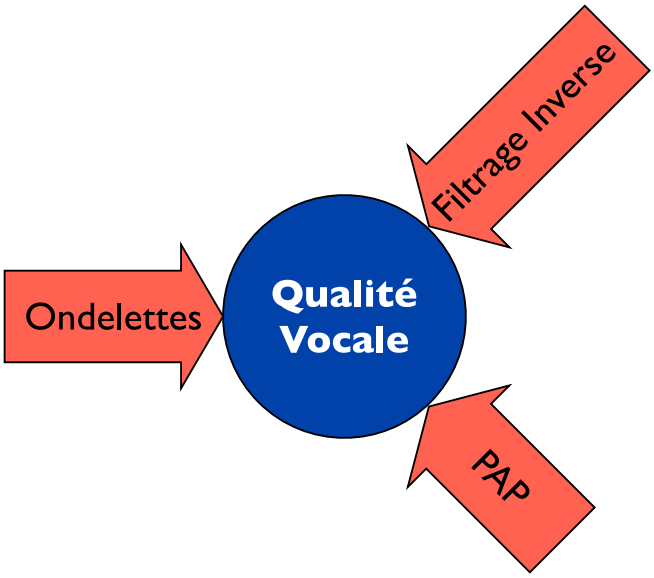
L'analyse de la parole expressive est difficile à évaluer par manque de valeurs de références. La majorité des analyses est donc faite en aveugle. En effet, les signaux qui pourraient servir de référence sont complexes à acquérir. La visualisation directe des plis vocaux par vidéo-endoscopie, par exemple, permet d'apporter des informations sur la segmentation des signaux vocaux, ou sur la configuration de l'appareil glottique mais au prix d'une instrumentation lourde. L'analyse électroglottographique (EGG) permet d'obtenir un signal approchant ces données physiologiques pour une instrumentation moindre. C'est pourquoi l'EGG est généralement considéré comme une référence suffisante. En dehors de ces données, aucune méthode d'analyse existante ne produit de données de qualité suffisante pour être considérée comme une vérité.

Pour pallier ces difficultés, l'analyse sur des signaux synthétiques sera donc utilisée au cours des chapitres de cette thèse. L'avantage des signaux synthétiques réside dans le contrôle du signal généré, donnant ainsi une connaissance parfaite de ses caractéristiques. Les résultats des méthodes étudiées peuvent donc être objectivement comparés aux caractéristiques imposées au signal synthétique. En contre partie, ces signaux synthétiques ne peuvent être qu'une simplification parfois grossière des phénomènes qui apparaissent au cours de la production d'une voix naturelle. Bien souvent, les analyses de parole naturelle ayant pour but de valider des méthodes ne peuvent que se reposer sur des modèles qui sont encore perfectibles.

A la fin de ce manuscrit, au chapitre 6, après avoir validé chaque méthode précédemment développée, des résultats seront présentés. Ces analyses complètes sur des signaux de parole expressive seront présentées dans le but d'établir les liens entre les valeurs des paramètres et les qualités vocales.

Organisation du document

Ce document est organisé en 3 parties. La première partie introduira les modèles et l'état de l'art sur les problèmes qui nous intéressent. Le deuxième partie présentera les contributions scientifiques en matière d'analyse du signal de parole. Afin de suivre une logique de progression dans le traitement des signaux, le premier chapitre de cette partie (chapitre 3) traitera de la segmentation des signaux vocaux par les ondelettes, et des informations qu'il est possible d'extraire de la représentation temps/fréquence de ces signaux. Une fois la segmentation effectuée, il s'agira de décomposer les signaux en parties périodique et apériodique (PAP, chapitre 4), il sera aussi vu dans quelle mesure la décomposition périodique/apériodique conserve ou non les apériodicités structurelles. Une fois connus la partie voisée du signal et ses instants d'excitation glottique, le dernier chapitre de la partie (chapitre 5) traitera de l'estimation des paramètres de l'onde de débit glottique par filtrage inverse. Au cours de cette deuxième partie, plusieurs approches seront donc présentées pour traiter une partie du problème de l'estimation de la qualité vocale :



Enfin, la troisième partie de cette thèse traitera de l'application des techniques d'estimation proposées sur un corpus de parole naturelle au chapitre 6, les différentes approches présentées dans la deuxième partie seront fusionnées pour proposer un système complet d'analyse. Le dernier chapitre conclura sur le travail effectué au cours de cette thèse et donnera ses perspectives.

Acronymes, Abréviations

Acronyme	Page de définition	Description succinte
ODG	26	Onde de Débit Glottique
EGG	30	Electroglottographie
DEGG	30	Dérivée du signal EGG
$\delta(t)$	31	fonction de Dirac
F_0	31	Fréquence fondamentale d'un signal
DODG	31	Dérivée de l'Onde de Débit Glottique
LF	32	Modèle LF du débit glottique
$argmax$	33	Argument du maximum, donnant $[y \| \forall x, f(y) \geq f(x)]$
ω_g	33	Pulsation de la modulation du modèle LF
F_g	33	Fréquence du formant glottique
O_q	33	Quotient ouvert
α_m	35	Quotient d'asymétrie
Q_a	35	Quotient de retour
E	36	Amplitude du modèle LF
ARMA	38	AutoRegressive Moving Average
GCI	47	Instant de fermeture glottique
MSP	51	Multi Scale Product
IAIF	55	Filtrage Inverse Itératif Adaptatif
CALM	56	Causal Anti-causal Linear Model
S^*	56	Conjugué du spectre S
ARX	66	AutoRegressive eXogenous
ARX-LF	66	modélisation ARX utilisant un modèle LF de débit glottique
NAQ	67	Quotient d'amplitude normalisé
bin	73	Element d'une transformée de Fourier discrète
TF ou FT	83	Transformée de Fourier
GOI	107	Instant d'ouverture glottique
MSPh	112	Mean Square Phase
RSB	124	Rapport Signal sur Bruit
RMS	125	Root Mean Square
RVB	140	Rapport Voisement sur Bruit
JND	151	seuil différentiel perceptif
API	176	Alphabet Phonétique International
HNR	182	Rapport harmonique sur bruit

Première partie

Modélisation et État de l'art

Chapitre 1

Modèle de la production vocale

Sommaire

1.1 Le signal vocal : production et modélisation 19
1.1.1 Du larynx à la bouche 19
1.1.2 Modélisation de la production vocale 20
1.2 Les différentes échelles du signal vocal 21
1.2.1 Structure temporelle 21
Macro échelle temporelle : l'énoncé 21
Briques linguistiques temporelles : les phonèmes 22
Micro échelle temporelle : la période du signal 22
1.2.2 Structure fréquentielle 22
Une structure harmonique 22
Une structure aléatoire 22
1.3 Le cycle glottique et sa caractérisation 23
Le masque de Rothenberg 23
Imagerie ultra-rapide 24
L'électroglottographie (EGG) 24
1.4 Le modèle du point de vue signal 24
1.4.1 Modélisation de la source glottique. 25
Modèles de la littérature 26
Choix du modèle 26
Le modèle LF plus en détails 26
Le quotient ouvert O_q 27
Le quotient d'asymétrie 28
Le quotient de retour Q_a 29
Le paramètre d'amplitude (E) 29
1.4.2 Modélisation du filtre 30
Formants et voyelles 30
Le triangle vocalique 31
Le filtre autorégressif 31

Les consonnes 32
1.5 Qualités vocales 32
1.5.1 La dimension serré / relâché 33
1.5.2 La dimension d'effort vocal 33
1.5.3 La dimension de raucité 34
1.5.4 La dimension de voisement 35
1.5.5 Récapitulatif 35
1.6 Conclusion 35

Ce chapitre expose les bases de la production vocale du point de vue articulatoire. Après avoir présenté les organes responsables de la production de la parole, puis les avoir caractérisés selon leurs fonctions de production, le fonctionnement des plis vocaux sera décrit plus en détails. Différentes méthodes d'observation et de mesure de leur comportement seront décrites. Un modèle linéaire de la production vocale sera présenté, basé sur les travaux antérieurs et proposant plusieurs critères de forme du débit glottique. Ce modèle de production sera expliqué en détails, notamment pour ses multiples variantes (un à trois critères de forme). Finalement, la notion de qualité vocale sera introduite, ainsi que les liens qu'elle entretient avec la configuration de l'appareil de production vocale.

1.1 Le signal vocal : production et modélisation

1.1.1 Du larynx à la bouche

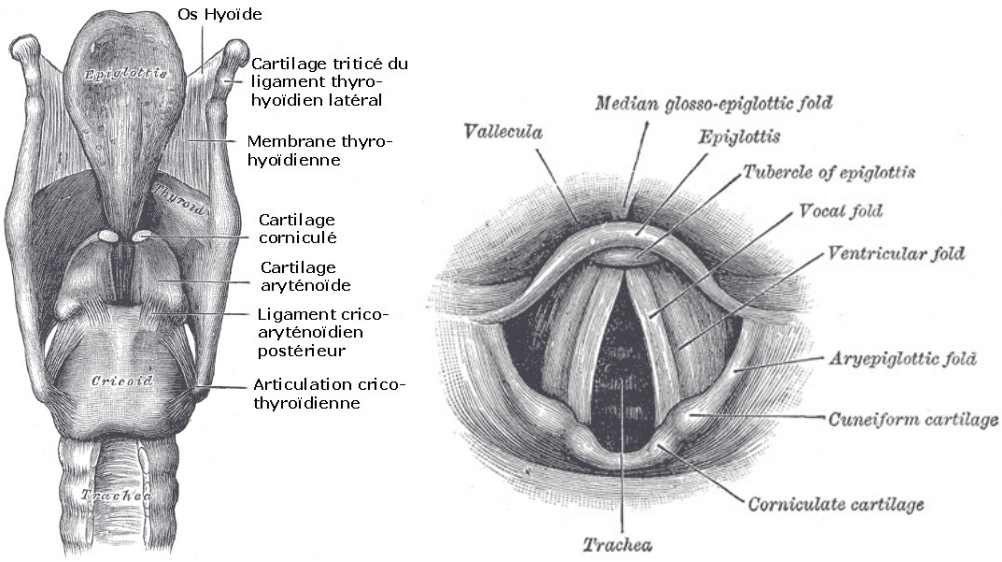


Figure 1.1 – Vue des éléments du larynx. Issu de *Gray's Anatomy*, 20ème édition (1918).

Produire de la parole, c'est tout d'abord expulser de l'air des poumons ; ce flux d'air va être mis en forme par la multitude de cavités, constrictions et orifices qui jalonneront son chemin depuis les poumons jusqu'aux lèvres et au nez. C'est au niveau du larynx que s'opère la première mise en forme de ce flux d'air. L'espace présent entre les deux masses de chair présentées sur la figure 1.1, les plis vocaux (anciennement appelés cordes vocales), est appelé glotte et on appelle alors le flux d'air passant par la glotte le *flux glottique*. Cette glotte oppose une résistance au flux d'air parcourant la trachée. Les plis vocaux vont donc vibrer en accord avec la configuration imposée par les muscles avoisinants. Plus ils seront tendus et plus la fréquence générée sera élevée à la manière d'un ballon de baudruche qu'on ferait vibrer en en pinçant le bout entre ses doigts.

Une fois ce débit glottique mis en forme par la glotte, il est filtré par différentes cavités au niveau de la gorge, de la bouche, du nez et des sinus. L'ensemble du trajet de la glotte jusqu'aux lèvres et au nez est appelé *le conduit vocal*. La configuration de ce conduit vocal va conditionner le filtrage appliqué à l'onde de débit glottique. C'est ainsi que notre voix change lorsque nous avons le nez bouché, car la configuration des cavités nasales est modifié.

Il est important de noter que ce modèle de production considère que la configuration du

conduit vocal n'a pas d'impact sur le signal de source - d'interaction entre la source et le filtre. Ce modèle est considéré linéaire. En pratique, cette interaction existe, mais la présence d'interactions source/filtre n'empêche pas l'analyse par un modèle linéaire s'il n'est pas trop contraint. En effet, si l'interaction dépend de la configuration du conduit vocal, les résultats obtenus par analyse linéaire devront être analysés en connaissance de la voyelle prononcée dans la mesure où l'interaction (si elle existe) est dépendante du conduit vocal.

1.1.2 Modélisation de la production vocale

On peut séparer l'appareil de production vocale en trois parties distinctes visibles sur la figure 1.2 :

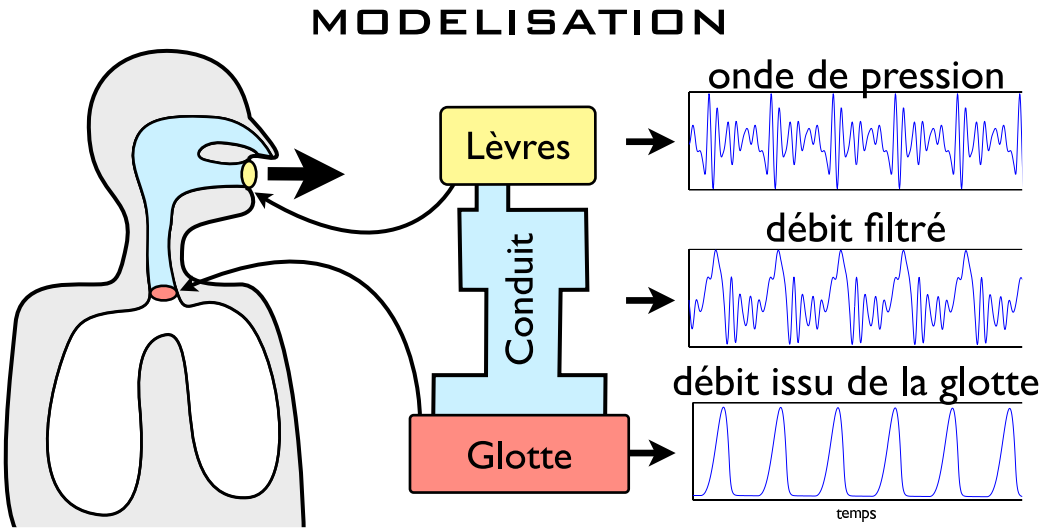


Figure 1.2 – Modélisation de la production vocale comme une succession de filtres linéaires. En parallèle est donné un modèle source/filtre comme celui utilisé en prédiction linéaire.

- La première partie est responsable de la production du flux d'air qui va servir de support à la voix. Composée des poumons et du larynx, cette partie s'arrête au niveau des plis vocaux d'où est issue l'onde de débit glottique (ODG).
- La deuxième partie, appelée le "conduit vocal", fonctionne comme une succession de guides d'ondes qui vont agir sur l'onde de débit glottique. Les cavités nasales jouent aussi un rôle à ce niveau, mais dépendant du phonème. Ainsi, on retrouve certaines voyelles dites "nasales" en français (le /ɔ̃/ de 'son' ou le /ɛ̃/ de 'pain'). L'effet des cavités nasales est plus marginal et n'est pas pris en compte par toutes les modélisations.
- La troisième partie, comportant uniquement l'ouverture au niveau des lèvres et du nez va transformer l'onde de débit en onde de pression acoustique. Cette transformation est assimilée, par simplification, à une dérivation dans le domaine signal.

La dernière partie a un effet important sur la manière dont nous mesurons le signal acoustique vocal. En effet, alors que c'est un débit qui est produit par les plis vocaux, nous ne mesurons généralement que l'onde de pression qui résulte de la diffusion par les lèvres. Il est donc habituel de voir une représentation faite uniquement à base d'ondes de pression, qui se traduit par la visualisation de l'onde de dérivée du flux glottique plutôt que par le flux lui-même. Cet artifice est permis par la linéarité du modèle utilisé.

Si des travaux récents permettent une représentation poussée de la configuration du conduit vocal d'après l'analyse des signaux de paroles [Laprie et Mathieu, 1998], le débit glottique reste encore très difficile à estimer. Il est donc nécessaire de comprendre comment les plis vocaux agissent sur le flux issu des poumons pour le mettre en forme. Avant cela, intéressons-nous au signal acoustique capté par un micro.

1.2 Les différentes échelles du signal vocal

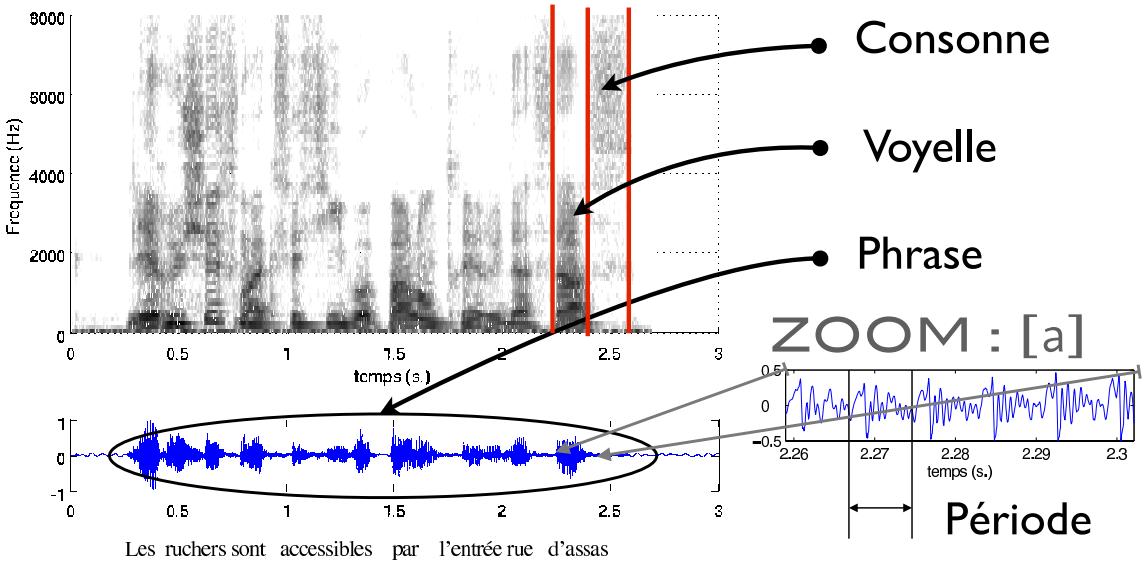


Figure 1.3 – Décomposition d'un signal de parole en temps au niveau de la phrase et de la période, et en fréquence.

1.2.1 Structure temporelle

La première direction d'analyse du signal vocal est la direction temporelle. En effet, beaucoup d'informations se développent au cours du temps.

Macro échelle temporelle : l'énoncé

Un premier découpage de la progression temporelle de la voix peut se faire au niveau de la phrase, comme montré sur la figure 1.3. Cette entité d'analyse permet d'extraire des informations lexicales mais aussi prosodiques, c'est à dire : sur les variations rythmiques, intonatives et d'intensité réalisées par le locuteur qui prononce la phrase. A ce niveau, il n'y a pas de description de la configuration de l'appareil vocal.

Briques linguistiques temporelles : les phonèmes

On décompose l'articulation d'une phrase en une succession de phonèmes. Le phonème est une entité abstraite qui correspond à la plus petite unité distinctive dans la parole. Leur analyse permet d'extraire les informations linguistiques de la phrase. La reconnaissance des phonèmes réalisée en traitement automatique du langage naturel nécessite une analyse poussée du signal qui se concentre sur l'évolution de la conformation du conduit vocal et laisse de coté beaucoup d'informations sur la source glottique.

Micro échelle temporelle : la période du signal

Le zoom de la figure 1.3 montre le son /a/. Il est composé de formes répétées dans le temps : les périodes. Le signal vocalique n'est pas strictement périodique, mais beaucoup d'analyses font une hypothèse de stationnarité du signal sur un temps d'environ 20ms. Cette hypothèse permet l'analyse des signaux vocaux en les considérant comme périodiques. L'intervalle de temps minimum nécessaire à la reproduction de la forme du signal est appelé période du signal, noté T_0. L'inverse de cette période $F_0 = \frac{1}{T_0}$ est la fréquence fondamentale du signal, exprimée en Hz.

L'analyse de la période du signal vocal ne tient aucunement compte des informations lexicales et se concentre essentiellement sur l'analyse du signal par des modèles physiques et/ou mathématiques. C'est à cette échelle temporelle d'analyse de la parole qu'est basé le travail présenté dans cette thèse. En retour, une telle analyse peut apporter des informations en dérivant des variations prosodiques sur la qualité vocale [Pfitzinger, 2006].

1.2.2 Structure fréquentielle

Les signaux peuvent se représenter dans deux espaces différents, un espace temporel et un espace fréquentiel. L'analyse de Fourier propose une représentation des signaux non pas du point de vue de leur évolution temporelle, mais de leur contenu en terme fréquentiel. La figure 1.3 donne une représentation temps/fréquence du signal, le spectrogramme. A chaque instant temporel, on visualise le contenu en fréquence (plus sombre = d'avantage d'énergie) du signal. Comme pour beaucoup de signaux acoustiques, on trouve dans la voix deux caractéristiques principales de ces signaux : une structure harmonique et une structure aléatoire.

Une structure harmonique

Les signaux vocaux, supposés périodiques en première approximation, présentent des maxima locaux réguliers en fréquence. Ces maxima sont présents aux fréquences multiples de la fréquence fondamentale du signal : les harmoniques. Elles comportent toute l'information nécessaire à la reconstruction du signal déterministe.

Une structure aléatoire

L'hypothèse de stationnarité du signal vocal sur une fenêtre d'environ 20ms reste une hypothèse, et une composante aléatoire est toujours présente. Cette composante est la seule présente lorsque le locuteur produit des sons non voisés, comme la consonne de la figure 1.3. Les sons dits non voisés ne présentent pas de structure harmonique, c'est le cas des consonnes fricatives comme /s/ par exemple.

Le problème réside dans le cas de sons dits "mixtes" ou une forme périodique et une forme aléatoire se superposent.

1.3 Le cycle glottique et sa caractérisation

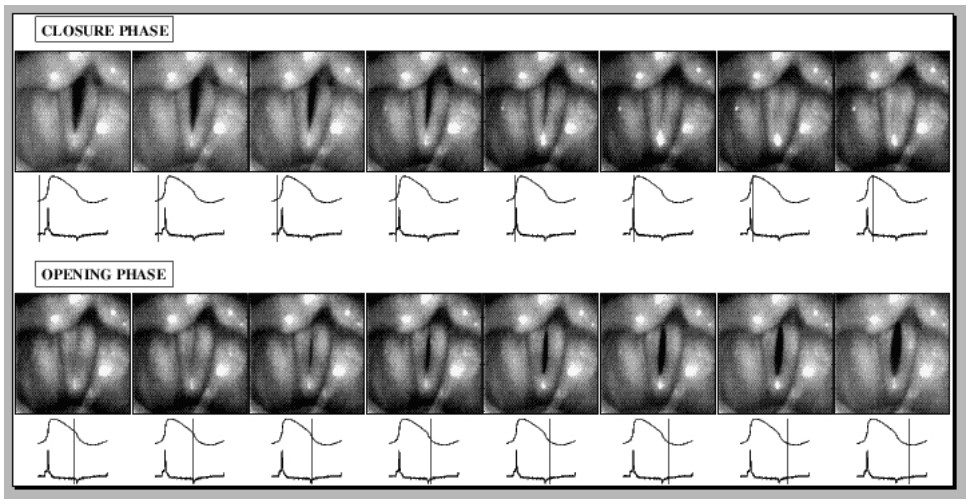


Figure 1.4 – Représentation du cycle glottique par vidéo ultra rapide, électroglottographie (EGG) et dérivée de l'EGG, extrait de [Henrich *et al.*, 2004].

Afin de chercher à modéliser plus finement la source de la production vocale, il faut préalablement comprendre le fonctionnement des plis vocaux. En reprenant la précédente métaphore du ballon de baudruche, on remarque que si le son change en fonction de la configuration, le mouvement et la dynamique des plis vocaux changent en conséquence. Ce sont donc les muscles et articulations (dont les cartilages aryténoïdes) du larynx qui vont conditionner la qualité de la production vocale : grave, aiguë, stridente, ample, etc...

Le mouvement des plis vocaux est quasi périodique. Le cycle de leur mouvement est décomposable en deux phases. Des études sur le débit d'air expulsé lors de la phonation [Rothenberg, 1977, Rothenberg, 1973, Gauffin et Sundberg, 1989] ont montré que ces deux phases présentaient principalement une différence dans le niveau de débit d'air expulsé. Durant la phase dite "fermée", les plis vocaux sont rapprochés au maximum et donc le débit est minimum, tandis que pendant la phase dite "ouverte" le débit d'air atteint son maximum, correspondant à l'espace maximal au niveau de la glotte.

Dans certains cas, la fermeture n'est pas complète et un débit d'air résiduel peut être présent lors de la phase fermée. Le fait que l'attention soit portée plus souvent sur le gradient du débit du flux glottique que sur le flux lui même rend ce détail insignifiant ; le débit résiduel étant bien souvent constant, il sera juste annulé par la dérivation. Il sera vu plus tard que lorsque ce débit résiduel cause des turbulences (par exemple, dans le cas du bruit d'aspiration), sa contribution est incluse dans la partie non périodique du signal vocal.

Plusieurs types d'instruments de mesure permettent de visualiser, représenter ou analyser ce cycle glottique : le masque de Rothenberg, l'imagerie ultra-rapide et l'éléctroglottographe.

Le masque de Rothenberg

La mesure du débit d'air expiré est réalisée à l'aide d'un masque pneumotachométrique (dit masque de Rothenberg [Rothenberg, 1973] visible sur la figure 1.5a). C'est un masque facial percé de trous équipés de grilles fines métalliques, que le sujet applique de manière étanche sur son visage. Le débit d'air est évalué par mesure de la chute de pression entre l'intérieur et l'extérieur

a) b) c)

Figure 1.5 – A) Un masque de Rothenberg équipé d'une mesure de pression intraorale (source : site web de l'INPG). B) Une visualisation de la glotte par endoscopie. C) Une locutrice instrumentée pour une acquisition EGG (source : site web du LIMSI/CNRS).

de la grille. Une mesure de pression intraorale peut être effectuée par adjonction d'un petit tube de polyéthylène qui s'insère à travers les lèvres et est relié à un capteur de pression basse fréquence. Ce système a cependant quelques limites qui le rendent inutilisable pour l'analyse de la voix expressive ; en particulier, il ne permet pas de décrire le flux glottique au delà de 1kHz [Badin *et al.*, 1990].

Imagerie ultra-rapide

L'imagerie ultra-rapide [Kendall, 2009], qui consiste en une caméra haute vitesse (4000 à 16000 images par seconde) filmant la glotte à l'aide d'un système endoscopique rigide. Ce système permet une visualisation très précise, mais provoque une forte gène au locuteur à cause de l'encombrement du système dans la gorge. De plus, la forte luminosité nécessaire provoque des échauffements dans la gorge du sujet et empêche les acquisitions sur de longues durées. Un exemple d'images obtenues par ce procédé est présenté sur la figure 1.4 ; de même, une mise en œuvre de ce mode de visualisation est présenté sur la figure 1.5b. Ces signaux obtenus par imagerie ultra-rapide présentent de nombreuses difficultés d'analyse automatique, mais permettent d'extraire des caractéristiques intéressantes sur la configuration glottique [Karakozoglou *et al.*, 2010].

L'électroglottographie (EGG)

L'électroglottographie consiste en un appareil de mesure de l'admittance électrique au niveau du larynx [Fabre, 1957]. Deux électrodes sont placées sur le cou du locuteur au niveau de la glotte (comme présenté sur la figure 1.5c) et mesurent l'admittance qui varie en fonction de l'éloignement des plis vocaux, entre autres. Ce signal est plus difficilement exploitable que des signaux vidéos mais a l'avantage de nécessiter une instrumentation beaucoup plus légère. Les signaux obtenus sont présentés sur la figure 1.4. L'EGG est le signal supérieur, sa dérivée (DEGG) est le signal inférieur.

Si l'avantage va à la facilité de mise en œuvre, le fait d'avoir une mesure sur une seule dimension ne permet pas d'avoir une confiance aveugle dans les résultats. La présence de mucus à l'intérieur du larynx, par exemple, peut modifier les courbes obtenues sur EGG et donc les mesures qui en découlent. La mesure est, en outre, sensible au positionnement des capteurs : des méthodes EGG à 3 dimensions ont été développées [Rothenberg, 1992] mais restent peu courantes.

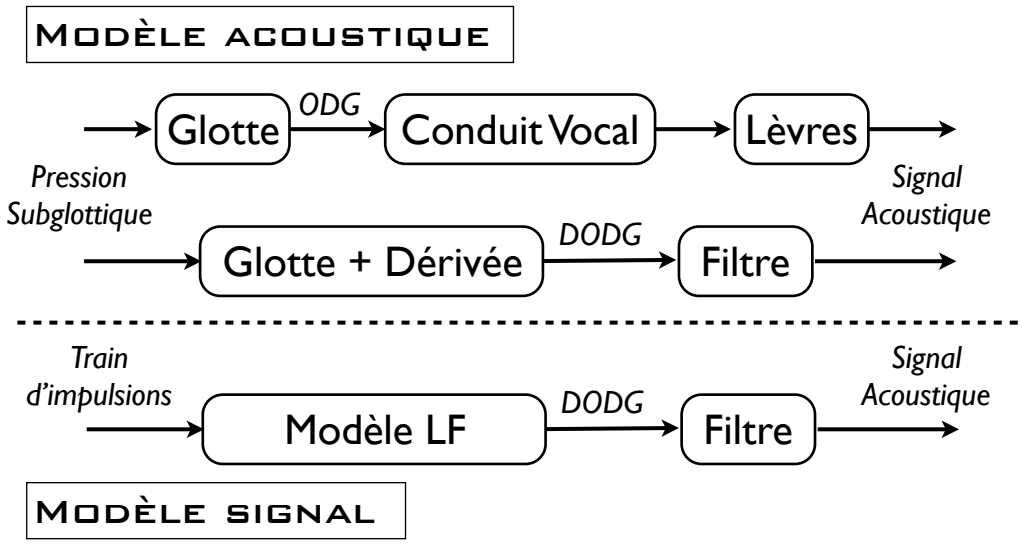


Figure 1.6 – Passage du modèle acoustique au modèle signal de la production vocale. Bien souvent, on associe la dérivation due au rayonnement du débit à l'onde de débit glottique (ODG). On considère donc généralement le modèle de la dérivée de l'onde de débit glottique (DODG) directement.

1.4 Le modèle du point de vue signal

Cela a été vu précédemment, la forme d'onde du débit glottique est une forme d'onde périodique. On peut donc considérer la source comme le résultat d'un filtre ayant pour réponse impulsionnelle une période $g(t)$ excitée par un train d'impulsions de période $\frac{1}{F_0}$: $\delta_{T_0} = \sum_{i=-\infty}^{\infty} \delta(t - \frac{i}{F_0})$ et filtré par le conduit vocal, comme montré sur l'équation 1.1.

$$s(t) = f(t) * [g(t) * \sum_{i=-\infty}^{\infty} (1 + sh(i))\delta(t - (1 + jitt(i))\frac{i}{F_0})] \tag{1.1}$$

Le train d'impulsions n'est en réalité pas constant, il comporte des apériodicités structurelles : modification en amplitude et en temps du train d'impulsion. On définit alors le shimmer sh comme la variation d'amplitude d'une période à une autre (en %) et le jitter $jitt$ comme la variation de fréquence d'une période à une autre (en %). Le train d'impulsion est donc porteur des informations d'énergie, de la fréquence fondamentale (F_0) et de sa micro-variation. Le filtre $g(t)$ transformant le train d'impulsion en onde de débit glottique porte quant à lui des informations de qualité vocale qui vont permettre de donner, pour une même fréquence fondamentale, des textures de voix différentes. Ensuite, l'onde de débit glottique passant dans le conduit vocal est filtrée par la réponse impulsionnelle $f(t)$ de ce dernier. On définit donc un deuxième filtre, portant les informations propres à la configuration articulatoire du conduit vocal. La figure 1.6 résume ce parallèle entre articulation et modélisation. Pour simplifier, on assimile souvent la dérivation et l'onde de débit glottique pour représenter et analyser la dérivée de l'onde de débit glottique (DODG).

Ainsi, en fonction des informations que l'on souhaitera extraire du signal de parole, on cherchera à estimer plutôt l'une ou l'autre des composantes spécifiées.

1.4.1 Modélisation de la source glottique.

La phase d'expérimentation qui a permis de décrire les différentes phases du cycle glottique permet l'établissement d'un modèle plus précis. La transcription de ce modèle sous forme de fonction mathématique autorise l'utilisation des outils de traitement du signal, et notamment son analyse spectrale [Doval *et al.*, 2006].

Modèles de la littérature

Des modèles ont été proposés pour tenter de décrire la forme d'onde du débit glottique par Rosenberg [Rosenberg, 1971] (Rosenberg C), Liljencrants et Fant [Fant *et al.*, 1985, Fant, 1995] (LF), Klatt [Klatt et Klatt, 1990] (KGLOTT88) et Veldhuis [Veldhuis, 1998] (R++) notamment. Tous ces modèles présentent des caractéristiques identiques à la forme de la figure 1.7, mais une paramétrisation différente.

- Rosenberg C est un modèle composé de plusieurs sinusoïdes avec toujours 4 paramètres dont 2 paramètres de formes, la fréquence fondamentale et l'amplitude de voisement.
- Le modèle LF présente 3 paramètres de forme en plus de la fréquence fondamentale et de l'amplitude de voisement, mais il est basé sur une expression analytique assez complexe avec notamment des équations implicites pour la détermination des paramètres.
- KGLOTT88 est un modèle qui a pour expression un polynôme cubique $(2at - 3bt^2)$ où a et b définissent la forme de l'onde. Ce modèle présente donc deux paramètres de forme, avec une asymétrie du débit fixée à $\frac{2}{3}$. A ces paramètres de forme on ajoute la fréquence fondamentale et l'amplitude de voisement (paramètre d'échelle).
- R++ présente cette fois 3 paramètres de forme en plus de la fréquence fondamentale et de l'amplitude de voisement. Basé sur le même jeu de paramètres que LF, il permet un calcul plus rapide.

Choix du modèle

Dans le travail de cette thèse, c'est le modèle de Liljencrants et Fant (LF) qui a été retenu, il présente 4 paramètres s'ajoutant à la fréquence fondamentale, permettant de contrôler précisément la forme de la source. Ce modèle présente plusieurs avantages par rapport aux autres, notamment :

- Il est explicitement formulé comme une sinusoïde modulée par une exponentielle suivie d'une exponentielle amortie (absolument décroissante). Cette séparation en deux phases (équation 1.2) permet des analyses et des prédictions spectrales pertinentes tout en s'approchant au mieux du comportement physiologique des plis vocaux et du débit qu'ils mettent en forme. Par ailleurs, ces deux phases sont liées par une équation implicite.
- Il est largement utilisé dans la littérature, notamment les travaux récents de [Vincent *et al.*, 2007, Degottex *et al.*, 2010, Drugman *et al.*, 2008].
- On peut lier les paramètres du modèle aux différentes dimensions de la qualité vocale. Ce point sera explicitement abordé en fin de chapitre.
- Les travaux de Boris Doval [Doval *et al.*, 2006] ont notamment montré que tous les modèles étaient équivalent du point de vu de leur propriétés. On décrira donc le modèle LF en utilisant des paramètres normalisés qui peuvent être définis pour les autres modèles.

Le modèle LF plus en détails

L'équation 1.2 décrit la forme de la dérivée du débit glottique $\frac{dg}{dt}(t)$ selon la figure 2 de [Fant *et al.*, 1985].

$$\frac{dg}{dt}(t) = \begin{cases} E_0 sin(\omega_g t)e^{\alpha t} & 0 < t < T_e \\ -\frac{E_0}{\epsilon}[e^{-\epsilon(t-T_e)} - e^{-\epsilon(T_a-T_e)}] & T_e < t < T_e + T_a \end{cases} \tag{1.2}$$

Les quantités ϵ et α sont déterminées en fonction des paramètres du modèle et d'une équation implicite : $\int_0^{T_0} \frac{dg}{dt}(t)dt = 0$. Afin d'utiliser des paramètres normalisés et de les exprimer en fonction de leur effet sur le spectre de la source, nous utilisons les paramètres de forme O_q, α_m, Q_a décrits à la figure 1.7 et explicités par la suite. L'expression du modèle reste identique, mais ces paramètres permettent une analyse plus simple des résultats.

Le spectre de la dérivée de l'onde de débit glottique présente un maximum ayant la même caractéristique qu'un formant du filtre vocalique. On l'appelle le formant glottique et on définit sa fréquence F_g telle que

$$F_g = \underset{\nu}{\operatorname{argmax}} |FT[\frac{dg}{dt}](\nu)|$$

.

Il est important de remarquer que tenant compte des notations, ω_g n'est pas égal à $2\pi F_g$. En fonction de l'amortissement α de l'exponentielle du modèle 1.2, le maximum du spectre n'est pas nécessairement placé sur cette fréquence de modulation. En effet, calculons le spectre $FT[\frac{dg}{dt}]$ de la dérivée du débit glottique, et son terme correspondant à la transformée du terme exponentiel de la partie ouverte $G_{exp}(\omega)$. On constate une interaction entre les parties (a) et (b) de l'équation 1.3, d'autant plus importante que l'atténuation est faible : pour $\omega = \omega_g$ on se retrouve avec $FT[\frac{dg}{dt}(\omega_g)] = \frac{1}{2i}[G_{exp}(0) - G_{exp}(2\omega_g)]$. En fonction de la bande passante du spectre G_{exp}, il se peut que la pulsation ω_g ne soit pas le maximum.

$$G_{exp}(\omega) = \frac{e^{(\omega i - a)T_e} - 1}{\omega i - a}$$

$$FT[\frac{dg}{dt}](\omega) = \frac{1}{2i}[\delta(\omega - \omega_g) - \delta(\omega + \omega_g)] * G_{exp}(\omega)$$

$$FT[\frac{dg}{dt}](\omega) = \frac{1}{2i}[G_{exp}(\omega - \omega_g) - G_{exp}(\omega + \omega_g)]$$

$$G_{exp}(\omega - \omega_g) \text{ (a)} \qquad G_{exp}(\omega + \omega_g) \text{ (b)} \tag{1.3}$$

Ainsi, on peut en déduire que $\omega_g \approx 2\pi F_g$ uniquement si la pente spectrale imposée par cette exponentielle est importante, c'est à dire pour des faibles valeurs de α.

Le modèle LF propose la substitution de chaque impulsion de la source par une forme d'onde en deux parties et 4 paramètres. Les paramètres peuvent être définis en unité de temps ou par rapport à la période fondamentale : le modèle et ces deux jeux de paramètres sont présentés sur la figure 1.7. Voyons quels sont leurs effets respectifs sur la forme d'onde et le spectre de la source d'après [Doval *et al.*, 2006].

Le quotient ouvert O_q

Le quotient ouvert est le rapport entre la durée de la phase d'ouverture T_e et la période du cycle glottique T_0. De nombreux travaux, dont récemment Henrich [Henrich, 2001] et Sundberg

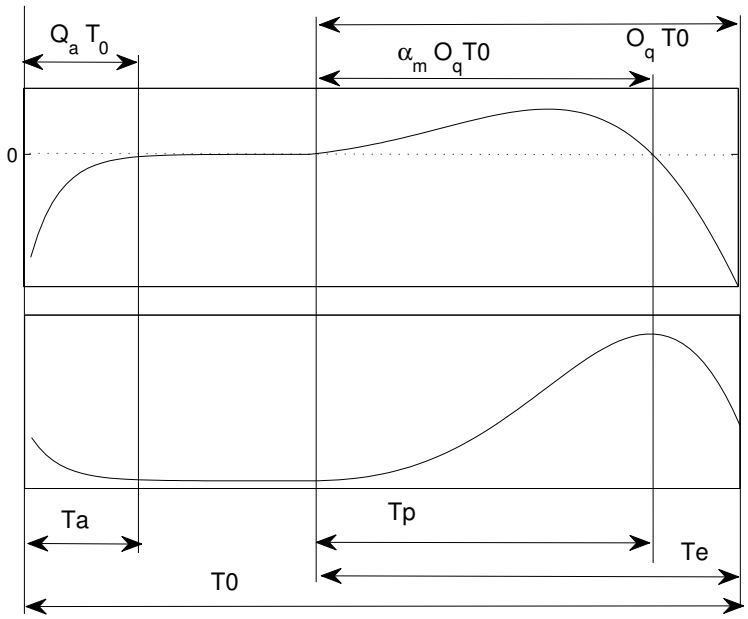


Figure 1.7 – Le modèle LF [Fant *et al.*, 1985] et ses paramètres. En haut les paramètres normalisés, en bas les paramètres temporels. Forme du haut : DODG, forme du bas : ODG.

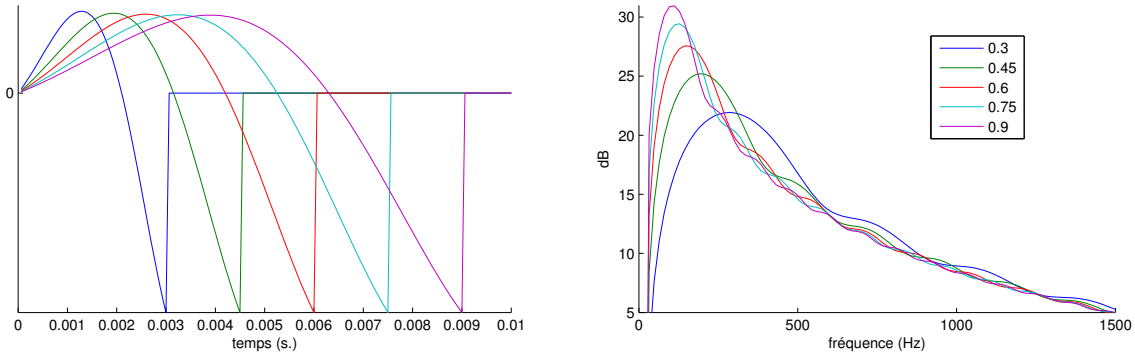


Figure 1.8 – Effet de la variation du quotient ouvert sur la forme d'onde dérivée du débit glottique. (inspiré de [Doval *et al.*, 2006]).

[Sundberg, 1994] ont montré que sa valeur estimée ou mesurée était fortement corrélée avec la notion de qualité vocale. O_q est donc l'un des paramètres majeurs à estimer pour qualifier la source glottique.

En dehors de la technique de filtrage inverse et d'estimation, il existe des méthodes de mesure directe de O_q. En effet, comme le quotient ouvert correspond à des événements articulatoires intervenants au niveau des plis vocaux, on peut le mesurer directement par différentes méthodes : vidéo ultrarapide ou EGG. Comme montré sur la figure 1.4, la visualisation du cycle glottique permet de mettre en lumière les instants de fermeture et d'ouverture sur la vidéo comme sur l'EGG, où ces instants correspondent à des extrema du signal DEGG.

Sur la figure 1.8 on représente les dérivées d'ondes de débit glottique, ainsi que les spectres associés pour différentes valeurs de O_q. On remarque que les variations de quotient ouvert ont pour principal effet le déplacement du maximum sur le spectre de la dérivée du flux glottique.

Ce maximum, le formant glottique, est positionné sur la fréquence F_g.

Le quotient d'asymétrie

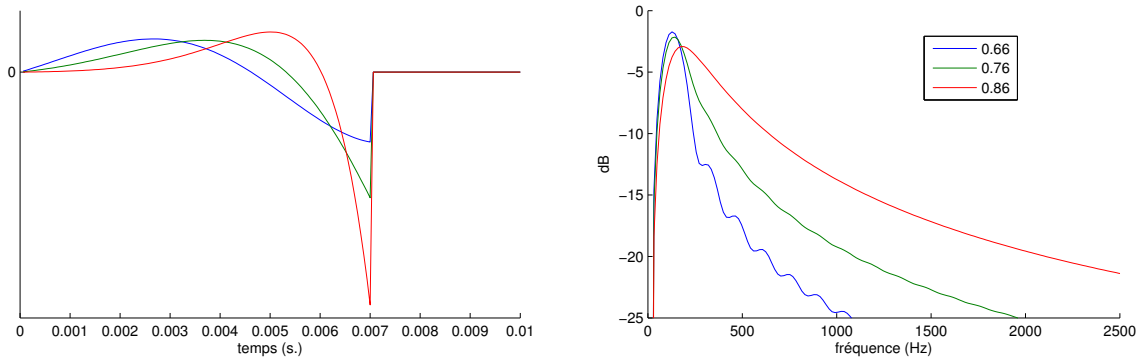


Figure 1.9 – Effet de la variation du quotient d'asymétrie de la forme d'onde dérivée du débit glottique (inspiré de [Doval *et al.*, 2006]).

La quotient d'asymétrie ($\alpha_m = \frac{T_p}{T_e}$) est, comme son nom l'indique, le paramètre qui détermine l'asymétrie de l'onde de débit glottique. Ce coefficient est le rapport entre les temps T_p et T_e de la figure 1.7. Pour être cohérent avec la réalité physique du modèle, on pose la contrainte $0.5 < \alpha_m < 1$, le cas limite $\alpha_m = 0.5$ correspond à un débit glottique sinusoïdal, et $\alpha_m = 1$ correspond à une impulsion parfaite. On notera, comme précisé dans [Fant *et al.*, 1985] que $T_p = \frac{\pi}{\omega_g}$.

Sur la figure 1.9 on représente les dérivées d'ondes de débit glottique, ainsi que les spectres associés pour différentes valeurs de α_m. On remarque sur les spectres que les variations du coefficient d'asymétrie ont pour principale conséquence la modification de la largeur de bande du formant glottique. Plus la largeur de bande est large, et plus le formant glottique est déplacé en haute fréquence par rapport à la pulsation ω_g.

Le quotient de retour Q_a

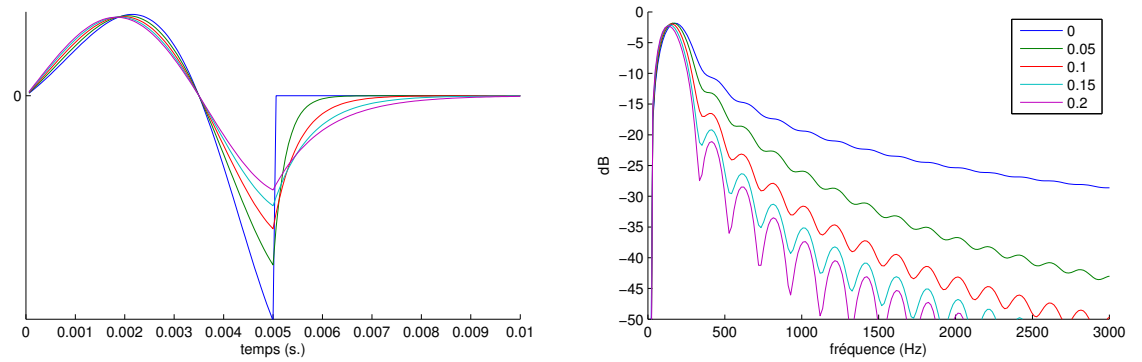


Figure 1.10 – Effet de la variation du quotient de retour sur la forme d'onde dérivée du débit glottique (inspiré de [Doval *et al.*, 2006]).

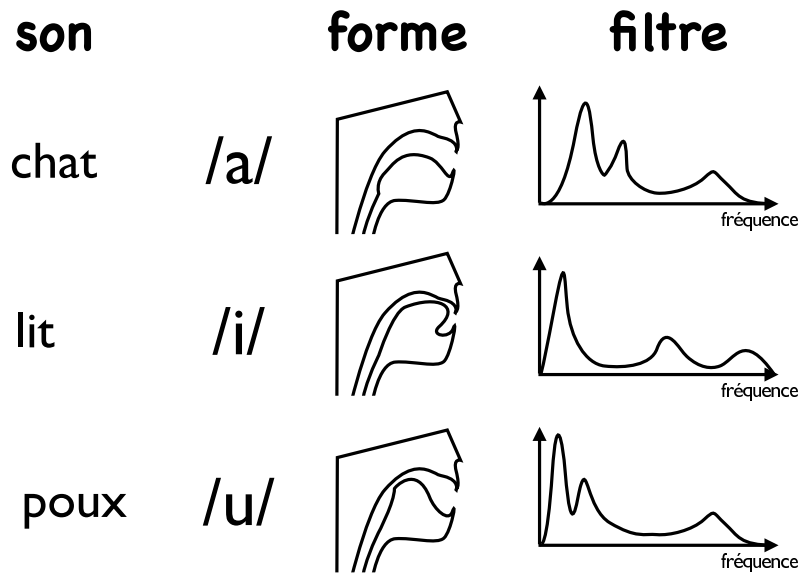


Figure 1.11 – Position de la langue et des résonateurs du filtre pour 3 voyelles différentes.

Le quotient de retour est un paramètre qui influe sur la phase de retour du débit glottique, comprise entre T_e et T_0. Cependant, dû à l'équation implicite de la dérivée du débit glottique - moyenne nulle, car dérivée d'un signal périodique -, ce paramètre a aussi un effet sur la forme de la phase ouverte de l'onde de débit glottique. On peut donc voir ce paramètre comme le contrôle d'un filtre passe-bas de premier ordre agissant sur l'onde de débit glottique sans phase de retour comme cela a été présenté dans [Klatt et Klatt, 1990] ou plus récemment pour le modèle causal-anticausal présenté au chapitre suivant [Doval *et al.*, 2003].

Sur la figure 1.10 on représente les dérivées d'ondes de débit glottique ainsi que les spectres associés pour différentes valeurs de Q_a. On constate l'effet passe-bas du premier ordre de ce paramètre, où sa valeur contrôle la fréquence de coupure et induit donc une pente spectrale plus ou moins marquée (et quelques ondulations par effet de fenêtrage).

Le paramètre d'amplitude (E)

Alors que les 3 paramètres précédents avaient une définition temporelle, le paramètre d'amplitude E est un coefficient multiplicateur de l'amplitude de l'onde. Du point du vue spectral, il n'a donc que peu d'intérêt mais un tel paramètre joue de manière prépondérante sur l'énergie du signal vocal.

1.4.2 Modélisation du filtre

Dans la présente étude, les formes et variations du filtre vocalique ne seront pas abordées en tant que descripteurs de la qualité vocale. Ils jouent cependant un rôle prépondérant dans le mécanisme de production vocale.

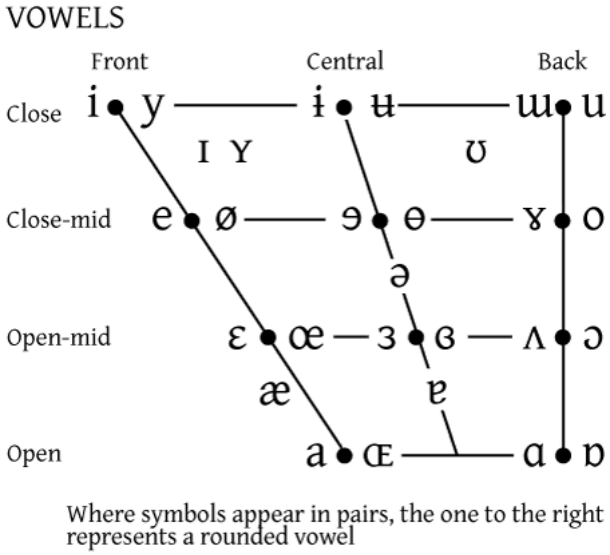


Figure 1.12 – Placement de voyelles par leur lieu d'articulation. (alphabet phonétique international)

Formants et voyelles

On appelle filtre vocalique l'ensemble des processus qui altèrent le débit issu des plis vocaux depuis le larynx jusqu'aux lèvres. En fonction de la configuration du conduit vocal, et de la position de la langue notamment, des plages de fréquences différentes vont être excitées dans le signal. Elles correspondent aux résonances du conduit vocal, encore appelés formants. Sur la figure 1.11 on peut visualiser la réponse en fréquence et la position des langue, mâchoire, lèvres pour les 3 voyelles cardinales.

A chaque voyelle correspond un jeu de maxima différents. Ces maxima spectraux sont appelés des formants et permettent de caractériser les voyelles du point de vue signal. Ainsi, en traitement du langage parlé, le contenu phonémique est en partie extrait de la détection des formants sur le signal, ou de critères en dépendant. Selon les conditions d'observations et la voyelle prononcée, le nombre de formants utiles varie, mais en français le nombre de formants perceptivement significatifs est généralement de 2 ou 3. On considère en moyenne 1 formant par tranche de 1000Hz entre 0 et 4000Hz.

Le triangle vocalique

On classe chaque voyelle en fonction de son lieu d'articulation (le lieu de plus forte constriction) dans le triangle vocalique. On retrouve à chaque extrémité du triangle les 3 voyelles dites "cardinales" : /a/, /i/ et /u/. Un trapèze reprenant cette représentation est donné sur la figure 1.12 en fonction de la position de la langue (antérieure, postérieure) et de l'ouverture de la bouche (ouverte, fermée). Comme la position des formants dépend aussi des lieux d'articulation, la position d'une voyelle dans le triangle donne aussi des informations sur sa réponse en fréquence. On peut donc lire sur ce trapèze une fréquence de premier formant décroissant avec les ordonnées, et une fréquence de deuxième formant décroissant avec les abscisses.

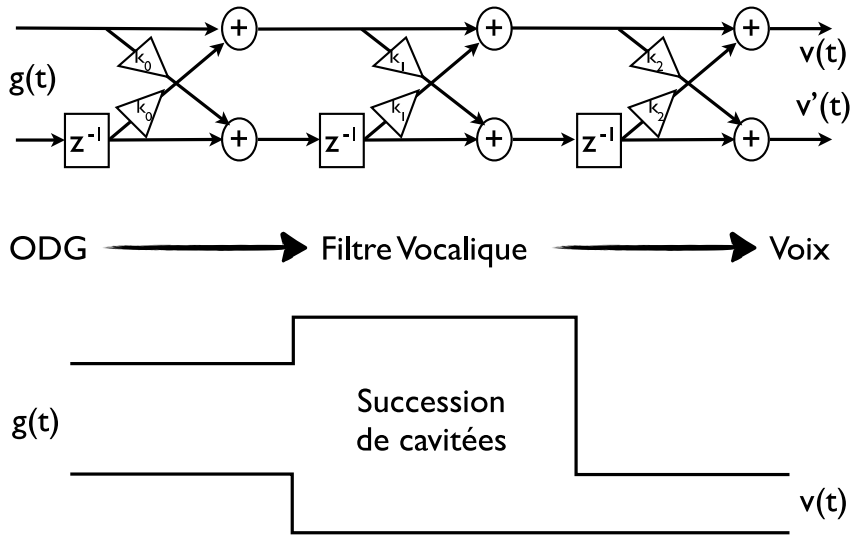


Figure 1.13 – Modélisation du conduit vocal, décrire le conduit sous la forme de résonateurs permet de modéliser le conduit comme un filtre en treillis, pour plus tard utiliser une modélisation autoregresive.

Le filtre autorégressif

La question qui se pose donc, en parallèle à l'analyse du signal vocal, est la modélisation du filtre vocalique. Nous l'avons vu précédemment, ce filtre est la cause de toutes les modifications du signal issu des plis vocaux. Dans une première approximation, on peut considérer que le trajet de l'ODG se fait dans le conduit vocal comme à travers une succession de tubes de diamètres et de longueurs différents, comme présenté sur la figure 1.13 où le signal d'entrée $g(t)$ est l'onde de débit glottique et le signal de sortie $v(t)$ est le signal acoustique résultant. Cette succession de cavités peut se modéliser par un filtrage en treillis [Kelly et Lochbaum, 1962] où les coefficients de réflexions k_i dépendent de la taille et section de chaque "tube". La structure d'un tel filtre est présentée sur la figure 1.13, le bloc z^{-1} représente un retard pur d'un échantillon dans le domaine discret.

Cependant, le filtrage en treillis est un filtrage itératif. Le pendant sous forme directe de ce filtre est sa modélisation autorégressive de coefficients $a(n)$. Dans cette modélisation, la sortie $v(n)$ du filtre ne dépend pas seulement du signal d'entrée $g(n)$, mais aussi des sorties précédentes telles que :

$$v(n) = g(n) + \sum_{i=1}^{N-1} a(i)v(n-i)$$

L'estimation des coefficients de ce filtre ainsi que la relation entre les coefficients $a(i)$ et les coefficients k_i du filtre en treillis seront traités au chapitre suivant.

Certains modèles complètent le filtre autorégressif (AR) par un filtre moyenneur (MA). Le filtre alors appelé ARMA [Fujisaki et Ljungqvist, 1987] permet une modélisation plus fine du conduit vocal, notamment pour les voyelles nasales (connues pour posséder des anti-résonances). Ce modèle permet ainsi de considérer le cas particulier vu précédemment pour lequel les cavités nasales font partie du conduit vocal ; il ne sera donc pas utilisé.

Les consonnes

Le conduit vocal ne sert bien entendu pas uniquement à la mise en forme des voyelles, mais aussi des consonnes. La phonétique propose ainsi une description des consonnes en rapport avec le lieu où se produit la constriction à l'origine de l'aspect particulier d'une consonne [Calliope, 1989]. On les différencie selon le lieu d'articulation (par exemple : labial, alvéolaire, dental, palatal), la quantité de voisement (voisé ou pas) et la qualité de la constriction : plosive lorsque le conduit vocal est entièrement fermé, fricative lorsqu'une constriction cause une turbulence, etc...

Ainsi, la consonne /t/ est une consonne non voisée, plosive et alvéolaire alors que la consonne /d/ est une consonne voisée, plosive et alvéolaire.

1.5 Qualités vocales

La notion de qualité vocale est difficile à définir de manière objective [Kreiman *et al.*, 2004], mais on peut la considérer comme ce qui a trait au contenu non linguistique et para linguistique du signal de parole. Un même contenu lexical peut-être compris différemment en fonction de la manière dont il est prononcé et du contexte d'énonciation. L'influence des informations para et non-linguistiques sur la réception du message parlé est donc considérable, et l'étude de la prosodie et de la qualité vocale reste primordiale pour la parole expressive [Fant, 1997]. La notion de prosodie ne sera pas étudiée dans son ensemble au cour de cette thèse. C'est sur la qualité vocale - parfois considérée comme une composante de la prosodie [Pfitzinger, 2006] - que se porte la présente étude.

Cette notion de qualité vocale peut être décrite selon plusieurs dimensions plus ou moins orthogonales. Selon [d'Alessandro C., 2006], on retrouve quatre composantes :

- La dimension serré / relâché
- La dimension d'effort
- La dimension de voisement : chuchoté - voisé
- La dimension de raucité de la voix (jitter et shimmer)

On peut aussi rajouter une dimension supplémentaire, le mécanisme de vibration des plis vocaux. Dans certains contextes (fin de phrase, modification volontaire de sa voix) un changement de mécanisme peut être la conséquence d'une volonté de changement de l'expressivité de la voix. Un changement de mécanisme s'accompagne systématiquement d'une modification des paramètres du modèle de production vocale [Childers et Lee, 1991].

Pour les 4 dimensions sus-citées, on retrouve une contribution importante de la configuration des plis vocaux. Une étude de Scherer [Scherer, 2003] a montré que l'état émotionnel du locuteur a une influence directe sur la configuration du larynx, et par voie de conséquence, sur la forme de l'onde de débit glottique qui en était issue. Un locuteur triste aura tendance à resserrer son larynx et à produire des ondes de débit glottique plus dissymétriques. Une voix de locuteur triste aura donc tendance à être plus riche en hautes fréquences.

Un des outils permettant de visualiser la corrélation entre qualité de la production vocale et forme de l'onde de débit glottique sur un signal de parole est le masque de Rothenberg. Sur la figure 1.14 tirée de [Sundberg, 1994], on visualise différentes formes d'ondes obtenues par ce procédé et la qualité vocale perçue.

On trouve aussi des études sur l'impact perceptif des paramètres du modèle de l'onde de débit glottique comme [van Dinther *et al.*, 2005] ou [Henrich *et al.*, 2003] qui tentent de quantifier le rapport entre modification de la forme de l'ODG et perception de la qualité de la voix.

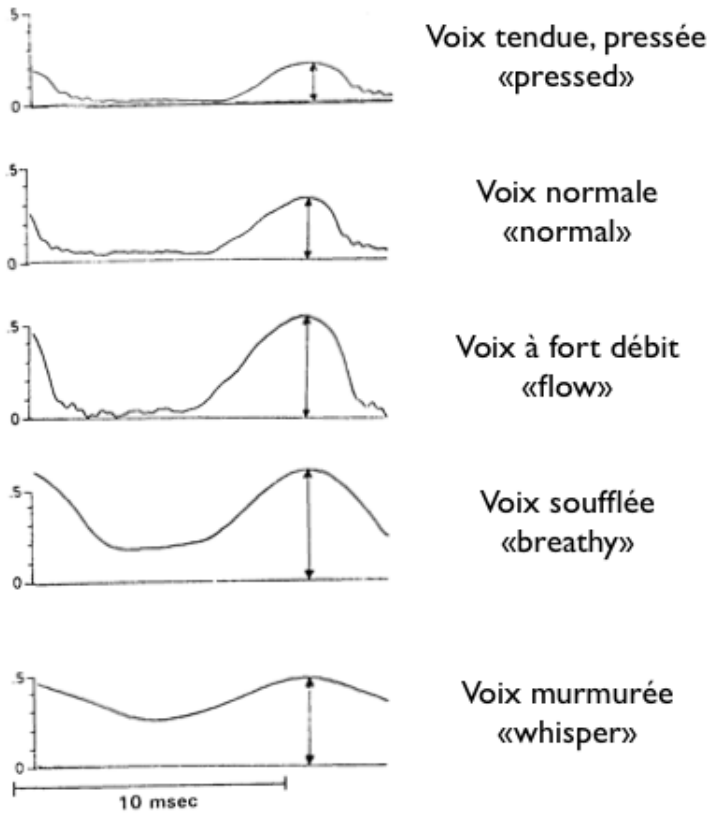


Figure 1.14 – Différents flux glottiques (en cm de mercure) obtenus avec un masque de Rothenberg pour différentes qualités vocales (figure 1 de [Sundberg, 1994]).

1.5.1 La dimension serré / relâché

La longueur de vibration des plis vocaux impacte directement le son produit. Plus cette longueur sera courte et plus les plis seront tendus, nécessitant une pression subglottique d'autant plus importante pour être séparés. Les travaux antérieurs [d'Alessandro C., 2006] ont montré que cette dimension était fortement corrélée avec la valeur du quotient ouvert O_q de l'onde de débit glottique, un quotient ouvert faible donnant un voisement qualifié de plus "serré". C'est pourquoi cette dimension est associée au quotient ouvert.

1.5.2 La dimension d'effort vocal

L'intensité perçue du signal vocal ne dépend pas uniquement de la pression sonore brute émise par le locuteur. En effet, le contenu spectral du signal joue aussi un rôle important à cet effet, et différents mécanismes entrent en jeu lorsque le locuteur cherche à augmenter l'intensité perçue de sa voix :

Le premier est l'adaptation du conduit vocal pour augmenter le rendement de la production. Ce phénomène est très présent chez les chanteurs lyriques, qui apprennent à ajuster la fréquence des résonateurs, causant le formant du chanteur [Winckel, 1954]. Cet effet augmente l'énergie du signal sans avoir besoin de modifier le débit glottique (en dehors de la possibilité d'interactions).

Le deuxième joue sur la configuration des plis vocaux. L'intensité perçue est aussi fonc-

tion de la teneur en hautes fréquences du signal. Plus le spectre d'un signal est riche en hautes fréquences et plus il est perçu comme étant "fort" [Fletcher et Munson, 1933]. Ce principe se retrouve lorsqu'un locuteur produit un effort vocal important, en "poussant" sur la voix pour augmenter l'intensité perçue, notamment afin de la faire porter plus loin [Traunmüller et Eriksson, 2000]. Une stratégie permettant cet effort est l'augmentation de la tension de plis vocaux qui a pour effet de rendre plus "pointue" la forme de la dérivée de l'ODG, augmentant par là même son asymétrie. On retrouve de tels résultats dans [C.Sapienza *et al.*, 1998] et [Guruprasad et Yegnanarayana, 2009].

1.5.3 La dimension de raucité

Une voix rauque, ou rugueuse, comme une surface, a un aspect non lisse, irrégulier. Elle est souvent associée à un "chat dans la gorge" ou au fait de "gratter", toujours dans cette idée d'irrégularités du signal. Ces irrégularités sont causées par des micro-variations du signal en temps et en amplitude [Kreiman et Gerratt, 2010] : les apériodicités structurelles. La première est le jitter : la variation de fréquence période à période. La deuxième est le shimmer : la variation d'amplitude période à période.

L'apparition de mucus, l'irrégularité du flux d'air issu des poumons ou encore des non-linéarités au niveau de la structure du mouvement des plis vocaux sont à l'origine de tels phénomènes. Une voix considérée saine présente toujours du "jitter" et du "shimmer" mais en faible quantité [Meike *et al.*, 2010]. Dans les cas extrêmes, ces apériodicités peuvent aussi avoir pour origine des dysfonctionnements de l'appareil vocal (comme c'est le cas pour la dyplophonie - doublement de la fréquence fondamentale).

1.5.4 La dimension de voisement

Une dernière dimension très importante joue sur la qualité vocale : il s'agit du voisement du signal. On décrit le voisement comme la mise en vibration des plis vocaux. Ce mouvement n'est pas nécessaire pour produire de la voix intelligible, nous cherchons à l'éviter dans le cas du chuchotement. Longtemps considéré comme un trait binaire "voisé / non voisé", des études [de Krom, 1993, Jackson et Shadle, 2001, d'Alessandro *et al.*, 1998] ont montré que cette dimension était plus fine que cela.

Lors de la phonation, une fermeture incomplète de la glotte laisse passer un débit d'air qui va causer des turbulences dans le conduit vocal. Ces turbulences se traduisent par un bruit aléatoire qui s'ajoute au signal harmonique du mouvement des plis vocaux. Le rapport entre l'énergie de chacune de ces contributions devient alors la quantité de voisement du signal. Une voix forte et claire sera généralement très voisée, alors qu'une voix douce, parfois faible, (sans pour autant être chuchotée, comme c'est le cas du murmure) aura un voisement bas.

1.5.5 Récapitulatif

On peut lier chacune des dimensions de la qualité vocale à des paramètres du modèle de production vocale. Ce récapitulatif est donné dans le tableau 1.1.

1.6 Conclusion

Ce premier chapitre a présenté les bases du travail de cette thèse en terme de contexte et de modèles utilisés. Il s'intéresse particulièrement à l'analyse périodique des signaux de parole, dans le but d'arriver à les décomposer et paramétriser les informations qu'ils contiennent.

Table 1.1 – Tableau récapitulatif des liens entre qualité vocale et paramètre du modèle de production

dimension	paramètres	effet qualitatif
serré / relâché	O_q	tension de la voix
fort / faible	α_m et Q_a	intensité sonore perçue
raucité de la voix	jitter et shimmer	irrégularité dans le signal vocal
quantité de voisement	rapport $\frac{\text{Voisé}}{\text{Non voisé}}$	clarté de la voix

Le modèle de production de la voix, tant du point de vue physique que du point de vue signal a été présenté. L'attention est particulièrement portée sur le modèle utilisé pour décrire l'onde de débit glottique, représentant le gradient de débit issu des plis vocaux lors de la phonation. Le modèle le plus complet [Fant *et al.*, 1985], proposant 3 critères de formes, a été retenu. Différentes méthodes d'observation du mouvement de ces plis vocaux ont été présentées. Elles servent de référence dans l'évaluation de la performance des méthodes d'analyse. La deuxième partie du modèle de production, le filtre vocalique, a aussi été présentée mais ne sera que très peu étudiée au cours de ce manuscrit. La modélisation auto-régressive du filtre vocalique demeure un *a priori* très fort sur l'analyse des signaux.

La notion de qualité vocale, exposée dans ce chapitre, est étroitement liée à la configuration de l'appareil de production vocale, tant au niveau de la source que du filtre. Le chapitre suivant fera un état de l'art des méthodes et modèles actuels utilisés pour l'étude des signaux de parole, en particulier du point de vue de l'analyse de la qualité vocale. Il présentera notamment une nouvelle approche du modèle linéaire de production.

Résumé

La notion de qualité vocale

La notion de qualité vocale est intimement liée à la manière dont la parole est produite. Cette production se situe à deux niveaux, repris dans la modélisation source/filtre. Le filtre modifie le signal issu de la source : le débit issu des plis vocaux.
Le modèle du débit glottique retenu, le modèle LF, est un modèle analytique qui a fait ses preuves et est largement utilisé, il présente trois critères de forme qui peuvent être reliés aux dimensions de la qualité vocale : la dimension d'effort vocal, la dimension de tension vocale, la dimension de voisement et la dimension de raucité. Ces quatre dimensions ont été présentées et modélisées.

On peut donc penser

A partir du modèle de la production vocale, l'analyse des signaux de parole devrait permettre de relier les paramètres de modèle à la notion de qualité vocale. Actuellement, les dimensions de cette production sont plus ou moins bien maitrisées en matière d'estimation, de décomposition de signaux et d'analyse en général. Un travail important reste à réaliser afin de mieux comprendre les liens qui existent entre la manière dont la voix est produite et la manière dont elle est perçue par un auditeur.

Chapitre 2

État de l'art de l'estimation des paramètres de la source

Sommaire

2.1 La détection des instants de fermeture glottique 41
2.1.1 Méthode naïve : résidu de la prédiction linéaire 42
2.1.2 Utilisation des propriétés de phase 42
2.1.3 Produit Multi-échelles 44
2.1.4 Filtrage en fréquence zéro 45
2.2 Filtrage inverse et caractérisation de la source 46
2.2.1 La prédiction linéaire (LPC) pour l'estimation du débit glottique 46
Histoire 46
Filtre autorégressif 46
Estimation des coefficients du filtre 46
Coefficient de préaccentuation 47
2.2.2 Filtrage inverse 47
Variation sur phase fermée 48
La modélisation tout-pôle discrète 49
Filtrage itératif adaptatif IAIF [Alku, 1992] 49
2.2.3 Le modèle causal/anticausal 49
Prédiction linéaire, phase et causalité 50
Avantages de l'approche CALM 51
2.2.4 La théorie ZZT 52
Choix de la forme de la fenêtre d'analyse 53
Choix de la taille de la fenêtre d'analyse 54
Position de la fenêtre d'analyse 54
Calcul des zéros du polynôme de la transformée en Z 55
Déplacement de zéros 56
2.2.5 Algorithme de la décomposition par ZZT 56
2.2.6 ZZT et Filtrage Inverse 57

2.2.7 Quelques exemples 57
2.2.8 Décomposition causale/anticausale par Cepstre Complexe 58
2.2.9 Vers l'estimation des paramètres du modèle de source 59
Estimation directe sur l'onde de débit glottique 59
Estimations contraintes ARX 59
2.2.10 D'autres critères à signification perceptive 60
Le quotient d'amplitude normalisé (NAQ) 60
La différence d'amplitude entre harmoniques 61
Le paramètre réduit R_d 61
2.3 Périodicités, Apériodicités 62
2.3.1 Origine des apériodicités 62
Apériodictés structurelles 62
Bruits Additifs 62
2.3.2 Décomposition périodique/apériodique des signaux de parole 63
Estimation du rapport entre harmoniques et bruit 63
Modèle de la voix périodique 63
Fréquence limite de voisement 65
Modèle sinusoïdal 65
2.3.3 Estimation de la partie voisée 65
Méthodes d'estimation statistiques 65
Utilisation d'un filtre harmonique 66
2.3.4 Filtre harmonique, optimisations diverses 66
Reconstruction du bruit par interpolation 66
Reconstruction du bruit de manière itérative 67
Décomposition par adaptation de la fenêtre d'analyse [Jackson et Shadle, 2001] 68
2.4 Conclusion 68

La qualité vocale est donc définie autour de quatre dimensions, chacune imputable à des propriétés physiques présentes dans le signal. Pour évaluer la relation entre voix expressive et qualité vocale, il convient de mesurer ces propriétés, qu'elles soient pour segmenter les signaux de parole, décomposer les contributions périodique et apériodique ou encore pour mesurer les paramètres du modèle de source. Le travail réalisé au cours de cette thèse cherche à améliorer ou développer des méthodes d'estimation des différentes propriétés de la qualité vocale avant de les appliquer à un corpus de parole expressive et d'analyser leurs valeurs et les liens existant entre elles.

Ce chapitre dresse l'état de l'art sur trois domaines, chacun abordant un aspect précis de l'analyse des signaux vocaux servant à en extraire des critères de qualité vocale :

- La première section traite de la segmentation des signaux vocaux, et en particulier de l'estimation des instants de fermeture glottique : différentes méthodes seront vues, liées au travail réalisé dans le chapitre 3.
- La deuxième section traite de l'analyse de l'onde de débit glottique et de son estimation. Seront vues dans cette section : la prédiction linéaire et l'utilisation de la modélisation autorégressive pour l'analyse des signaux vocaux, la méthode de décomposition source/filtre par modèle causal/anticausal ainsi que la paramétrisation de l'ODG. Ces travaux font référence aux méthodes présentées dans le chapitre 5.
- La troisième section traite de la décomposition des signaux vocaux en parties périodique et apériodique, et notamment du cadre technique permettant cette décomposition. Une nouvelle méthode de décomposition sera présentée au chapitre 4.

Compte tenu de la quantité importante de travaux dans chacun de ces trois domaines, l'intérêt est particulièrement porté sur les travaux en relation directe avec les axes et paradigmes de recherche explorés lors ce travail de thèse.

2.1 La détection des instants de fermeture glottique

Les instants de fermeture glottique (GCI) sont les événements caractéristiques permettant la détermination de la fréquence fondamentale, leur estimation permet d'obtenir facilement et précisément la fréquence fondamentale du signal mais aussi de réaliser des analyses dites *pitch synchrones* (en synchronie avec la fréquence fondamentale du signal). Ainsi, avant même de chercher à estimer la source du signal, en extraire les instants de fermeture glottique semble une priorité. La connaissance de ces instants permet notamment d'informer certaines techniques d'estimation du débit glottique comme il sera vu à la section suivante.

Au cours d'un cycle glottique, le point de rapprochement maximal entre les plis vocaux est un instant très caractéristique. La rupture (plus ou moins brutale) du flux d'air provoque un afflux énergétique à la fois important et bref, dont la dérivé seconde est sensiblement comparable à une impulsion de Dirac ayant les propriétés énoncées dans l'équation 2.1.

$$\delta(t) = 0 \text{ pour } t \in \mathbf{R}^* \quad \text{et} \quad \int_{-\infty}^{\infty} \delta(t)dt = 1 \tag{2.1}$$

Ainsi, la majeure partie des méthodes cherchant à estimer des GCI se penchent sur l'estimation d'un point caractérisé par des changements dans les propriétés du signal [Ananthapadmanabha S. et Yegnanarayana, 1979, Cheng et O'Shaughnessy, 1989, Moulines et R., 1990, Smits et Yegnanarayana, 1995] : quantité de passage par zéros, différences entre les phases, propriétés statistiques. On s'intéresse aussi à la présence d'une énergie concentrée en un point notamment par l'alignement des phases. Dans tous les cas, on cherche

donc un point dans le temps provoquant un changement important dans le signal. Après avoir montré la difficulté de retrouver ces points par application directe du modèle source/filtre linéaire, trois méthodes seront présentées :

1. une approche basée sur l'analyse de la pente de la phase
2. une approche basée sur l'analyse temps fréquence des signaux de parole
3. une approche basée sur un filtrage en fréquence zéro [Murty *et al.*, 2009].

Dans un souci de concision, seules ces trois méthodes seront abordées, car elles résument bien les approches utilisées pour contourner la difficulté de l'estimation directe des GCI sur le signal de parole. Certaines méthodes utilisant ces approches pratiquent une estimation *contrainte*, notamment des instants de fermeture glottique [Degottex *et al.*, 2010], ces méthodes seront abordées en même temps que l'estimation des paramètres du débit glottique. En effet, on peut considérer que le GCI fait partie de ces paramètres, mais dans bien des cas (tout comme la fréquence fondamentale) il fait aussi partie du jeu d'informations nécessaires à une bonne analyse des signaux de parole.

2.1.1 Méthode naïve : résidu de la prédiction linéaire

Par définition du modèle linéaire de production vocale, déconvoluer le signal de parole par une estimée des fonctions de transfert de la glotte et du conduit vocal devrait permettre de trouver les instants d'excitation glottique. L'équation 2.2 montre comment arriver à retrouver le train d'impulsions de TF δ_{F_0} par déconvolution (division spectrale) du signal par les filtres de fonction de transfert G et F respectivement pour la glotte et le conduit vocal. Un telle déconvolution peut être opérée par prédiction linéaire, présentée en section 2.2.1.

$$\begin{aligned} S &= GF\delta_{F_0} \\ \delta_{F_0} &= \frac{S}{GF} \end{aligned} \tag{2.2}$$

La figure 2.1 illustre un tel principe, où le résidu de la prédiction linéaire du signal vert est donné en bleu. On constate sur cette figure que seuls les instants où l'excitation est bien marquée peuvent servir à la détection de GCI. De plus, des excitations secondaires peuvent apparaitre sur le résidu rendant d'autant plus difficile la localisation des instants d'excitation. Les travaux publiés dans [Ananthapadmanabha S. et Yegnanarayana, 1979] ont notamment traité cette question et ont montré que de nombreuses ambiguïtés apparaissaient lors de l'analyse directe du résidu. Parmi les méthodes proposées pour contourner ces ambiguïtés : une analyse des propriétés de la phase du résidu.

2.1.2 Utilisation des propriétés de phase

Une méthode proposée plus tard par Smits et Yegnanarayana [Smits et Yegnanarayana, 1995] propose d'utiliser la pente de la phase déroulée pour déterminer la position des instants de fermeture glottique. En effet, un système linéaire et causal, comme c'est le cas pour la modélisation du filtre vocalique (système passif) possède tous ses pôles à l'intérieur du cercle unité, et constitue un système à phase minimale. Dans le cas du placement de la fenêtre d'analyse sur le GCI, la pente de la phase est donc nulle. Tout déplacement de la fenêtre d'analyse verra donc un terme linéaire s'appliquer à la pente de la phase, terme proportionnel au déplacement par rapport au GCI. Un exemple est donné sur la figure 2.2 pour un simple train d'impulsion. Selon

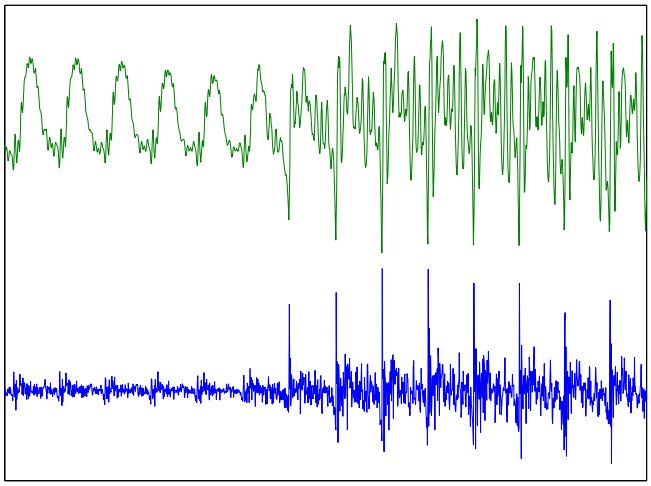


Figure 2.1 – Utilisation du filtrage inverse pour la détection des GCI (fenêtre glissante de 20ms toutes les 10ms, signal échantillonné à 16kHz, 18 pôles estimés pour le filtre AR). Signal original en vert, résidu de la prédiction linéaire en bleu.

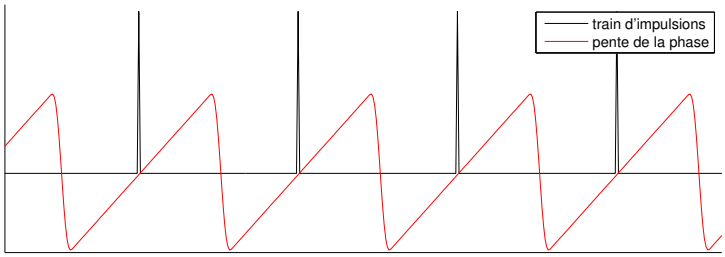


Figure 2.2 – Illustration de l'utilité de la pente de la phase pour déterminer l'emplacement des impulsions dans un signal.

[Smits et Yegnanarayana, 1995] une telle détection est opérée sur le résidu du filtrage inverse et demande une connaissance *a priori* de la fréquence fondamentale moyenne du signal afin d'appliquer un fenêtrage optimal. La figure 2.3 présente la même méthode appliquée à un signal réel, on remarque que des fausses détections du GCI sont causées par des oscillations de la pente de la phase (ovale en pointillés). Afin d'éviter des sur-détections, une méthode plus robuste a été développée.

La méthode DYPSA - pour DYnamic programming Phase Slope Algorithm - développée par Kounoudes [Kounoudes *et al.*, 2002], Naylor[Naylor *et al.*, 2007] et Brookes, utilise cette propriété pour estimer les instants de fermeture glottique. Une programmation dynamique est appliquée à l'estimation en tenant compte d'une fonction de coût de déviation de période à période, ainsi qu'un critère de similarité entre formes d'ondes et d'une correction a posteriori des GCI candidats.

Cette méthode est extrêmement efficace et présente généralement plus de 90% de bonnes détections dans les 0.5ms autour de la référence extraite du signal EGG. Mais la méthode localise les GCI avec une précision de 0.25ms selon [Kounoudes, 2001]. De plus, la quantité importante de traitements en amont et en aval rend cette méthode peu appropriée aux cas atypiques de

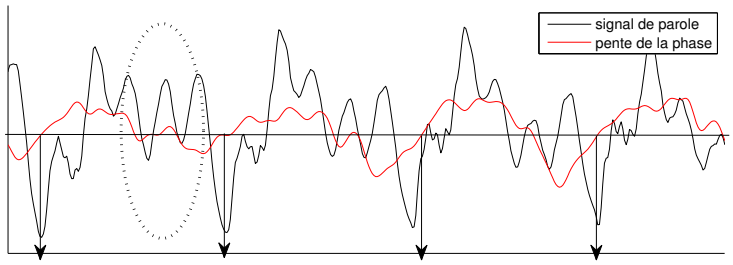


Figure 2.3 – La méthode de Smits et al. [Smits et Yegnanarayana, 1995] appliquée à un signal réel, des oscillations peuvent causer des fausses détections.

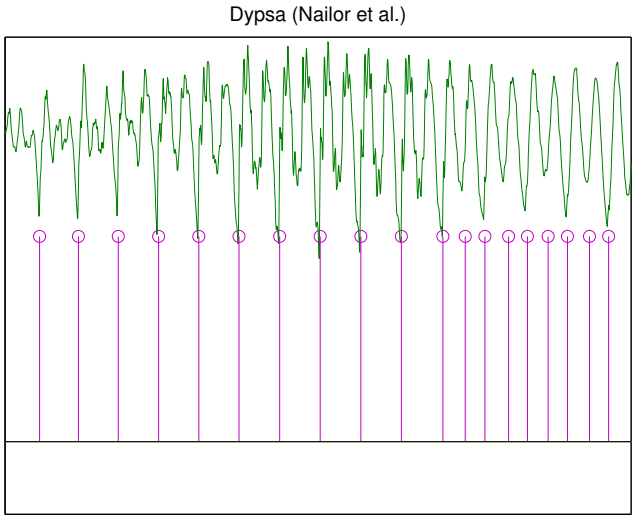


Figure 2.4 – Exemple problématique d'estimation des GCI sur un signal de parole avec la méthode DYPSA (DYnamic programming Phase SLope Algorithm) [Naylor *et al.*, 2007] avec le signal en vert, et les GCI détectés en mauve.

signaux vocaux et peut notamment mener à des sur ou sous-estimations de GCI comme c'est le cas sur la figure 2.4 où le changement de qualité vocale (harmoniques moins marqués) sur le signal (en vert) induit l'algorithme en erreur et cause une sur-détection des GCI (en mauve).

Dans la même idée de minimisation de phase, Degottex et al. [Degottex *et al.*, 2010] ont proposé un estimateur permettant de localiser à la fois le GCI et de déterminer la forme de l'onde de débit glottique. Cette démarche étant similaire à celle qui sera entreprise au chapitre 3, ces travaux seront discutés plus en détail à cet endroit.

2.1.3 Produit Multi-échelles

L'idée d'une analyse multi résolution a été introduite par Mallat [Mallat et Hwang, 1992]. Répandue en traitement des images, cette idée a mené aux travaux de Kadambe et al. [Kadambe et Boudreaux-Bartels, 1992], qui utilisaient un produit entre bandes d'analyse en ondelettes d'un signal de parole pour détecter la position des instants de fermeture glottique. Récemment, le produit multi-échelles développé par Bouzid et al . [Bouzid et Ellouze, 2007] a été présenté, il utilise aussi une décomposition du signal par bancs de filtres en ondelettes, mais

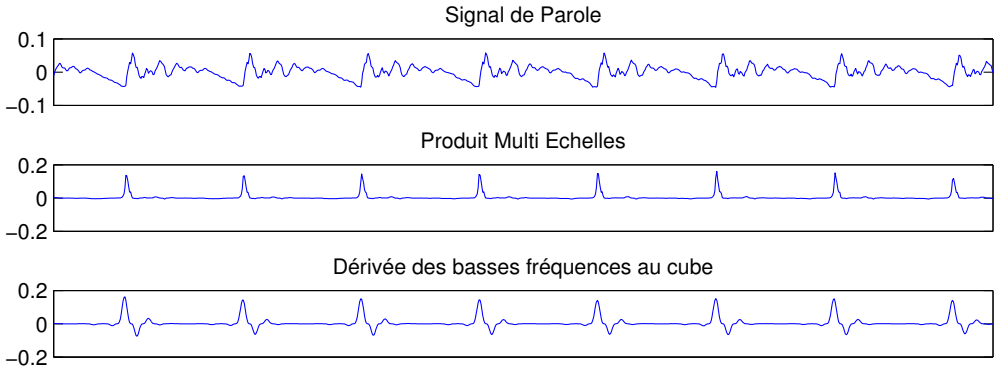


Figure 2.5 – Le produit multi-échelles face à un produit de la partie basse fréquence du signal de parole. On constate que les résultats obtenus sont similaires, mais plus contrastés dans le cas du MSP

alors que les travaux de Kadambe utilisent les hautes fréquences du banc de filtres, le produit multi-échelles de Bouzid se contente des trois échelles de plus basses fréquences.

Dans cette section, seul le produit échelle le plus récent [Bouzid et Ellouze, 2007] sera étudié. Les 3 signaux issus des bandes basses fréquences du banc de filtres en ondelettes sont multipliés entre eux. L'ondelette utilisée pour la décomposition possède le spectre défini à l'équation 2.3.

$$\Psi(\omega) = i\omega \left(\frac{sin(\omega/4)}{\omega/4} \right)^4 \tag{2.3}$$

Il semblerait que le produit multi-échelles (MSP - Multi Scale Product) se contente de créer des harmoniques supplémentaires (par non linéarité du produit) à partir des informations basses fréquences pour détecter avec précision l'instant de fermeture glottique. Pour démontrer cela, regardons la différence entre MSP et la dérivée du cube de la partie basses fréquences (0-800Hz) du signal. Sur la figure 2.5 sont présentés un signal de parole, son produit multi-échelles ainsi que la dérivée du cube de la partie basses fréquences du signal (0-800Hz). On peut voir que les résultats obtenus sont similaires et se demander légitimement si on peut extraire suffisamment d'informations sur la partie basses fréquences du signal pour déduire une localisation précise des instants de fermeture glottique. Les récents travaux de Murty et al. [Murty *et al.*, 2009] ont cependant montré que la partie basses fréquences des signaux vocaux véhicule probablement une information suffisante, ces travaux seront discutés au paragraphe suivant.

Le produit multi-échelles de Bouzid et al. prend tout son sens en ce qui concerne l'instant d'ouverture glottique. L'énergie de l'instant d'ouverture glottique étant principalement contenue dans les basses fréquences. La difficulté réside alors dans la détection de ces instants, car ils produisent des maximas locaux bien moins marqués.

Les méthodes décrites dans [Kadambe et Boudreaux-Bartels, 1992, Bouzid et Ellouze, 2007] se contentent de ramener les propriétés de la représentation temps-fréquence (2D) en une seule dimension. Une approche différente développée au LIMSI [Tuan et d'Alessandro, 1999], prenant davantage en compte les propriétés bi-dimensionnelles des représentations temps fréquence sera discutée plus en détails dans les chapitres 3 et 5.

2.1.4 Filtrage en fréquence zéro

Récemment, une nouvelle proposition de détection conjointe de la fréquence fondamentale et des instants de fermeture glottique a été proposée par Murty et al. dans [Murty *et al.*, 2009]. Basée sur les propriétés basses fréquences du signal, elle propose l'utilisation de la sortie d'un double résonateur en fréquence zéro (affaiblissement de 24 dB par octave) pour détecter à la fois la fréquence fondamentale, les instants de fermeture glottique, et l'activité de voisement. Cette mesure s'apparente cependant davantage à une mesure d'énergie à court terme.

2.2 Filtrage inverse et caractérisation de la source

2.2.1 La prédiction linéaire (LPC) pour l'estimation du débit glottique

Histoire

Lors des premiers travaux sur les signaux vocaux, il est vite apparut que l'utilisation de filtres résonants du second ordre placés à des fréquences bien précises donnait l'illusion de prononcer des voyelles. Ces filtres, alors appelés formants [Fant, 1960] on été mis en relation avec les études sur l'appareil de production vocale pour déboucher sur une modélisation source/filtre. L'idée décisive a été de modéliser le canal emprunté par l'air issu de la glotte comme une série de cylindres de longueurs et de sections différentes.

Transposée en modèle signal, cette modélisation acoustique donne une réponse impulsionnelle infinie, modélisée par un filtre autorégressif. Ce sont précisément les coefficients de ce filtre que la prédiction linéaire cherche à estimer.

Filtre autorégressif

En première approximation on peut considérer le filtre du conduit vocal comme un filtre linéaire et causal. En conséquence, la réponse s du filtre de coefficients $a(n)$ à une entrée e peut être écrite comme suit :

$$s(n) = e(n) + \sum_{i=0}^{N-1} s(n-i)a(i)$$

Où a est le vecteur des coefficients du filtre dit autorégressif, de dimension N représentant l'ordre du filtre. L'adjectif *autorégressif* vient du fait que pour calculer $s(n)$, il faut procéder à des opérations sur les échantillons précédemment calculés de s. Ces filtres peuvent donc être instables, car une mauvaise série de coefficients peut mener à une série s divergente.

Á l'aide de la représentation de Laplace adaptée au domaine discret (Transformée en Z), on montre que pour que la série a représente les coefficients d'un filtre linéaire, causal et stable il est nécessaire que toutes les racines X_n du polynôme $P_a(x) = \sum_{i=0}^{N-1} a_i x^{-i}$ soient à l'intérieur du cercle unité ($|X_n| < 1$).

Comme il sera montré par la suite, le fait de considérer le conduit vocal comme un filtre autorégressif est une hypothèse très importante en matière d'analyse/synthèse des signaux vocaux. Cette hypothèse montre tout de même certaines limites car elle contraint l'expression du filtre. Dans la mesure où les estimations du filtre et de la source sont liées, l'utilisation d'une modélisation de la source comme CALM (pour Causal/Anti-causal Linear Model [Doval *et al.*, 2003]), pourrait être une alternative intéressante à la modélisation autorégressive. En effet, les filtres autoregressifs ne sont qu'une petite partie des filtres à phase minimale. Cette modélisation sera vue par la suite.

Estimation des coefficients du filtre

Soit $\mathbf{s}$ un vecteur de L échantillons de sortie du système dont on cherche à estimer les coefficients :

$$\mathbf{s_L(n)} = \left[\begin{array}{cccc} s(n) & s(n-1) & \dots & s(n-L+1) \end{array} \right]^T$$

On cherche à minimiser l'erreur de prédiction à chaque échantillon, c'est à dire :

$$\begin{aligned} e(n) &= s(n) - \sum_{i=0}^{N-1} a(i)s(n-i) \\ &= s(n) - \mathbf{a}_L^T \mathbf{s}(n-1) \end{aligned}$$

Pour réaliser cette minimisation, on peut utiliser le critère des moindres carrés et minimiser la grandeur $E[e^2(n)]$, ce qui mène aux équations (où $r(i)$ est le coefficient d'autocorrelation) :

$$\begin{aligned} \sum_{i=1}^{L} a_i r(i-j) &= -r(j) \\ r(i) &= E[x(n)x(n-i)] \end{aligned} \tag{2.4}$$

Les équations 2.4 sont appelées équations normales ou encore équation de Yule-Walker [Yule, 1927].

Coefficient de préaccentuation

Le signal de source dérivée possède une pente de -6dB par octave. Afin de se rapprocher de l'hypothèse d'excitation par un bruit blanc (énergie identique sur toute la plage des fréquences), on peut utiliser un dérivateur appliqué au signal. Ce dernier présente donc une pente de +6dB par octave, ce qui annule la pente de la source [Paliwal, 1984]. Par ailleurs, on peut aussi régler la fréquence de coupure de ce dérivateur. Nous avons vu précédemment que si la source présentait une pente spectrale, elle n'était pas la même pour toutes les configurations de paramètres, en particulier en basse fréquence. Un ajustement de cette fréquence demande un *a priori* sur la source du signal à estimer, mais permet de réaliser des estimations très efficaces. Il existe par exemple une méthode d'estimation du filtre qui utilise cette propriété pour ajuster l'intégrateur de préaccentuation de manière itérative (IAIF[Alku, 1992], vu par la suite).

Il est possible de rencontrer des filtres de préaccentuation d'ordre supérieur à 1 [Nordstrom et Driessen, 2006]. Le choix du filtre se révèle alors plus délicat et nécessite un ajustement, mais permet des estimations de formants plus constantes avec notamment une indépendance vis à vis de l'effort vocal.

2.2.2 Filtrage inverse

La manière la plus évidente d'atteindre le flux glottique, source du signal de parole, est d'appliquer l'inverse du filtre du conduit vocal estimé par prédiction linéaire [Rabiner et Schafer, 1978]. Si l'inversion d'un filtre autorégressif est une opération très simple, elle nécessite tout de même quelques précautions et notamment la préaccentuation du signal. Une sélection peut être opérée sur les pôles du filtre estimé afin de ne conserver que ceux qui ont une réalité physique (i.e. : représentent un formant). Dans ce cas là, la règle d'un formant par bande 1000Hz est retenue, en fonction des observations faites sur le conduit vocal. De plus, il est nécessaire de choisir un nombre de pôles en rapport avec la fréquence d'échantillonnage : la méthode de prédiction linéaire cherchant à minimiser l'erreur sur tout le spectre, elle a tendance à répartir les pôles en

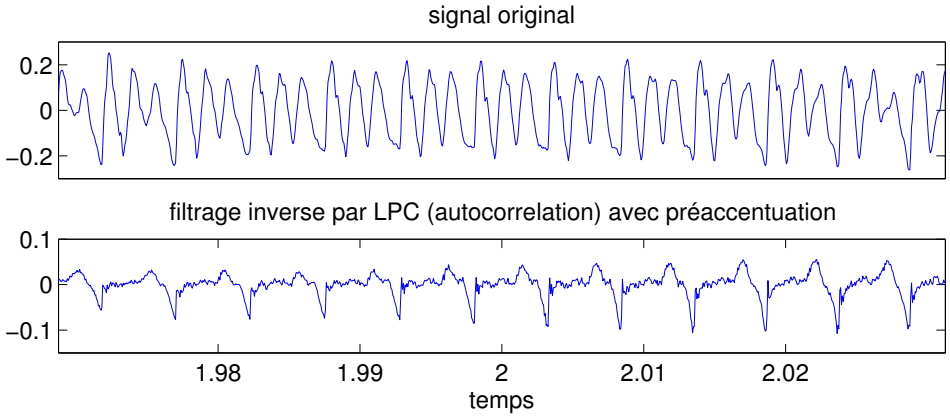


Figure 2.6 – Exemple d'estimation LPC sur un signal de parole. 18 pôles, fenêtre de 20ms (pondération Hanning), superposition de 10ms. Temps en secondes, amplitudes arbitraires.

conséquence. On choisit donc de suivre la formule $n_{ples} = \frac{F_s}{1000} + 2$, soit un pôle par bande de 1000Hz + 2 pôles. Pour une fréquence d'échantillonnage de 16kHz, qui sera la valeur par défaut dans ce document, on choisira 18 pôles.

On estime donc $\hat{a}(n)$ les coefficients du filtre vocalique par LPC sur le signal préaccentué par le paramètre a, généralement pris à la valeur $a = 0.98$ [Paliwal, 1984] sur le signal $\hat{s}(n)$:

$$\hat{s}(n) = s(n) - as(n-1)$$

Dans certains cas où l'ordre choisi est plus important que la valeur de $\frac{F_s}{1000} + 2$, on construit alors le filtre corrigé $a(n)$ par sélection des pôles sur $\hat{a}(n)$:

$$a(n) = \hat{a}(n)|_{\text{une paire de pôles par 1000Hz}}$$

Soit a(n) le jeu de coefficients du filtre vocalique estimés par LPC, on retrouve alors le résidu de ce filtrage inverse $g(n)$ par :

$$g(n) = \sum_{i=0}^{N} a(i)s(n-i)$$

On note que la préaccentuation ne sert qu'à l'estimation du filtre vocalique, et que le filtrage inverse est appliqué sur le signal original. Un exemple de filtrage inverse par LPC utilisant un filtre de préaccentuation de fonction de transfert $1 - 0.98z^{-1}$ est utilisé sur la figure 2.6. Un signal de parole a été analysé par LPC (autocorrelation) sur des sections de 20ms avec une pondération par la fenêtre de Von Hann [Harris, 1978], et une superposition de 10ms.

Variation sur phase fermée

La partie du débit glottique correspondant à la phase fermée est un instant propice à l'estimation de la réponse du filtre du conduit vocal. Durant ce temps là, le débit glottique est considéré constant, et les oscillations du signal de parole ne sont que la réponse du filtre à l'excitation glottique. En faisant l'hypothèse que l'essentiel de l'excitation se fait sur le GCI, on observe alors une réponse impulsionnelle. C'est donc sur ce laps de temps qu'on réalise en théorie l'estimation la plus fidèle de la réponse impulsionnelle du débit glottique [Deller, 1981, Wong *et al.*, 1979, Alku *et al.*, 2009].

Cependant, des problèmes apparaissent :

- Il faut connaitre la valeur du quotient ouvert (ou du quotient fermé) ainsi que la position du GCI pour pouvoir appliquer cette technique.
- Dans le cas d'une fréquence fondamentale élevée ou d'une phase fermée très courte, le nombre d'échantillons disponibles peut limiter l'estimation de la réponse impulsionnelle (taille du filtre). Ainsi, on se retrouve limité par la règle des 18 pôles pour 16kHz de fréquence d'échantillonnage à partir de $O_q = 0.65$ et $F_0 = 300Hz$ environ.
- Bien souvent, on observe un repliement de la réponse impulsionnelle du filtre d'une période à l'autre, ce qui cause une mauvaise estimation du filtre sur la phase fermée.

Cette méthode jouit d'une certaine popularité, notamment en vertu de ses excellents résultats sur des voix à faible fréquence fondamentale et dans des conditions idéales d'enregistrement (notamment par l'ajout d'un signal EGG synchrone [Krishnamurthy et Childers, 1986]).

La modélisation tout-pôle discrète

La modélisation tous-pôles discrète (DAP - Discrete All Poles [El-Jaroudi et Makhoul, 1991]) calcule la réponse impulsionnelle du filtre sur les harmoniques du signal uniquement. Dans le cas de signaux à haute fréquence fondamentale, la période est plus courte que la réponse impulsionnelle du filtre, et le signal se retrouve replié sur lui même. Le principe de la DAP est de résoudre ce problème en modélisant l'enveloppe spectrale uniquement sur les points fréquentiels correspondants aux multiples de la fréquence fondamentale. La minimisation de l'erreur E_{DAP} se fait par le calcul de la distance modifiée de Itakura-Saito où $S(\omega_m)$ et $\hat{S}(\omega_m)$ sont respectivement les spectres observés et estimés pris aux points contenant les harmoniques (de fréquence ω_m) :

$$E_{DAP} = \sum_m \frac{S(\omega_m)}{\hat{S}(\omega_m)} - log(\frac{S(\omega_m)}{\hat{S}(\omega_m)}) - 1$$

Filtrage itératif adaptatif IAIF [Alku, 1992]

Cette méthode tire profit de l'ajustement du filtre de pré-accentuation pour le filtrage inverse. La méthode IAIF [Alku, 1992] (pour *Iterative Adaptive Inverse Filtering*) procède donc à un ajustement itératif de ce filtre (d'ordre 1, 2 ou 4). Le critère de minimisation repose alors sur le résidu de l'estimation LPC du signal pré-accentué. Cette méthode ne propose pas une estimation à part entière, mais un contexte itératif dans lequel procéder à cette estimation. Cet algorithme peut donc s'appliquer à différentes méthodes dont la prédiction linéaire à base d'autocorrélation ou la modélisation tout pôles discrète par exemple.

Une illustration du principe d'ajustement itératif de IAIF est donné sur la figure 2.10.

2.2.3 Le modèle causal/anticausal

A partir de l'observation du modèle LF, Doval et al. [Doval *et al.*, 2003] ont proposé un nouveau paradigme d'approche du modèle de la source glottique en décomposant phase ouverte et phase fermée à partir de la position des pôles du signal par rapport au cercle unité.

$$\frac{dg}{dt}(t) = \begin{cases} E_0 sin(\omega_g t)e^{\alpha t} & 0 < t < T_e \\ -\frac{E_0}{\epsilon}[e^{-\epsilon(t-T_e)} - e^{-\epsilon(T_a-T_e)}] & T_e < t < T_e + T_a \end{cases} \tag{2.5}$$

En effet, l'analyse de l'équation 2.5 du modèle LF de la source glottique révèle deux éléments indépendants. La phase ouverte entre 0 et T_e est décrite par une exponentielle croissante modulée par un sinus. Cette onde peut se voir comme la réponse tronquée d'un filtre causal stable du

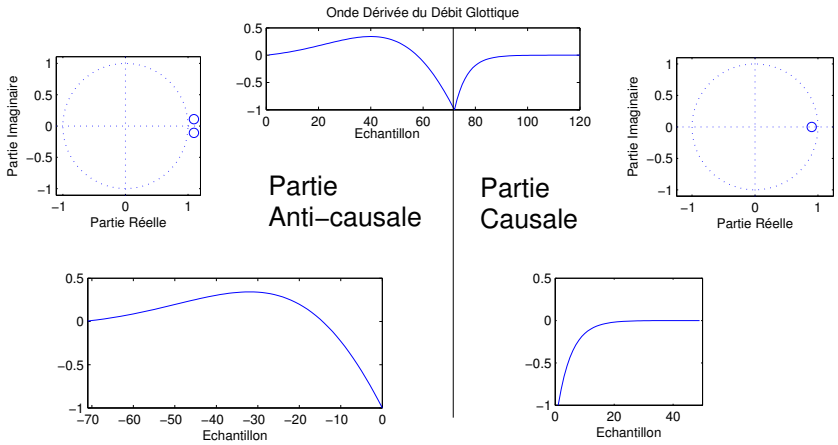


Figure 2.7 – Illustration du positionnement des pôles pour les deux parties du modèle CALM. La forme d'onde de la partie ouverte ressemble bien à la réponse impulsionnelle d'un filtre du deuxième ordre anticausal stable. De même, la forme d'onde de la partie causale est la réponse impulsionnelle d'un filtre du premier ordre.

deuxième ordre à une impulsion placée à l'instant de fermeture glottique mais dans une progression inversée du temps. On utilisera donc le terme "anti-causal" pour parler de cette phase ouverte, ce terme restant purement descriptif du comportement du modèle afin de garder la progression du temps dans son sens naturel. Un tel filtre présente deux pôles, situés en symétrie Hermitienne à l'extérieur du cercle unité. La phase ouverte est illustrée sur la partie gauche de la figure 2.7.

La phase fermée entre T_e et T_0 est, quant à elle, décrite par une exponentielle décroissante, peut être vue comme la réponse d'un filtre causal stable du premier ordre à une impulsion placée à l'instant de fermeture glottique. Un tel filtre présente un pôle situé sur l'axe réel à l'intérieur du cercle unité. Cette phase fermée est illustrée sur la partie droite de la figure 2.7.

Ce constat a permis de décrire le modèle de la source glottique selon l'analyse de la phase du signal, et en particulier de définir la fréquence F_g du formant glottique (fréquence de résonance du filtre anticausal stable du premier ordre décrivant la phase ouverte). Un filtre anticausal stable possède une progression de phase opposée à celle d'un filtre causal. Les études de Bozkurt et al. [Bozkurt *et al.*, 2004b] notamment, ont permis de mettre en lumière cette propriété et de différencier formant glottique et formants vocaliques lors de l'analyse des signaux de parole. Le modèle causal/anticausal sera décrit par l'acronyme CALM (pour Causal Anticausal Linear Model) utilisé aussi en synthèse [D'Alessandro *et al.*, 2007].

Prédiction linéaire, phase et causalité

La partie dite de "phase ouverte" de l'ODG possède donc une évolution qualifiée d'anticausale. La phase augmente brusquement au niveau de la fréquence de résonance au lieu de chuter. Soit un signal réel $s(t)$ de spectre $S(\nu)$, alors le spectre du signal retourné en temps $s(-t)$ est le conjugué du spectre original, soit $S^*(\nu)$.

Or la prédiction linéaire est basée sur une estimation de l'énergie du spectre. Quelles que soient les composantes convoluées d'un signal $s(t) = [a * b * c](t)$, l'estimateur de la corrélation

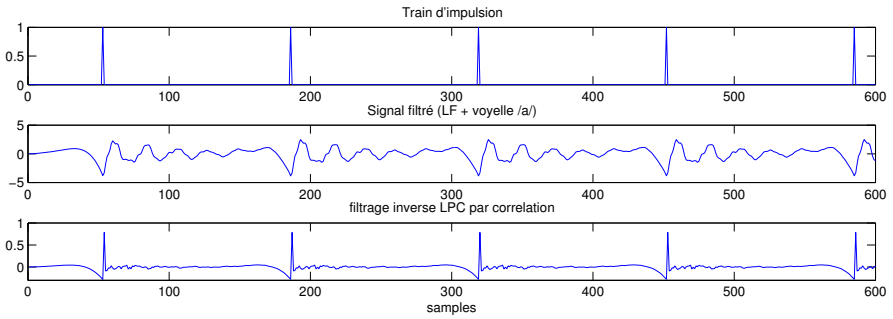


Figure 2.8 – Illustration du défaut d'estimation des phases anti-causales par la prédiction linéaire. Le signal résiduel n'est pas un train d'impulsion synchrone, mais est déphasé.

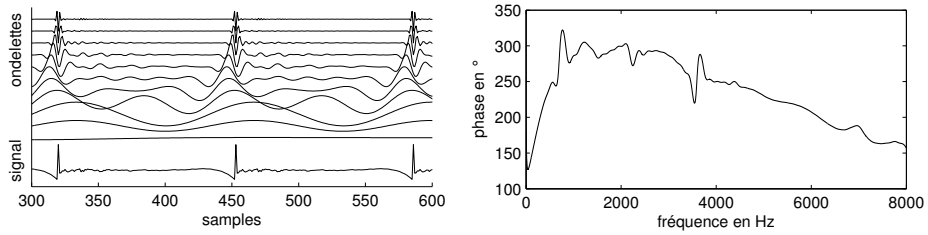


Figure 2.9 – Analyse du signal résiduel de la figure 2.8 par les ondelettes (sur la gauche) et son spectre de phase (sur la droite). Un maximum du déphasage est observé vers 2000Hz.

est la transformée de Fourier inverse de $S(\nu)S^*(\nu) = A(\nu)B(\nu)C(\nu)A^*(\nu)B^*(\nu)C^*(\nu)$. Ainsi, la corrélation (il en est de même pour la covariance) généralement utilisée pour estimer les filtres par prédiction linéaire ne distingue pas le sens de progression du temps et ce, même si les composantes convoluées du signal ont des sens de progression différents. On s'attend donc à ce que le résidu de l'analyse par prédiction linéaire ne rende pas exactement à l'identique le signal source.

Cette propriété est illustrée sur la figure 2.8. Le train d'impulsions (en haut) a été filtré par un modèle LF et la réponse impulsionnelle d'une voyelle /a/. Le signal de voix synthétisé est donné au milieu de la figure. On constate que le résidu de l'analyse par LPC, donné en bas de la figure 2.8, n'est pas exactement égal au train d'impulsions initial, malgré la synthèse strictement autorégressive utilisée pour générer le signal. On constate aussi que le résidu possède la réponse très caractéristique d'un retard de groupe assez prononcé en basse fréquence. Cette observation est confirmée par l'analyse du spectre de phase et de la réponse à un banc de filtre en ondelettes sur la figure 2.9. Remarquons que ceci est vrai pour tout modèle possédant une progression anticausale. La source glottique étant un système génératif, l'hypothèse d'une phase anticausale sur l'onde de débit glottique est totalement cohérente vis à vis de la réalité physique du système.

Une méthode couramment utilisée pour le filtrage inverse des signaux vocaux est l'utilisation de la préaccentuation. Bien que l'argument justifiant cette méthode est de blanchir le spectre, on s'aperçoit surtout que l'avantage de la préaccentuation est de "gommer" le formant glottique pour que l'analyse par LPC n'estime pas de pôles à sa proximité. Cette méthode est efficace, mais uniquement dans le cas où ce formant est effectivement distinguable des formants vocaliques - et plus bas en fréquence. Il s'agit donc d'un *a priori* fort sur le signal et ses caractéristiques.

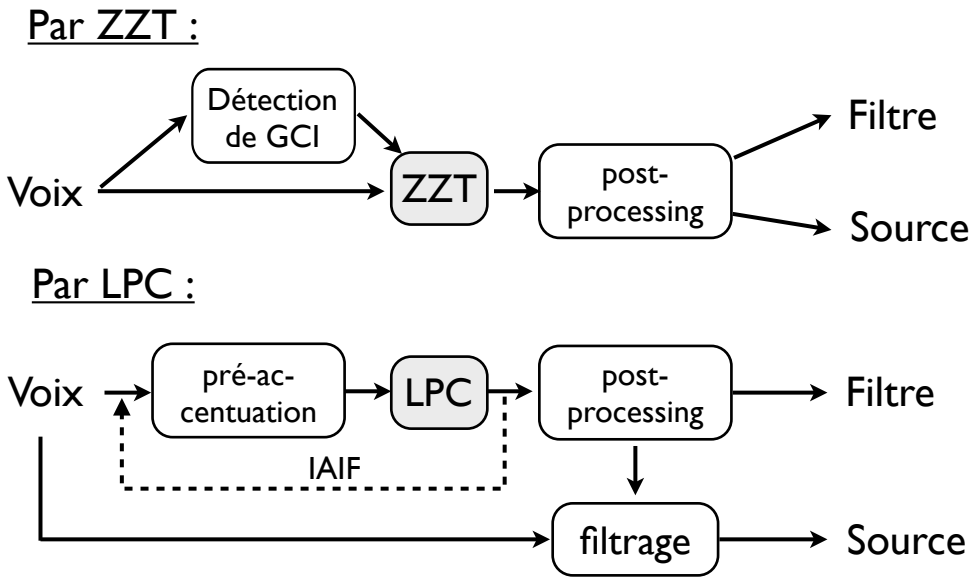


Figure 2.10 – Comparaison des étapes de filtrage inverse par LPC et ZZT. La ZZT utilise moins d'*a priori* avant et après l'estimation.

Avantages de l'approche CALM

Les méthodes de prédiction linéaire (LPC) reposent sur des *a priori* forts à la fois sur la forme du signal de source et sur la forme du filtre du conduit vocal. En effet, d'un côté le modèle imposé pour le filtre est un modèle autorégressif, de l'autre côté l'excitation (le résidu) est censé être une impulsion après pré-accentuation. De tels *a priori* permettent de stabiliser et de corriger les défauts naturels de la prédiction linéaire. On peut citer par exemple le nombre de formants par bande de fréquence ou la préacccentuation du signal. L'ajustement du filtre de préaccentuation fait d'ailleurs partie intégrante de la méthode de filtrage inverse IAIF proposée par Alku [Alku, 1992].

Le modèle CALM tend à s'affranchir de ces *a priori* en permettant une séparation basée sur la seule progression de la phase de chaque contribution. En conséquence il ne contraint pas la forme du filtre vocalique, et devrait donc permettre indirectement d'estimer des filtres vocaliques complexes - comme des voyelles nasalisées - possédant notamment des zéros d'anti-résonances. Ce paradigme permet donc la déconvolution directe du signal acoustique vocal sans ajustement. En pratique, une analyse préalable du signal sera nécessaire afin de déterminer des instants caractéristiques dans le signal. Ces instants - les GCI - offrent une origine des temps locale nécessaire et suffisante pour faire le séparation causale / anticausale. La différence entre avec le filtrage inverse par LPC réside donc dans les informations *a priori* sur la décomposition : d'un côté un filtre autorégressif + pré-accentuation, de l'autre la connaissance des GCI. Le détail de la mise en pratique de l'analyse sera vu à la section suivante.

Sur la figure 2.10 sont illustrées les différentes étapes avant et après estimation pour la prédiction linéaire (incluant le cas de la méthode IAIF) ainsi que pour le modèle CALM par la ZZT. On remarque que le signal n'est pas modifié en amont de l'analyse par ZZT alors qu'il est pré-accentué pour l'estimation du filtre par LPC.

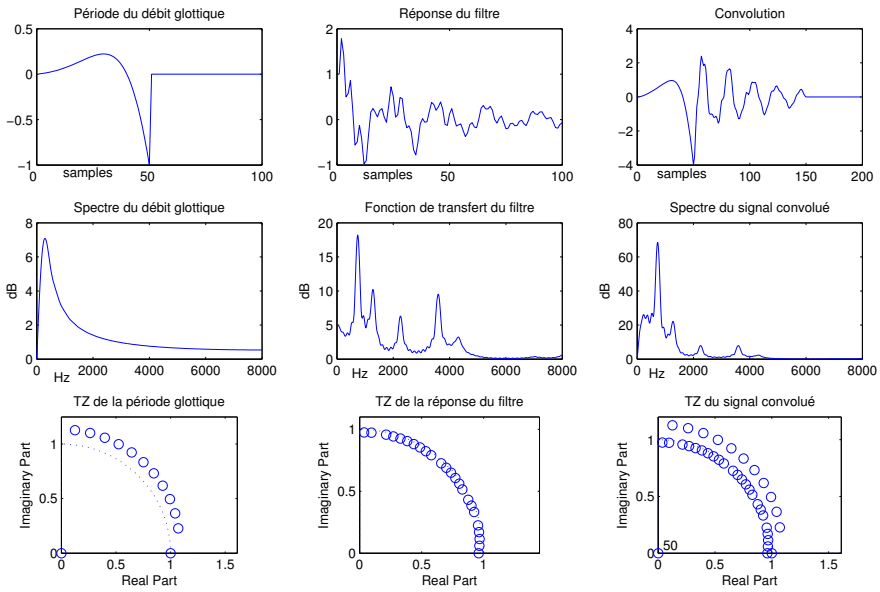


Figure 2.11 – Illustration du paradigme de la ZZT : Les signaux temporels (première ligne) de la source (première colonne) et de la réponse du filtre (deuxième colonne) sont convolués (troisième colonne) selon le modèle de production linéaire de la parole. En deuxième rang on peut visualiser leurs spectres d'énergie et enfin le troisième rang représente comment les zéros sont combinés par convolution.

2.2.4 La théorie ZZT

On pourrait naturellement penser qu'une prédiction linéaire convenablement mise en oeuvre pourrait fournir des pôles de part et d'autre du cercle unité, et donc permettre l'application directe du modèle CALM. Mais comme il a été présenté précédemment, les méthodes de prédiction linéaire reposent sur des estimateurs indépendants du sens de progression du temps, et contraignent donc les pôles à être à l'intérieur du cercle unité - hors instabilité de calcul.

L'idée est donc apparue de séparer les composantes d'un signal à partir des zéros calculés sur ce signal. La ZZT se base donc sur le calcul des zéros de la transformée en Z d'un segment de signal. Le travail de thèse de Baris Bozkurt [Bozkurt, 2005] présente ce paradigme pour la décomposition source-filtre et l'estimation des formants glottiques et vocaliques. Sur la figure 2.11 est présentée une illustration du principe de déconvolution par ZZT. Les signaux temporels (première ligne) de la source (première colonne) et de la réponse du filtre (deuxième colonne) sont convolués (troisième colonne) selon le modèle de production linéaire de la parole. En deuxième rang de la figure 2.11 on peut visualiser leur spectre d'énergie et enfin le troisième rang représente comment les zéros sont combinés par convolution. La ZZT cherche donc à réaliser le processus inverse : obtenir les formes d'onde à partir de la séparation des zéros. L'algorithme de décomposition tel que présenté dans la thèse de B. Bozkurt ne propose cependant pas de validation approfondie sur des signaux réels.

Des travaux récents [Drugman *et al.*, 2009b] sur le sujet ont montré que le paradigme de la ZZT pouvait être transcrit (avec des résultats identiques) dans le domaine du cepstre complexe. Ces travaux critiquent et commentent abondamment la forme de la fenêtre d'analyse nécessaire à une décomposition ZZT cohérente - prochaine section -, et montrent le parallèle entre ZZT

et cepstre complexe. La présente étude se limitera cependant sur la ZZT présentée par B. Bozkurt dans la mesure où la principale différence entre ZZT et cepstre complexe (CCD : *Complex Cepstrum Decomposition*) réside dans la charge de calcul à l'avantage de CCD.

Choix de la forme de la fenêtre d'analyse

Afin de pouvoir exploiter les zéros de la transformée en Z du signal, il est nécessaire de convenablement choisir la forme de la fenêtre d'analyse [Bozkurt *et al.*, 2004a]. Une fenêtre de forme exponentielle $e^{-a|t|}$ va permettre de favoriser le placement des zéros d'un côté ou de l'autre du cercle unité en fonction du placement par rapport à l'instant de fermeture glottique, mais va introduire une distorsion dans le spectre résultant. Pour la décomposition source filtre, les expériences ont montré que la fenêtre de Blackman permet de rester au plus près de la forme originale tout en permettant une séparation efficace des zéros. Dans les publications récentes traitant de la ZZT, cette fenêtre fait office de référence [Drugman *et al.*, 2008, Drugman *et al.*, 2009a, Sturmel *et al.*, 2007].

Cependant, les travaux récents de Drugman et al. [Drugman *et al.*, 2009b] ont mis en lumière l'impact du choix de la forme de la fenêtre sur la qualité de la décomposition par ZZT de manière plus précise. Ainsi pour les fenêtres de taille N de la famille donnée par l'équation 2.6, on retrouve les fenêtres de Hann et Blackman pour le coefficient α de valeur 1 et 0.84 respectivement mais ne sont pas optimales. La valeur optimale de α pour une décomposition ZZT à partir d'une fenêtre de taille $2T_0$ est $\alpha = 0.75$ selon [Drugman *et al.*, 2009b], mais ces résultats n'ont pas été systématiquement confirmés par les expériences, notamment celles présentées au chapitre 5.

$$w(t) = \frac{\alpha}{2} - \frac{1}{2}cos(\frac{2\pi n}{N-1}) - \frac{1-\alpha}{2}cos(\frac{4\pi n}{N-1}) \tag{2.6}$$

Choix de la taille de la fenêtre d'analyse

Forme et largeur de la fenêtre d'analyse sont liées pour donner une décomposition ZZT de qualité. Selon le travail de Bozkurt, pour une fenêtre de Blackman, l'analyse se fait sur une largeur de deux périodes, mais toujours centrée sur l'instant de fermeture glottique. Dans le cas où les deux périodes ont des durées différentes, la taille de la fenêtre retenue est le double de la plus petite des deux périodes. L'utilisation de fenêtres asymétriques se révélerait problématique dans la mesure où on perdrait la linéarité en phase du processus de fenêtrage.

Position de la fenêtre d'analyse

Enfin, l'emplacement de la fenêtre d'analyse est d'une importance prépondérante. A l'instant de fermeture glottique, une impulsion acoustique va exciter le conduit vocal ; cet instant est aussi la transition entre la phase ouverte et la phase fermée du débit glottique. Baris Bozkurt [Bozkurt *et al.*, 2005] a montré que l'ordonnancement des zéros était beaucoup plus stable lorsque l'analyse était placée autour d'un instant de fermeture glottique.

Les travaux récents de Drugman et al. [Drugman *et al.*, 2009a] ont proposé une opération a postériori sur les zéros pour détecter et corriger une erreur de position de la fenêtre d'analyse. Si les résultats obtenus sont encourageants, il ne sera pas tenu compte de ce travail dans le présent chapitre, la détection des instants de fermeture glottique n'étant pas un problème grâce à la présence de données électroglottographiques pour les signaux analysés dans ce chapitre.

Le problème de la position des zéros et de leur déplacement vis-à-vis de l'instant de fermeture glottique a aussi été observé par Daalsgard et al. [Dalsgaard *et al.*, 2008].

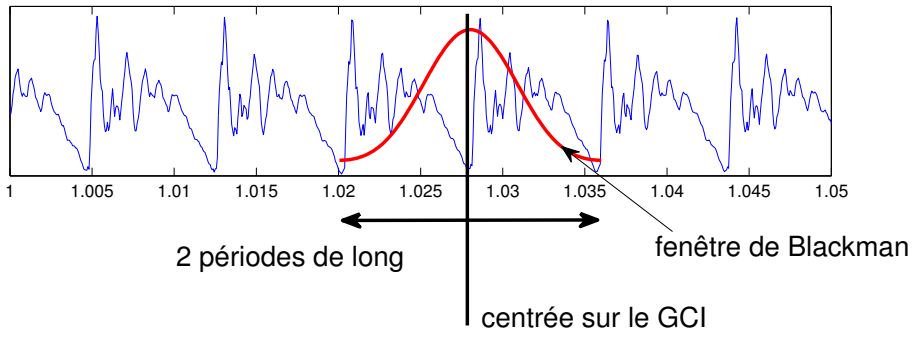


Figure 2.12 – La forme, position et taille de la fenêtre d'analyse utilisée en ZZT.

Calcul des zéros du polynôme de la transformée en Z

A partir des échantillons a du signal d'analyse fenêtré, on compose un polynôme P dont chaque coefficient correspond à un échantillon, dans l'ordre chronologique des puissances croissantes (coefficient $a(i)$ pour la puissance $-i$). Dans des conditions de parole naturelle, il est courant de calculer des racines de polynômes d'ordre 200 ou supérieurs, nécessitant donc une résolution numérique.

$$\begin{aligned} P(z) &= \sum_{i=0}^{N-1} a(i) z^{-i} \\ P(z) &= \frac{1}{z^{N-1}} \sum_{i=0}^{N-1} a(i) z^{N-1-i} \\ P(z) &= \frac{1}{z^{N-1}} \hat{P}(z) \end{aligned} \tag{2.7}$$

Pour déterminer les zéros de ce polynôme, nous utilisons la fonction "roots" de MATLAB© . Cette fonction est basée sur la diagonalisation de la matrice dite "associée" au polynôme[1]. Actuellement, c'est l'algorithme le plus stable permettant de calculer les zéros pour des polynômes de très grand ordre. La matrice associée au polynôme $P(z)$ de l'équation 2.7 s'exprime :

$$\mathbf{H} = \begin{bmatrix} 0 & 0 & \dots & 0 & -a(0)/a(n) \\ 1 & 0 & \dots & 0 & -a(1)/a(n) \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \ddots & 0 & -a(n-2)/a(n) \\ 0 & 0 & \dots & 1 & -a(n-1)/a(n) \end{bmatrix}$$

On notera que la forme de la matrice $\mathbf{H}$ est de type Hessenberg (triangulaire supérieure à laquelle on ajoute une série de valeurs sous la diagonale ou inversement), ce qui rend le processus de diagonalisation beaucoup plus simple. Les valeurs propres de cette matrice sont donc les solutions de l'équation $P(x) = 0$, les racines du polynôme P. On obtient alors les zéros z_i par factorisation, tels que :

$$P(z) = \frac{1}{z^{N-1}} \prod_{i=0}^{N-1} (z - z_i)$$

1. compagnon matrix [Werner, 1983]

La partie la plus gourmande en ressources réside dans le calcul des zéros. Cet algorithme est lourd, à la fois en ressources mémoire et en temps de calcul : c'est un algorithme de complexité $O(n^2)$. Par le biais d'une programmation en parallèle (chaque période analysée étant indépendante des autres), il est possible de tirer partie des nouvelles architectures de micro-processeurs. Dans une implémentation optimisée, on arrive peu à peu à approcher un temps de calcul équivalent au temps réel, ce qui peut ouvrir de nouvelles perspectives d'applications pour la ZZT, et atteindre des performances qui étaient jusque là réservées à l'analyse par LPC : analyse de grands corpus, décomposition en temps réel, analyse/synthèse...

Les avancées sur l'utilisation du cepstre complexe pour procéder à une décomposition causale/anticausale permettent une décomposition encore plus rapide. Ce point sera vu ultérieurement (2.2.8).

Déplacement de zéros

Dans l'algorithme original une étape de la décomposition par ZZT consiste en un ajustement du nombre de zéros de part et d'autre du cercle unité sur l'axe réel (fréquence nulle). Cette étape, qui donne des résultats plus stables d'une période à l'autre, semble être liée à la pente spectrale de la source glottique. En effet, rajouter un zéro sur l'axe réel consiste à appliquer un filtre linéaire du premier ordre au signal. L'expérience montre que cette étape est nécessaire pour compenser un défaut de la ZZT, qui n'estime pas toujours les composantes causales et anticausales avec le bon ordre de dérivation/intégration : une dérivation ou une intégration revient à rajouter un zéro sur l'axe réel, proche de 1, respectivement à l'intérieur et l'extérieur du cercle unité.

Ce défaut pourrait se transformer en avantage, si le placement de ce zéro se révèle effectivement lié au terme de rayonnement du modèle de production vocale, son estimation correcte mènerait à une estimation correcte de ce rayonnement.

2.2.5 Algorithme de la décomposition par ZZT

L'algorithme de la ZZT est présenté sur la figure 2.13. Sur la droite de la figure sont illustrées les différentes étapes de l'algorithme. Les analyses présentées sont réalisées sur un signal réel : une voyelle /a/ prononcée à 130Hz environ par un locuteur masculin.

2.2.6 ZZT et Filtrage Inverse

Ce n'est pas l'opération de calcul des zéros en elle même qui donne la décomposition source/filtre, mais le fait de séparer ces zéros. Si le signal s est factorisé en zéros z_i dans le domaine spectral tel que :

$$\sum_{i=0}^{N-1} s(i)z^i = \prod_{i=1}^{N-1} (z - z_i)$$

Alors enlever un zéro z_j revient à pratiquer le produit spectral suivant :

$$\frac{\prod_{i=1}^{N-1}(z - z_i)}{(z - z_j)}$$

Ainsi, l'opération de séparation de zéros revient à une déconvolution par un filtre donc l'expression est imposée par le modèle : ce filtre est composé des zéros intérieurs au cercle unité. Retrouver l'expression du débit glottique en filtrant le signal par un filtre dont la forme est imposée par un modèle, revient à faire un filtrage inverse. La différence entre le filtrage inverse par prédiction linéaire et le filtrage inverse par ZZT réside donc dans la manière dont le filtre est estimé, et en particulier la forme qui lui est imposé.

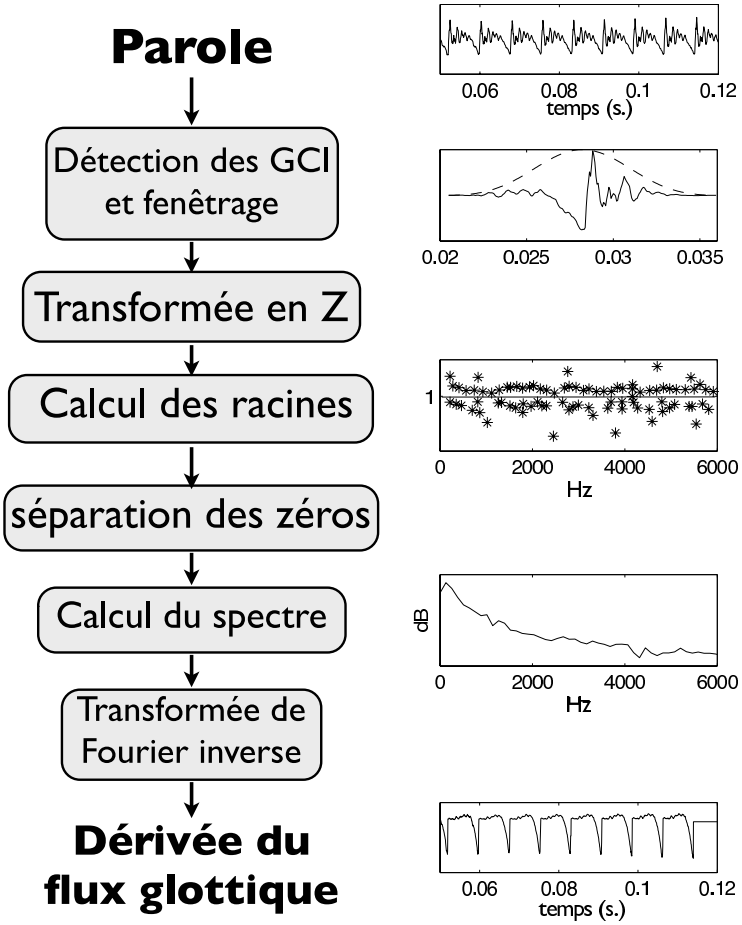


Figure 2.13 – Algorithme de la décomposition source filtre par ZZT, point par point.

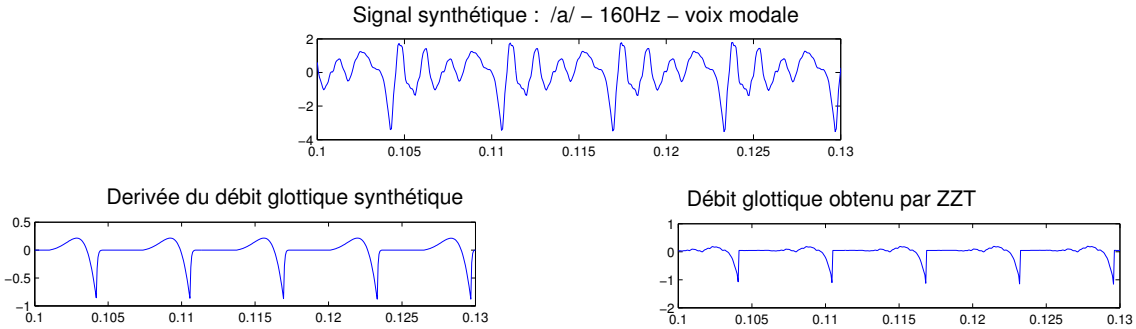


Figure 2.14 – Exemple de décomposition par ZZT sur un signal synthétique (/a/, 160 Hz, voix modale - $0_q = 0.5; \alpha_m = 0.8$).

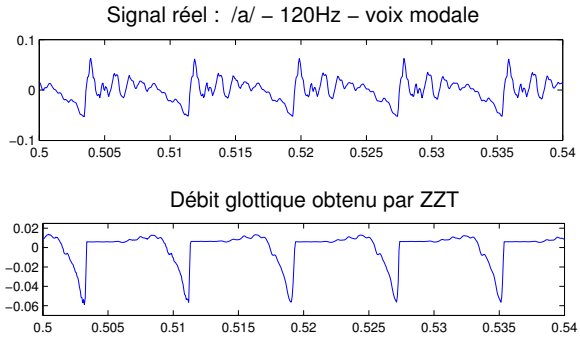


Figure 2.15 – Exemple de décomposition par ZZT sur un signal réel (/a/, 120 Hz, voix modale - $0_q \approx 0.5; \alpha_m \approx 0.8$).

2.2.7 Quelques exemples

Avec uniquement les instants de fermeture glottique et sans *a priori* sur la forme du filtre il est possible de séparer efficacement source et filtre sur un signal synthétique. Un exemple est présenté sur la figure 2.14. Sur cette figure est la décomposition d'une voyelle /a/ synthétisée à la fréquence fondamentale de 160Hz pour des paramètres de voix modale ($0_q = 0.5; \alpha_m = 0.8$) on remarque que la décomposition est fidèle à l'exception du "lobe" positif de la dérivée du débit glottique qui se retrouve tassé par décomposition ZZT. Heureusement, nous verrons par la suite que ce genre de distorsion temporelle n'a que très peu d'impact dans le domaine de Fourier.

De toute évidence, l'application à la parole naturelle (dans des conditions réelles d'enregistrement) n'est pas aussi simple, en particulier du fait de la présence du rayonnement des lèvres qui ne suit pas la forme d'un dérivateur parfait. Un exemple est présenté sur la figure 2.15 pour une voix modale d'un locuteur masculin prononçant un /a/ à une fréquence fondamentale d'environ 120Hz. La qualité de la décomposition est cependant suffisante pour y visualiser le quotient ouvert de 0.5 fourni par les données EGG. Cependant, la mesure automatisée semble problématique.

Cela dit, des travaux réalisés par [Bozkurt *et al.*, 2004b, Drugman *et al.*, 2008, Sturmel *et al.*, 2006] ont déjà permis de montrer que la ZZT était une méthode convenable pour l'estimation des paramètres du filtre (suivi de formants) et de la source (estimation de paramètres) :

– Les travaux de Baris Bozkurt [Bozkurt *et al.*, 2004b] ont mis en lumière l'applicabilité

du paradigme du modèle causal-anticausal pour le suivi de formants avec davantage de précision que la prédiction linéaire, en particulier pour le suivi du premier formant.
- Les travaux de Thomas Drugman et al. [Drugman *et al.*, 2008] ont démontré par validation perceptive que l'estimation des paramètres de la source glottique était rendu possible grâce à la combinaison du paradigme du modèle causal-anticausal et de la prédiction linéaire par un ajustement non linéaire.
- Nos travaux réalisés pour la conférence AQL [Sturmel *et al.*, 2006] ont permis de montrer qu'un pavage dans le plan de variation OQ-AlphaM permettait aussi de lier la forme d'onde glottique estimée par ZZT, la fréquence du formant glottique ainsi que la valeur du quotient ouvert pour approcher une valeur de l'asymétrie correspondant au modèle. Le chapitre 5 présentera une méthode alternative d'estimation de l'asymétrie du débit glottique.

Mais parmi toutes ces études, aucune ne s'est intéressée à comparer l'estimation du débit glottique par ZZT à celui obtenu par prédiction linéaire. Les méthodes contraintes (qui imposent une forme paramétrique tant au filtre qu'à la source) vues par la suite reposent toutes sur une modélisation auto-régressive du conduit vocal, et un tel travail pourrait montrer que la modélisation causale/anticausale est utilisable dans les même contextes d'estimation contrainte.

2.2.8 Décomposition causale/anticausale par Cepstre Complexe

L'utilisation du cepstre complexe permet de réaliser une séparation causal/anti-causale à la manière de la ZZT. Ces travaux ont été publiés par Drugman et al. dans [Drugman *et al.*, 2009b]. Soit $\hat{x}(n)$ le cepstre complexe de $x(n)$ de taille N, on le définit selon l'équation 2.8.

$$\begin{aligned} X(\omega) &= TF(x(n)) \\ log[X(\omega)] &= log|X(\omega)| + \Phi[X(\omega)] \\ \hat{x}(n) &= \frac{1}{2\pi}\int_{-\pi}^{\pi}[log(X(\omega))]e^{j\omega n}d\omega \end{aligned} \tag{2.8}$$

Dans [Steiglitz et Dickinson, 1977] il est montré que le cepstre complexe peut être factorisé par la transformée en Z de telle manière que les premiers éléments de $\hat{x}(n)$ (en temps négatif) appartiennent à la partie anticausale et que les éléments suivants (en temps positif) appartiennent à la partie causale. Ainsi, les parties anticausales ac et causales c de $\hat{x}(n)$ sont déduites par les équations 2.9 et 2.10.

$$ac(n) = \hat{x}(n)|_{n<N/2} \text{ et } ac(N/2) = \frac{\hat{x}(N/2)}{2} \tag{2.9}$$

$$c(n) = \hat{x}(n)|_{n>N/2} \text{ et } c(N/2) = \frac{\hat{x}(N/2)}{2} \tag{2.10}$$

On retrouve les mêmes problèmes que pour la transformée en Z : une perte du facteur d'échelle de chaque partie décomposée, ainsi qu'une inconnue sur la polarisation de chaque partie décomposée. La décomposition vue sous cet angle permet tout de même un calcul plus rapide pour une stabilité numérique légèrement supérieure. Dans la suite de cette étude, la décomposition par cepstre complexe ne sera pas utilisée mais peut être considérée comme interchangeable avec la ZZT.

2.2.9 Vers l'estimation des paramètres du modèle de source

Estimation directe sur l'onde de débit glottique

La première approche est d'utiliser les signaux obtenus par filtrage inverse pour déterminer les paramètres du débit glottique. De multiples études se sont penchées sur le problème (notamment [Alku, 2003, Riegelsberger et Krishnamurthy, 1993, Lehto *et al.*, 2007]), mais il est compliqué d'utiliser de manière systématique les ondes de débit glottique déterminées par prédiction linéaire.

Estimations contraintes ARX

Le filtrage inverse par estimation du filtre du conduit vocal n'utilise pas l'intégralité des connaissances dont nous disposons sur le modèle de production. Aucune information sur le modèle de la source n'y est utilisée. On peut donc chercher à contraindre l'estimation en injectant un modèle de source en plus du modèle du filtre. C'est l'idée de la modélisation ARX (pour *AutoRegressive eXogenous*).

La méthode ARX-LF [Vincent *et al.*, 2007] utilise comme information injectée à l'estimation, le modèle LF de l'ODG. Cette modélisation est présentée sur l'équation 2.11 où $s(n), a_k, g(n)$ et $\epsilon(n)$ sont respectivement le signal, les coefficients du filtre autorégressif, l'onde de débit glottique et l'erreur de prédiction, procède à une optimisation non linéaire des paramètres de la source sur le résidu du filtrage inverse par LPC. Comme le problème d'estimation est non linéaire, le calcul n'est pas formalisé analytiquement. Une approche permettant d'atteindre le minimum d'erreur pour l'estimation consiste à se baser sur un dictionnaire prédéterminé de paramètres de source glottique θ, et à chercher l'item θ_{min} du dictionnaire présentant l'erreur la plus basse pour les variations de θ et a_k (équation 2.12) . On peut ensuite affiner l'estimation par itérations successives autour du jeu de paramètres trouvés précédemment.

$$s(n) = -\sum_{k=1}^{P} a_k s(n-k) + g_\theta(n) + \epsilon(n) \tag{2.11}$$

$$s(n) = \tilde{s}_{\theta,a_k}(n) + \epsilon(n)$$

$$\theta_{min} = \underset{\theta,a_k}{\operatorname{argmin}}\, \epsilon(n) \tag{2.12}$$

Le principe de la minimisation est très intéressant du point de vue de l'analyse-synthèse [Agiomyrgiannakis et Rosec, 2009, Audibert *et al.*, 2006] de la parole expressive. Cependant, même si les valeurs des paramètres de source rendues sur des signaux synthétiques restent fidèles [Vincent *et al.*, 2005], les résultats sur signaux réels sont systématiquement décrits du point de vue global (analyse et resynthèse, ou discussion sur le résidu) et sur peu d'exemples. Cette méthode se révèle cependant efficace dans la détection des GCI [Vincent *et al.*, 2006].

Le principal désavantage de cet algorithme est qu'il contraint à la fois la forme du filtre (autorégressive) et la forme de la source (LF) au cours de l'estimation. Du point de vue d'une représentation compacte des signaux, de telles contraintes permettent de décrire de manière optimale le signal : le résidu de l'estimation est d'ailleurs très faible avec ARX-LF.

2.2.10 D'autres critères à signification perceptive

La difficulté d'estimation et de modélisation de la source glottique a mené vers des approches empiriques de l'établissement du lien entre paramètres de la source glottique et qualité vocale,

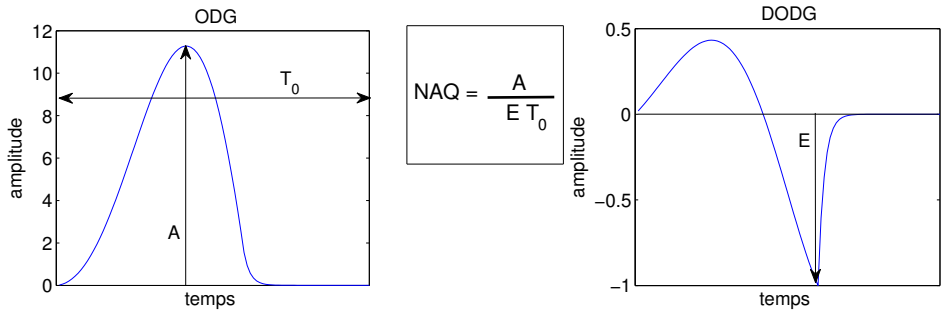


Figure 2.16 – Calcul du quotient d'amplitude normalisé sur l'onde de débit glottique. Le calcul se fait traditionnellement sur l'ODG et la DODG obtenues par filtrage inverse LPC sur le signal vocal.

compte tenu des propriétés spectrales de la source [Henrich *et al.*, 2001]. On trouve dans la littérature deux indicateurs extraits du résidu d'analyse par LPC du signal vocal. D'une part, sur le signal de l'onde de débit glottique et de sa dérivée : le quotient d'amplitude (généralement normalisé [Alku et Bäckström, 2002]) et d'autre part, sur le spectre du signal vocal. On trouve aussi un paramètre réduit du conduit vocal, R_d définit par Fant [Fant, 1995] comme une fonction des trois paramètres de forme O_q, α_m et Q_a.

Ces indicateurs se sont révélés très fortement corrélés à la configuration glottique et notamment à la valeur du quotient ouvert, mais tout en étant sensiblement dépendant des autres paramètres, notamment l'asymétrie.

Le quotient d'amplitude normalisé (NAQ)

Proposé par Alku [Alku et Bäckström, 2002], le quotient d'amplitude normalisé est le rapport entre l'amplitude de l'ODG et le minimum de sa dérivée, normalisé par la période fondamentale. Ce calcul est illustré sur la figure 2.16.

$$NAQ = \frac{A}{ET_0}$$

Les résultats montrent que ce quotient d'amplitude est fortement corrélé à la phase fermée du cycle glottique $(1 - O_q)$ et constitue un bon descripteur de la qualité vocale en terme de clarté de la phonation d'après des résultats statistiques sur un corpus de parole réelle. En réalité, on ne peut exclure une dépendance entre NAQ et α_m, ce qui rend ce paramètre hybride vis-à-vis du modèle LF de la source glottique, comme montré dans [Doval *et al.*, 2006] et illustré par la figure 2.17. Le quotient d'amplitude a notamment été utilisé pour l'analyse de l'effort vocal [Alku *et al.*, 2006].

La différence d'amplitude entre harmoniques

L'analyse du spectre de la source glottique a permis d'acquérir deux connaissances fondamentales :

1. La fréquence ainsi que la bande passante du formant glottique dépendent de la valeur du quotient ouvert (O_q) ainsi que de l'asymétrie (α_m).
2. La fréquence de ce formant glottique est comprise grossièrement entre $0.8F_0$ et $1.5F_0$.

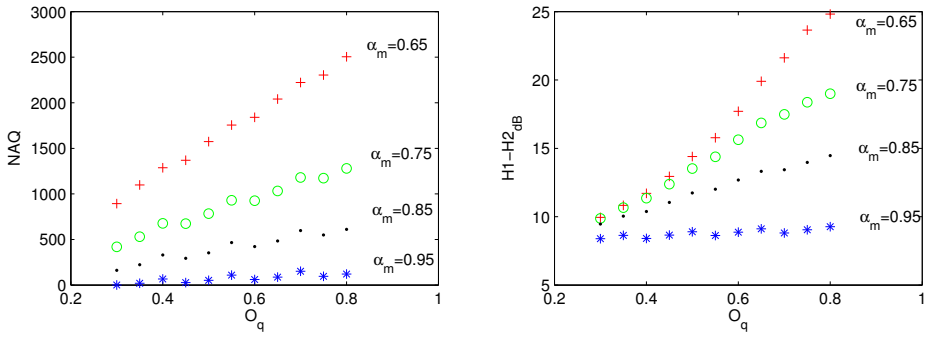


Figure 2.17 – Variation de la valeur des paramètres NAQ et H1H2 pour différentes valeurs de O_q et α_m. Tests sur signaux synthétiques par le modèle LF.

Les premier et deuxième harmoniques du signal se trouvent donc toujours dans le lobe résonant du formant glottique. Le calcul du rapport entre l'amplitude de ces harmoniques (ou la différence de leur logarithme) permet donc en théorie d'établir une valeur proche de la valeur du quotient ouvert associé à la qualité vocale prise en considération.

Dans sa thèse, Helen Hanson [Hanson, 1995] a pu démontrer que le rapport entre les deux premiers harmoniques avec correction des formants (H1*-H2* [Hanson, 1994]) était fortement corrélé à la qualité vocale associée au signal. Dans une première approximation, on peut même lier ce rapport entre harmoniques à la valeur du quotient ouvert [Henrich *et al.*, 2001]. Cependant, on ne peut exclure une dépendance du rapport H1-H2 à la valeur de l'asymétrie, à l'instar du quotient d'amplitude normalisé vu précédemment.

La figure 2.17 donne une idée de la variation des paramètres NAQ et H1-H2 pour différentes valeurs de O_q et α_m. Ces données ont été obtenues pour des signaux synthétiques d'onde de débit glottique par modèle LF.

Le paramètre réduit R_d

Le paramètre R_d [Fant, 1995] propose la réduction à une dimension des paramètres de forme du modèle LF (O_q, α_m, Q_a). Selon l'expression donnée par Fant lui même, ce paramètre R_d a donné naissance à NAQ, il est proposé avec l'expression de l'équation 2.13 - équation 1 de [Fant, 1995] adaptée aux notations du document.

$$R_d = \frac{A}{E}\frac{F_0}{110} \tag{2.13}$$

Ce paramètre limite les formes possibles du débit glottique, mais permet aussi de rendre plus cohérente la variation d'un tel paramètre. Il a été récemment montré que R_d est efficace pour l'analyse/synthèse des signaux vocaux dans une dimension serré / lâché proche de la dimension portée par le quotient ouvert [Degottex *et al.*, 2010]. Cependant, ce paramètre ne permet pas un éventail large d'analyse et de synthèse de débits glottiques et nécessite de nombreuses hypothèses sur la valeur du trio (O_q, α_m, Q_a). Ainsi, dans l'effort de qualifier au plus près les rapports entre qualité vocale et configuration glottique, le présent travail ne considérera pas ce paramètre réduit mais pourra servir à affiner les hypothèses liées à son utilisation.

2.3 Périodicités, Apériodicités

Le modèle de la voix ne serait pas complet sans prendre en compte la possibilité de générer du bruit lors de la phonation. L'origine du bruit peut être diverse, mais porte souvent des informations importantes du point de vue de la qualité vocale ou de l'information phonétique. L'ensemble du bruit généré est désigné par le terme "apériodicités" en opposition à une voix purement périodique, dont le spectre serait un spectre harmonique de raies.

2.3.1 Origine des apériodicités

Les apériodicités de voix possèdent deux origines distinctes. D'une part, les apériodicités dites structurelles proviennent de la non périodicité de la vibration des plis vocaux, causant les phénomènes de jitter $jitt(i)$ et de shimmer $sh(i)$, définis localement sur une période. D'autre part, la présence de turbulences dans le conduit vocal, notamment lors de la fermeture glottique ou de la présence de constrictions (consonnes), amène l'apparition d'un bruit à phase aléatoire $b(t)$. Ainsi, on peut adapter le modèle linéaire classique afin qu'il tienne compte de ces micro-perturbations comme sur l'équation 2.14.

$$s(t) = f(t) * [g(t) * \sum_{i=-\infty}^{\infty} (1 + sh(i))\delta(t - (1 + jitt(i))\frac{i}{F_0}) + b(t)] \tag{2.14}$$

Apériodictés structurelles

Causées par des non linéarités au niveau des plis vocaux, elles causent des variations de la fréquence fondamentale, modifiant par la même occasion la quantité d'énergie transmise au niveau de la glotte.

On décompose ces perturbations selon deux dimensions : les perturbations d'ordre temporel sont appelées Jitter alors que les perturbations de l'amplitude du signal sont appelées Shimmer. On les lie souvent à l'état de santé des plis vocaux, mais aussi à l'effort exercé lors de la phonation. Les voix possédant de telles apériodicités sont souvent perçues comme "éraillées".

Bruits Additifs

Tout comme les apériodictés structurelles, le bruit additif est issu de phénomènes non linéaires et causé durant l'écoulement d'air de la glotte aux lèvres. Ces turbulences sont toujours présentes dans la voix, mais peuvent être amplifiées par la présence de constrictions sur le trajet du flux d'air. Ces turbulences créent un bruit aléatoire (aussi dénommé *stochastique*) coloré qui s'ajoute au signal de source comme sur l'équation 2.14.

Lorsque la voix est chuchotée, ces apériodicités stochastiques constituent l'intégralité du signal de source.

2.3.2 Décomposition périodique/apériodique des signaux de parole

Faire une distinction, et surtout une séparation voisée/ non voisée sur de la parole naturelle est extrêmement difficile. Au delà de la distinction *tout ou rien* faisant la distinction entre les signaux voisés (produits par une activité glottique) et les signaux non voisés, on trouve de nombreuses méthodes [Bachu *et al.*, 2010, Murty *et al.*, 2009, de Cheveigné et Kawahara, 2002] basées sur des propriétés différentes des signaux de parole. Cependant, dans le cadre de l'étude de cette thèse, il est souhaitable de produire une séparation voisé/non voisée lors de l'activité

glottique : c'est à dire séparer une contribution déterministe d'une contribution aléatoire. Afin de résoudre ce problème, on peut partir du premier principe que chaque contribution (voisée et non voisée) dispose de son propre domaine spectral [Kim et Hahn, 2007], mais on peut aller encore plus loin et tenir compte des propriétés pseudo-périodiques des signaux de parole pour utiliser un modèle harmonique + bruit ; notamment illustré par les modèles de voix sinusoïdaux [McAulay et Quatieri, 1986].

Après avoir présenté un modèle de voix périodique, ces deux méthodes seront illustrées : d'un côté la décision de la fréquence limite de voisement, et de l'autre les algorithmes de séparation voisé/non voisé.

Estimation du rapport entre harmoniques et bruit

Afin de pouvoir caractériser la qualité vocale, la connaissance du rapport énergétique entre harmoniques et bruit dans un signal de parole peut-être une information importante. Afin d'estimer cette valeur, Guus de Krom [de Krom, 1993] a proposé d'utiliser le cepstre réel [Childers *et al.*, 1977]. Dans le domaine cepstral les signaux périodiques présentent un pic d'énergie pour la quéfrence correspondant à la période fondamentale du signal. L'extraction de ce pic d'énergie permet de déterminer les points en fréquence comportant l'énergie de la partie harmonique afin de déterminer le rapport entre la partie harmonique du spectre et sa partie bruitée.

Cette méthode est toutefois sensible aux apériodicités structurelles. Une autre approche basée sur la transformée de Hilbert d'un banc de filtres du signal de parole a été présentée par Michaelis et al. [Michaelis *et al.*, 1997]. Le paramètre présenté permet de mesurer le niveau de bruit présent dans le signal indépendamment de la présence de jitter ou de shimmer.

Cependant, pour pousser plus loin l'analyse des parties périodique et apériodique il est bon de ne pas se limiter à la seule connaissance du niveau de bruit, mais de l'estimer comme une réalisation particulière d'un processus aléatoire dans le but d'observer, par exemple, sa distribution en temps ou en fréquence.

Modèle de la voix périodique

Dans une première approximation, les parties voisées peuvent être assimilées aux parties présentant une périodicité. La transformée de Fourier d'un signal périodique de période réduite $\frac{1}{\nu_0}$ observée sur un temps infini est donnée comme :

$$S(\nu) = \sum_{i=0}^{N} \delta(\nu - i\nu_0)A(i)$$

où $A(i)$ est l'amplitude complexe pour chaque harmonique et ν la fréquence réduite, comprise entre 0 et 1 avec $\nu = \frac{f}{F_e}$ où f est la fréquence réelle (en Hz) et F_e la fréquence d'échantillonnage du signal. Evidemment, il n'est pas physiquement possible d'observer un tel signal infini, c'est donc la procédure de fenêtrage qui permet l'observation et l'établissement du critère de sélection des parties périodique et apériodique dans un signal.

Soit la fenêtre d'oubli $w(n)$ de transformée de Fourier $W(\nu)$ de largeur N, alors la transformée de Fourier du signal devient :

$$S_w(\nu) = \sum_{i=0}^{N} W(\nu - i\nu_0)A(i)$$

Chaque harmonique $A(i)$ est donc étalé spectralement par la fenêtre W. Le choix de cette fenêtre d'analyse est crucial, car son étalement conditionnera la résolution spectrale des harmoniques

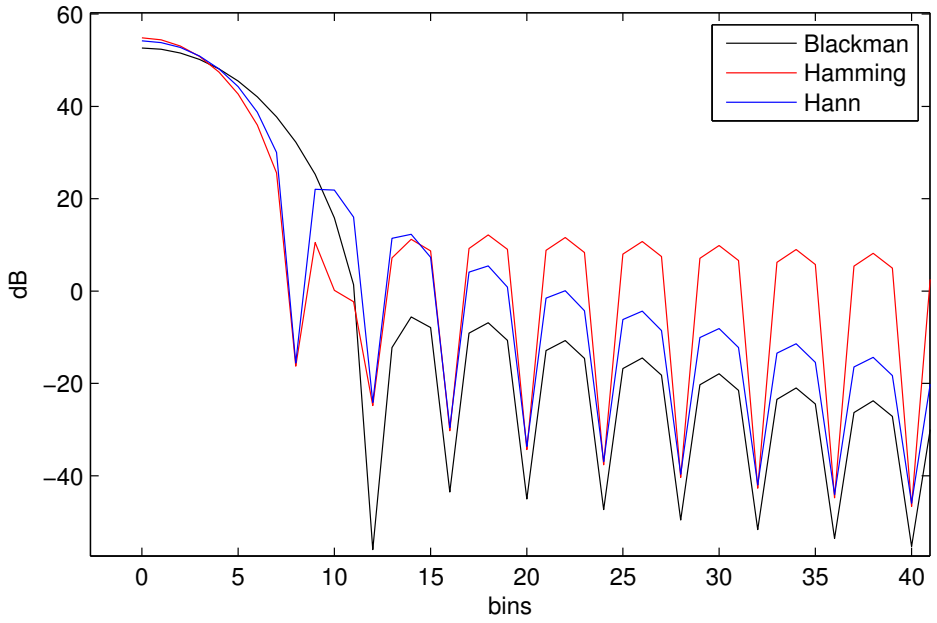


Figure 2.18 – Transformée de Fourier discrète de quelques fenêtres d'analyse, fenêtre de 1024 points, transformée sur 4096 points par complément de zéros. Seuls les bins de 0 à 40 sont représentés.

observés. Ainsi, dans le cadre des fenêtres d'analyse conventionnelles, la largeur du lobe principal ainsi que la puissance des lobes secondaires est à prendre en compte [Harris, 1978]. La transformée de Fourier de quelques fenêtres est présentée sur la figure 2.18. On considère les trois fenêtres les plus utilisées et décrites dans [Blackman et Tukey, 1959] :

- La fenêtre de Hann : $w(n) = cos(\frac{\pi n}{N-1})^2$, possède un lobe principal étroit avec une forte atténuation des lobes suivants, mais un lobe secondaire assez haut en puissance.
- La fenêtre de Hamming : $w(n) = 0.54 + 0.46cos(\frac{\pi n}{N-1})$, possède un lobe principal étroit, un lobe secondaire faible en puissance, mais une faible atténuation des lobes suivants.
- La fenêtre de Blackman : $w(n) = \frac{21}{50} + \frac{1}{2}cos(\frac{\pi n}{N-1}) + \frac{2}{25}cos(\frac{2\pi n}{N-1})$, possède un lobe secondaire très faible, des lobes suivants s'atténuant rapidement, mais un lobe principal plus large que les deux autres fenêtres.

Il est aussi fait usage de la fenêtre rectangulaire (1 sur la durée d'observation, 0 autrement), qui présente un lobe principal très fin mais des lobes secondaires élevés. Cette fenêtre ne présente aucun intérêt dans le cadre de cette étude.

Fréquence limite de voisement

Le fréquence limite de voisement [Stylianou, 1996, Kim et Hahn, 2007] est une solution simple au problème de la décomposition harmonique + bruit sur les signaux vocaux. Comme l'effet purement périodique est conditionné par le spectre de la dérivée de la source, caractérisé par une pente de -6dB/octave (sans phase de retour), arrive un point en haute fréquence où les harmoniques sont "noyés" par le bruit. On peut considérer alors que ce point correspond à l'endroit où n'est perçu que la contribution non voisée et que par effet de masquage, en amont de cette fréquence n'est perçue que la contribution voisée.

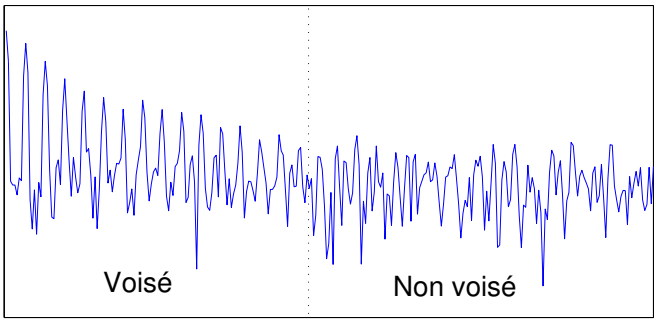


Figure 2.19 – Séparation du spectre d'un signal de parole en une partie voisée et une partie non voisée.

Cette prise de décision sur la fréquence limite de voisement est illustrée sur la figure 2.19.

Modèle sinusoïdal

Dans le cas de signaux purement harmoniques mais bruités, on peut quantifier la quantité de signal harmonique par estimation des paramètres d'un modèle sinusoïdal + bruit. Un tel modèle pour K harmoniques est donné par l'équation 2.15, on constate que chaque harmonique est identifié par son amplitude a_i, sa fréquence ν_i et sa phase ϕ_i, un bruit blanc Gaussien de variance σ est ajouté aux harmoniques afin de compléter le modèle.

$$s(n) = \sum_{i=0}^{K} a_i sin(2\pi n \nu_i + \phi_i) + b_\sigma(n) \tag{2.15}$$

Il suffit alors de déterminer les paramètres de ce modèle pour retrouver la partie voisée et par soustraction, la partie non voisée. Une adaptation de cette modélisation pour une transformée en harmonique a été proposée dans [Zubrycki et Petrovsky, 2010].

2.3.3 Estimation de la partie voisée

Méthodes d'estimation statistiques

Les méthodes d'estimations statistiques cherchent à minimiser l'erreur d'estimation $e = E(x - \tilde{x_\Phi})$ entre le signal original x et son estimée $\tilde{x}_\Phi$ pour un jeu de paramètres Φ. Pour cela il existe deux approches différentes. L'une basée sur l'estimation par le maximum de vraisemblance (minimisation d'une erreur) et une basée sur la projection des observations sur deux sous-espaces (bruit et signal) on peut citer par exemple la méthode ESPRIT [Roy et Kailath, 1989].

L'utilisation d'un modèle de sinusoïdes de fréquences ω_k et de phase Φ_k amorties d'un facteur d_k et d'amplitude a_k prédisant le signal x est présenté sur l'équation 2.16. Ce modèle, combiné à des méthodes d'analyse par décomposition en sous-espaces a été étudiée par Jensen dans [Jensen *et al.*, 1999], dans un contexte de codage des signaux de parole. Ces méthodes demandent tout de même une charge de calcul importante et ne tiennent pas compte des apériodicités structurelles présentes dans les signaux réels qui rendent difficile l'observation d'un segment de signal comme une superposition d'harmoniques.

$$x(n) = \sum_{k=1}^{K} a_k e^{-d_k n} cos(\omega_k n + \Phi_n) \tag{2.16}$$

Utilisation d'un filtre harmonique

La première idée consiste à chercher à séparer sur le spectre la contribution périodique de la contribution apériodique du signal [Serra et Julius Smith, 1990]. Pour cela, il faut supprimer les composantes du spectre considérées comme harmoniques. Comme chaque trame d'analyse est pondérée par une fenêtre, l'information de chaque harmonique est étalée sur le spectre entier. En première approximation on considère qu'on ne cherche à séparer que le lobe principal pour chaque harmonique. Après avoir déterminé la largeur du lobe principal, il s'agit de mettre tous les bins - éléments d'une TF discrète - correspondants de chaque harmonique à zéro afin de conserver la partie alors dénommée "apériodique". Un tel filtre est représenté en vert sur la figure 2.20.

Soit F_0 la position fréquentielle (en bins) de chaque harmonique, N_{fft} le nombre de points de la FFT et N la largeur de la fenêtre d'oubli - $N_{fft} \geq N$; $F_0 = \frac{\nu_0}{N}$. Soit L la largeur du lobe principal de la fenêtre : $L = 4\frac{N_{fft}}{N}$ (ou 2 ou 8, en fonction de la fenêtre), alors on peut définir un filtre $F(n)$ harmonique tel que :

$$\begin{aligned} F(n) &= 1 \text{ pour } kF_0 - \frac{L}{2} < n < kF_0 + \frac{L}{2} \text{ avec k entier} \\ &= 0 \text{ sinon} \end{aligned}$$

Ainsi, en appliquant ce filtre harmonique au signal, on retrouve une structure harmonique. Pour $L = 0$, on retrouve un filtre en peigne comme celui utilisé dans [Serra et Julius Smith, 1990, Jackson et Shadle, 2001]. Soit le vecteur $P(n)$ le spectre de la partie harmonique et $A(n)$ le spectre de la partie apériodique :

$$P(n) = S(n)F(n) \tag{2.17}$$

$$A(n) = S(n)(1 - F(n)) \tag{2.18}$$

La figure 2.20 illustre cet algorithme, où sont présentés les spectres S, A et P.

2.3.4 Filtre harmonique, optimisations diverses

Reconstruction du bruit par interpolation

L'utilisation du filtrage harmonique ne permet pas de dissocier la contribution du bruit de la contribution harmonique pour les bins harmoniques ($F(n) = 1$). En première approximation le bruit est considéré ayant une variation d'amplitude faible dans le domaine spectral : ainsi, si le filtre F est en peigne ($L = 0$) on peut estimer l'énergie du bruit au point n par interpolation des informations aux point $n - 1$ et $n + 1$. Différentes méthodes existent pour résoudre ce problème.

Reconstruction du bruit de manière itérative

L'algorithme précédent, bien que donnant des résultats intéressants, ne permet pas une décomposition optimale. En effet, on ne peut pas considérer que les régions contenant les informations sur les harmoniques du spectre soient dénuées de bruit, il s'agit donc de trouver un

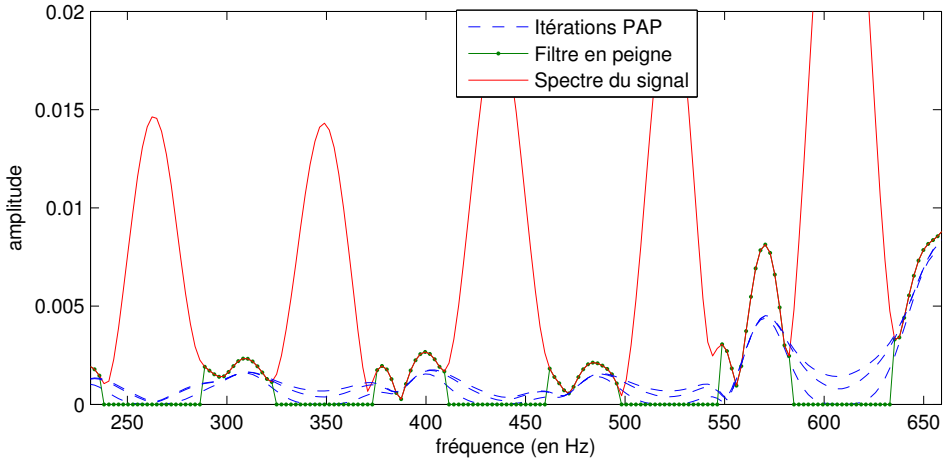


Figure 2.20 – Illustration de la décomposition périodique-apériodique sur le spectre du signal. Le trait continu représente le spectre du signal, le trait pointillé représente le filtre en peigne. Les itérations successives de la méthode PAP sont désignées par le trait discontinu.

moyen de reconstruire le bruit entre les harmoniques autrement que par moyennage. La méthode à base d'interpolation précédente ne respecte que l'hypothèse de l'enveloppe spectrale du bruit, mais pas sa cohérence de phase. Il semble plus naturel d'utiliser une reconstruction à base de contraintes sur le signal pour déterminer la partie bruitée. Un tel algorithme a été proposé [Yegnanarayana *et al.*, 1998] où deux contraintes (une spectrale, une temporelle) sont appliquées sur la partie apériodique de manière itérative afin de reconstruire ce qui se trouve en lieu et place du vide laissé par l'extraction des harmoniques par le filtre en peigne. L'avantage d'un algorithme de reconstruction est d'agir simplement sur la phase de la transformée de Fourier en plus d'agir sur les amplitudes comme pour les méthodes d'interpolation.

La première contrainte est d'ordre temporel. La transformée de Fourier $\hat{S}(\nu)$ du signal $s(n)$ de taille N est calculée avec un supplément de zéros ($N_{fft} = kN$ avec $k = 2$ ou $k = 4$). La transformée de Fourier inverse de la partie apériodique du filtre ($F(n) = 0$) donne un signal $\tilde{s}$ possédant des bins non nuls apres N. La première contrainte se situe donc sur la taille de la transformée inverse $\tilde{s}$, cette étape est réalisée à chaque itération (i) où l'on retrouve le signal s_i et sa transformée S_i :

$$\tilde{s_i}(n > N) = 0$$
$$\tilde{S_i} = TF[\tilde{s_i}]$$

La deuxième contrainte se situe sur la transformée de Fourier $\tilde{S_i}(n)$ du signal $\tilde{s_i}(n)$. En forçant les bins nuls du spectre en peigne F aux valeurs initiales de $\hat{S}(n)$ (équation 2.19), on apporte une deuxième contrainte à la reconstruction de la partie apériodique. De plus, cette étape permet de compenser la perte d'énergie apparue lors de la contrainte temporelle. Cette étape est réalisée à chaque itération (i), et permet de passer à l'itération suivante :

$$\tilde{S_i}(n) = \begin{cases} \hat{S}(n) & \text{si } F(n) = 0 \\ \tilde{S_i}(n) & \text{si } F(n) = 1 \end{cases} \tag{2.19}$$

$$\tilde{s}_{i+1} = TF^{-1}[\tilde{S_i}]$$

L"évaluation effectuée [Yegnanarayana *et al.*, 1998] montre que la reconstruction converge au bout de quelques dizaines d'itérations seulement. Cette méthode est illustrée sur la figure 2.20.

Décomposition par adaptation de la fenêtre d'analyse [Jackson et Shadle, 2001]

Dans les algorithmes précédents, le point commun se retrouve dans la taille de la fenêtre d'analyse, toujours fixée. Jackson [Jackson et Shadle, 2001] a proposé d'adapter la taille de la fenêtre à la fréquence fondamentale courante tout en conservant la proposition originale de recomposition du spectre par interpolation. Le travail en question donne la valeur de 4 périodes par fenêtre comme étant une valeur optimale empirique, car l'estimation des amplitudes et phases de chaque harmonique se fait sur le spectre du signal pondéré par une fenêtre carrée - aux lobes primaires étroits. Ensuite la partie voisée synthétisée est soustraite au signal pour obtenir la partie non voisée.

Au cours du chapitre 4, cette méthode sera utilisée comme point de comparaison.

2.4 Conclusion

Dans ce chapitre a été établi un état de l'art des méthodes et modèles existants pour l'analyse des signaux vocaux nécessaires à la bonne compréhension du manuscrit. Aussi bien l'estimation des instants de fermeture glottique, des instants d'ouverture glottique (ou de son pendant, le quotient ouvert) ainsi que la décomposition périodique ont été abordés.

L'estimation des instants de fermeture glottique repose sur un modèle éprouvé, la détection des impulsions sous-jacentes au voisement. Deux approches ont été présentées. La méthode DYPSA présente des résultats intéressants mais au prix de nombreux *a priori* sur la structure du signal et d'une dépendance forte à la possibilité d'appliquer le filtrage inverse aux signaux. Le principe de minimisation de la phase reste tout de même un domaine à explorer en vertu de sa robustesse. Le produit multi-échelles quant à lui ne tire pas parti de la totalité des propriétés de l'analyse par ondelettes, mais uniquement des informations basses fréquences présentes dans le signal original. Le protocole expérimental de validation de l'algorithme s'est limité aux voix "modales" sans distinction des valeurs de O_q pour la pertinence de l'estimation. De plus, toutes ces méthodes réduisent les informations des instants de fermeture glottique à une seule dimension : il conviendrait d'explorer dans quelle mesure une représentation multi-échelles des signaux vocaux en deux dimensions (temps et fréquence) pourrait fournir d'avantage d'informations en les combinant par exemple au principe de minimisation de la phase.

Concernant la décomposition périodique-apériodique, les méthodes actuelles ne tiennent que peu ou pas compte des variations locales du signal vocal, enfermées dans le modèle purement périodique des signaux. S'il est vrai que le terme "périodique" de la décomposition a pour vocation d'extraire des harmoniques, les signaux vocaux présentent très probablement des variations locales trop importantes pour utiliser un filtre harmonique à fenêtre de taille constante. Les travaux de Jackson ont prouvé que l'utilisation d'une fenêtre adaptative permet d'améliorer grandement la décomposition. Ce manque d'harmonicité des signaux vocaux empêche l'application directe des méthodes paramétriques (notamment de décomposition en sous-espaces) et rend difficile l'utilisation d'un modèle harmonique comme le HNM dans un cadre d'analyse. Dans le but d'améliorer la décomposition périodique/apériodique il conviendra dans un premier temps d'appliquer le principe de fenêtre adaptative à la reconstruction itérative du bruit et d'explorer dans quelle mesure les problèmes qui surviennent peuvent être améliorés.

Dans le cas de la paramétrisation de la source glottique, la majorité des méthodes se repose sur l'utilisation du filtrage inverse, un modèle certes éprouvé et robuste, mais qui impose une forme précise à la fonction de transfert du conduit vocal. Une méthode originale basée sur la position des zéros de la transformée en Z du signal a montré les bénéfices d'une décomposition basée uniquement sur la phase. Cependant, les résultats les plus convaincants sont généralement

obtenus au prix de nombreux ajustements et les paramètres qui en sont extraits ne sont généralement pas colinéaires aux paramètres du modèle de débit glottique généralement utilisés. On retrouve ainsi une combinaison de plusieurs paramètres pour NAQ (dépendant à la fois de O_q, α_m, mais aussi Q_a). Une telle approche rejoint celle de l'utilisation d'un seul critère de forme (R_d) pour caractériser l'onde de débit glottique. On peut toutefois déplorer cette réduction dans le cas de l'analyse de la parole expressive. Les analyses non linéaires comme ARX-LF imposent quant à elles une double contrainte sur le signal. Dans le cadre de cette thèse, nous préférons opter pour une approche de type causal/anticausal, plus libre et à la base de l'analyse par ZZT. Il s'agit donc dans un premier temps d'étudier les avantages de la décomposition ZZT par rapport aux méthodes "classiques" et d'explorer dans quelle mesure elle peut être utilisée pour estimer les paramètres du débit glottique.

L'omniprésence d'*a priori* sur les signaux et de contraintes fortes lors de l'estimation rend difficile l'application des méthodes existantes à des signaux de parole atypiques comme c'est le cas pour la parole expressive. Le travail présenté dans les chapitres à venir se concentrera sur l'adaptation des méthodes existantes ou le développement de nouvelles méthodes et de nouveaux modèles pour réaliser des décompositions et des estimations plus fidèles tant pour la parole modale que pour une variation importante des configurations glottiques.

Résumé

Les méthodes existantes

Il existe différentes approches pour l'estimation des paramètres de la qualité vocale. Un travail important a déjà été réalisé concernant l'estimation de l'onde de débit glottique - principalement par prédiction linéaire -, la détection des instants de fermeture glottique, la mesure du rapport harmonique/bruit pour déterminer la quantité de voisement, la décomposition périodique-apériodique ou encore l'estimation des paramètres de l'onde de débit glottique. Ces méthodes sont parfois largement utilisées (prédiction linéaire, dypsa) mais ne présentent des résultats fiables que dans des conditions bien précises qui ne sont pas tout le temps satisfaites pour de la parole expressive :

- La prédiction linéaire ne fonctionne pas sur les voyelles nasalisées (filtre non tout-pôles) et nécessite une pré-accentuation dont le coefficient conditionne la qualité de l'estimation du débit glottique par filtrage inverse. Une nouvelle approche à base de décomposition ZZT (Zéros de la Transformée en Z) nécessite d'être approfondie.
- Les méthodes de segmentation des signaux de parole en périodes fondamentales ne fonctionnent convenablement que lorsque la fermeture glottique est nettement marquée. Deux approches, basées sur la minimisation de la phase d'une part et sur l'analyse multi-échelles d'autre part, mériteraient d'être mises en commun pour déterminer dans quelle mesure les informations en deux dimensions de la représentation multi-échelles peuvent être utilisées pour extraire des informations connexes aux GCI.
- La décomposition périodique-apériodique nécessite soit des signaux stationnaires sur un laps de temps suffisant pour l'utilisation d'un filtre en peigne approprié soit des méthodes qui s'appliquent mal aux signaux non harmoniques comme les signaux de parole. Il s'agit d'explorer dans quelle mesure la reconstruction itérative peut être adaptée aux dernières avancées en matière de décomposition, notamment l'utilisation de fenêtres adaptatives.

Vers une analyse de la parole expressive

Il convient de revoir les méthodes existantes, éventuellement de proposer de nouveaux modèles, afin d'adapter l'estimation des paramètres de la qualité vocale à la parole expressive. Les méthodes doivent pouvoir être robustes vis-à-vis de fortes variations de fréquence fondamentale, d'amplitude de voisement ou encore de quantité de voisement. De nouveaux modèles doivent être proposés afin de limiter le nombre de contraintes sur l'estimation des paramètres. En effet, la parole expressive est un domaine mal connu par la communauté de l'analyse des signaux de parole et une grande plage de variation des paramètres permettra d'observer de nouveaux phénomènes.

Deuxième partie

Outils pour l'analyse de la qualité vocale

Chapitre 3

Ondelettes pour l'analyse des signaux vocaux

Sommaire

3.1 Méthode multi-échelles et application aux signaux vocaux. 75
3.1.1 Les ondelettes, principe 75
3.1.2 Détection de singularités par analyse multi-échelles 76
3.1.3 Application aux signaux de parole 76
3.2 Etude prospective : ondelettes appliquées aux signaux de parole . . 77
3.2.1 Observations 77
3.2.2 Liens avec le modèle de production linéaire 78
3.2.3 Parcourir les lignes 79
3.2.4 Forme des lignes 80
3.3 Méthode LoMA pour la détection de GCI 82
3.3.1 Méthode 82
Principe 82
Choix de l'ondelette 82
Suivi des maxima 83
Principe de l'Algorithme 83
3.3.2 Protocole de validation de l'algorithme 85
3.3.3 Résultats de l'évaluation 86
3.3.4 Discussion des résultats 88
3.3.5 Relation entre GCI signal et GCI articulatoire 89
Définition 89
Exemple 90
Analyse approfondie 90
3.4 LOMA pour la mesure de l'énergie relative 90
3.4.1 Normalisation des signaux 92
3.4.2 Energie cumulée sur la LoMA 92
3.4.3 Barycentre d'énergie de la LoMA 93
3.4.4 LoMA et distribution d'énergie : conclusion 94

3.5 Shimmer et jitter par les ondelettes 94
3.5.1 Base de données de signaux synthétiques 94
3.5.2 Mesure du jitter 95
3.5.3 Mesure du shimmer 97
3.5.4 Conclusion sur l'aptitude de la méthode LoMA à déterminer des apériodicités structurelles 98
3.6 Quotient ouvert et ondelettes 99
3.6.1 Forme des lignes et forme du débit glottique 99
Réponse en basse fréquence 99
Réponse en haute fréquence 100
3.6.2 Principe de l'extraction de la configuration glottique par ondelettes . . . 100
3.6.3 Validation par un signal synthétique 103
3.6.4 Tester l'algorithme plus en détail 105
3.7 Parallèle avec Mean Square Phase 105
3.8 Conclusion 106

Les approches courantes cherchant à segmenter les signaux vocaux tentent souvent d'extraire des informations annexes liées aux GCI. Ainsi, le filtrage en fréquence zéro [Murty *et al.*, 2009] tente aussi d'extraire l'amplitude de voisement lors de la segmentation alors que d'autres méthodes comme Dypsa [Naylor *et al.*, 2007] ou MSP [Bouzid et Ellouze, 2007] tentent d'y estimer l'instant d'ouverture glottique. L'approche MSP utilisant l'analyse par ondelettes est intéressante mais ramène les deux dimensions (temps et fréquences) de l'analyse multi-échelles en une seule, par l'utilisation d'un produit. En outre, l'utilisation des échelles exclusivement basses fréquences est contestable.

Ce chapitre présente un modèle de détection des instants de fermeture glottique par méthode d'analyse multi-échelles en deux dimensions. Il sera montré que cette représentation utilise aussi l'approche de minimisation de la phase. La connaissance de ces instants de fermeture permettra d'extraire des informations sur leur énergie grâce à l'utilisation de l'intégralité des informations contenues dans le plan temps/fréquence. L'étude du décalage temporel du premier harmonique permettra aussi d'approcher les valeurs de quotient ouvert. Des expériences de mesure du GCI, des apériodicités structurelles, des pentes spectrales et du quotient ouvert seront présentées et commentées.

3.1 Méthode multi-échelles et application aux signaux vocaux.

3.1.1 Les ondelettes, principe

Le problème de l'analyse temps/fréquence peut s'apparenter à l'incertitude de Heisenberg : plus on connait un signal précisément en fréquence et moins on le connaît précisément en temps, ceci bien sûr du point de vue de la résolution de la transformée de Fourier. La précision en fréquence est directement proportionnelle à la durée d'observation (taille de la fenêtre d'analyse). Cependant, pour de grandes fenêtres d'analyse, les informations données par une transformée de Fourier (TF) ne peuvent pas être localisées en temps à l'intérieur de cette fenêtre sans utilisation de connaissances sur le signal (modèle).

Une solution trouvée pour contourner ce problème est l'analyse multi-échelles. On conçoit un banc de filtres tel que la taille de la fenêtre d'analyse dépende de la fréquence centrale de la bande d'analyse. Plus on monte en fréquence, et plus on augmente la largeur de bande, tout en diminuant la taille de la fenêtre d'analyse. En gardant le rapport bande passante sur fréquence centrale constant, on réalise un banc de filtre à facteur de qualité constant.

Il existe de multiples manières de concevoir un tel banc de filtres, mais la règle générale consiste, pour faciliter la conception et l'analyse, d'utiliser une progression logarithmique de la fréquence centrale de chaque bande. Soit une ondelette $\Psi(t)$. $\Psi(t)$ doit être une fonction à énergie finie mais pas nécessairement à bande limitée, ou à support limité. On calcule la bande i du filtre multi-échelles par la relation de compression temporelle suivante. Soit a le coefficient de progression du filtre, alors la bande i (aussi appelée l'échelle a^i) a pour expression Ψ_{a^i}, présentée sur l'équation 3.1.

$$\Psi_{a^i}(t) = \frac{1}{\sqrt{a}^i}\Psi(\frac{t}{a^i}) \tag{3.1}$$

Le cas particulier $a = 2$ donne un banc de filtres en ondelettes dit "dyadique". Spectre et ondelettes d'un tel banc de filtre ayant comme ondelette primitive un sinus cardinal modulé sont présentés sur la figure 3.1. On remarque que sur une visualisation logarithmique, la largeur de bande reste constante mais se déplace en fréquence selon l'échelle. Cette propriété remarquable

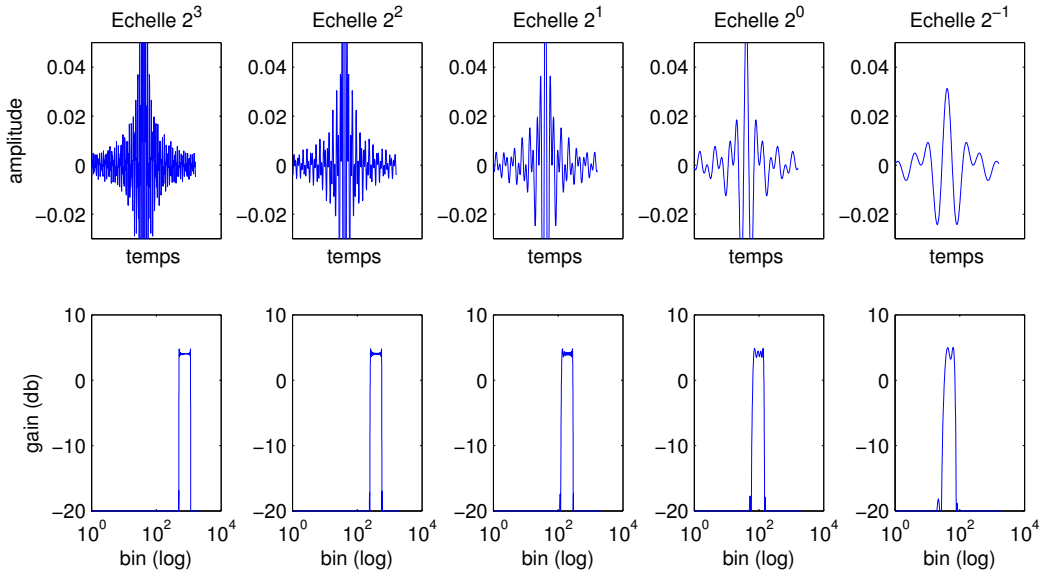


Figure 3.1 – Différentes réponses impulsionnelles et TF d'un banc de filtres en ondelettes dyadiques.

donne à l'analyse par ondelette des propriétés proches de l'analyse par bandes critiques du modèle d'audition [Sekey et Hanson, 1984] (pour les fréquences supérieures à 1000Hz).

3.1.2 Détection de singularités par analyse multi-échelles

Une singularité est un point de variation très bref sur un signal caractérisé par une non existence de la dérivée à tout ordre. La détection des singularités par analyse multi-échelles a été présentée par Mallat [Mallat et Hwang, 1992]. Dans un signal, une singularité peut être modélisée par une impulsion de Dirac ou par un échelon. Dans les deux cas, ces signaux possèdent un spectre très riche en énergie sur la totalité des fréquences. Un sursaut d'énergie permet de détecter une singularité.

Ceci fonctionne pour n'importe quelle représentation temps-fréquence, y compris la transformée de Fourier. Cependant, le problème de l'opposition entre localisation temporelle / localisation fréquentielle limite l'application de la détection de singularités par l'analyse de Fourier. Comme présentée sur la figure 3.2, l'utilisation de la sortie d'un banc de filtre en ondelettes permet une localisation bien plus précise et systématique qu'une analyse par TF à court terme (spectrogramme de la figure 3.2).

Le module de transformée de Fourier ou celui de la représentation en ondelettes ne donne pas la totalité des informations nécessaires pour la détection de singularités. En effet, si l'une des propriétés d'une impulsion de dirac est de présenter une énergie importante sur la totalité du spectre, une seconde propriété très importante est la synchronisation des phases sur la singularité. Sous condition d'utiliser un banc de filtres à phase nulle, notamment, pour la représentation en ondelettes, la visualisation directe des formes d'ondes issues de chaque bande permet de détecter beaucoup plus facilement les singularités. Un alignement des phases entre ondelettes visibles sur la figure 3.2 est très caractéristique de la présence d'une singularité dans un signal.

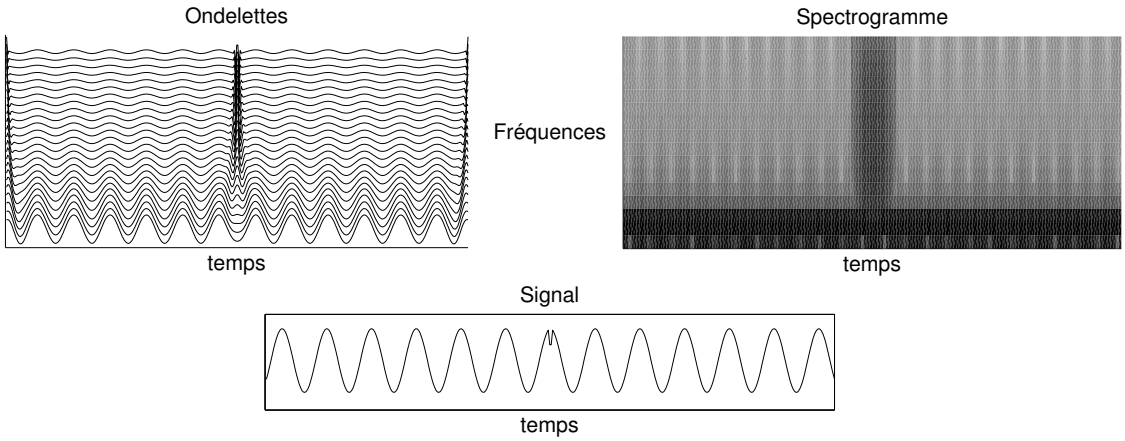


Figure 3.2 – Détection de singularité sur un signal (en bas) par la TF (à droite) puis par la sortie d'un banc de filtres en ondelettes (à gauche).

3.1.3 Application aux signaux de parole

Notre oreille perçoit naturellement les sons sur une dimension fréquentielle non linéaire. On peut donc s'inspirer de bancs de filtres modélisant la perception humaine des sons [Sekey et Hanson, 1984, Smith et Abel, 1999] pour analyser les signaux vocaux. Les bancs de filtres en ondelettes permettent de retrouver les propriétés de l'oreille humaine tout en conservant une simplicité analytique.

L'utilisation des ondelettes n'est pas nouvelle dans le domaine de l'analyse des signaux vocaux [Quatieri, 2001], mais son exploitation pour la détection de singularités dans le signal vocal a été possible grâce aux progrès dans le traitement des images [Mallat et Hwang, 1992]. Des travaux [Kadambe et Boudreaux-Bartels, 1992, Bouzid et Ellouze, 2007, Tuan et d'Alessandro, 1999] ont adapté le paradigme de détection de singularité sur les images pour les transposer aux signaux vocaux. En introduction de ce document une méthode récemment proposée [Bouzid et Ellouze, 2007] a été commentée. Mentionnons le travail de Kadambe et al. [Kadambe et Boudreaux-Bartels, 1992] qui est à l'origine de l'approche adoptée par Bouzid et al. mais qui réalise le produit spectral sur des échelles en hautes fréquences.

3.2 Etude prospective : ondelettes appliquées aux signaux de parole

Avant d'aborder une méthode développée au LIMSI, certaines observations des propriétés de l'analyse en ondelettes des signaux vocaux vont être présentées.

3.2.1 Observations

Observons l'analyse par ondelettes des deux signaux de parole assez caractéristiques présentés en figure 3.3. Les deux signaux sont voisés mais le deuxième présente clairement moins d'harmoniques : sa pente spectrale est plus prononcée. Cette observation se confirme sur la représentation en ondelettes, où les échelles supérieures présentent davantage d'informations dans le cas du premier signal. L'ondelette utilisée pour les analyses présentées par la suite suit l'équation 3.3 explicitée par la suite.

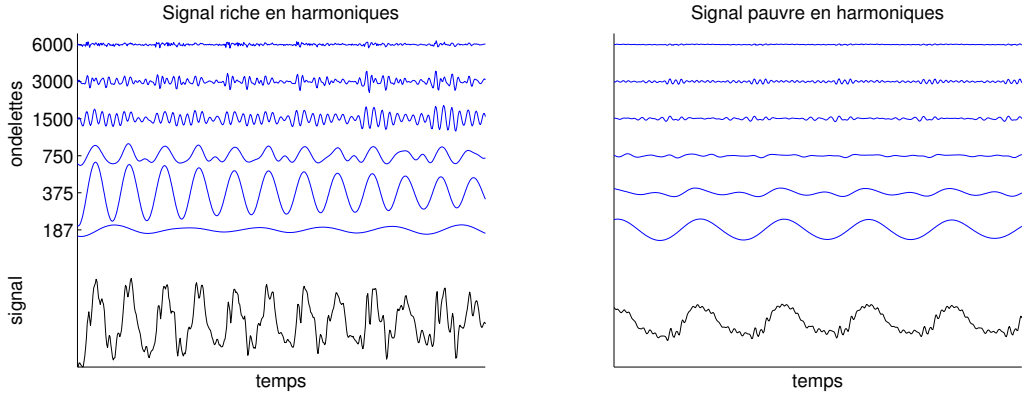


Figure 3.3 – Analyse par ondelettes de deux segments de signaux vocaux. La richesse spectrale visible sur le signal se traduit par l'observation des différentes échelles de décomposition.

Sur ces deux signaux, la périodicité se retrouve à toutes les échelles, par un effet de modulation des harmoniques à la fréquence fondamentale du signal. On ne retrouve cependant pas de strict alignement des maxima à travers les échelles. De plus, l'impulsion acoustique conséquente à la fermeture glottique est masquée par le filtrage de l'onde de débit glottique et par le conduit vocal atténuant alors très fortement les hautes fréquences. Peut-être n'y a-t-il même pas de fermeture ni d'impulsion dans ce cas ?

3.2.2 Liens avec le modèle de production linéaire

La transformée en ondelettes d'un train d'impulsions est présentée sur la figure 3.4. On retrouve directement les résultats prédits par Mallat [Mallat et Hwang, 1992], à savoir que chaque impulsion (vue comme une singularité dans le signal) provoque une oscillation sur toutes les échelles et que les maxima locaux sont synchronisés au moment de l'impulsion.

Pour un train d'impulsions de période T $\delta_T(t)$, les harmoniques sont synchronisés, le retard est donc nul à toutes les fréquences. Mais lorsque le signal est filtré par un filtre à phase non linéaire comme une période glottique $g(t)$, un retard apparaît qui peut décaler les maxima à travers les échelles. Pour un filtre à phase linéaire, ce retard temporel est constant pour toutes les fréquences, les maxima ne sont donc pas décalés. C'est ce qu'on visualise sur la figure 3.5 qui présente le train d'impulsions précédent filtré par une onde dérivée de débit glottique.

$$s(t) = [g * \delta_T](t)$$

Contrairement à la proposition de Mallat, il est choisi de visualiser directement la sortie des bancs de filtres à chaque échelle et non l'énergie. Ce choix apporte plus de robustesse à la visualisation, sous condition que la polarisation du signal soit connue. En effet, les extrêmes qui pointent sur les singularités du train d'impulsions sont maintenant les minima du signal ($\underset{t}{\operatorname{argmin}}\,(s(t))$), le gain du filtre étant négatif.

On constate que de cette manière la détection des instants de fermeture glottique par les ondelettes est extrêmement sensible à la polarisation du signal, mais beaucoup plus robuste sous condition que le signal soit convenablement polarisé. On observe aussi sur les figures 3.7 et 3.8 que les basses fréquences (dans la région proche du formant glottique) sont décalées par rapport aux hautes fréquences. Ce phénomène est tout à fait attendu : le retard de groupe appliqué au train d'impulsions par le filtre de l'onde de débit glottique n'est pas nul et présente un maximum au voisinage du formant glottique comme pour tout filtre du second ordre.

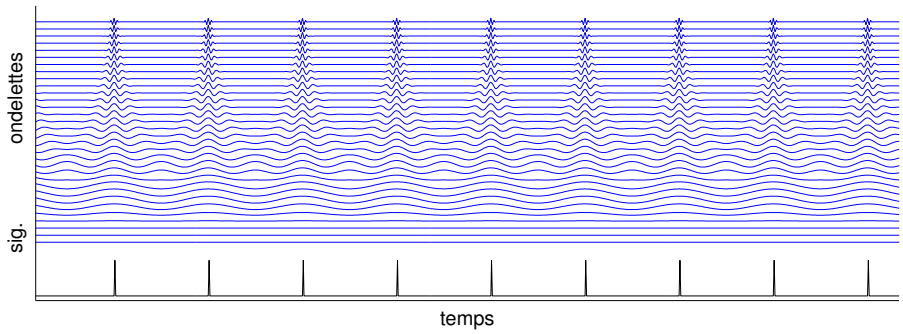


Figure 3.4 – Analyse par ondelettes d'un train d'impulsions, on retrouve bien l'alignement des maxima à travers les échelles.

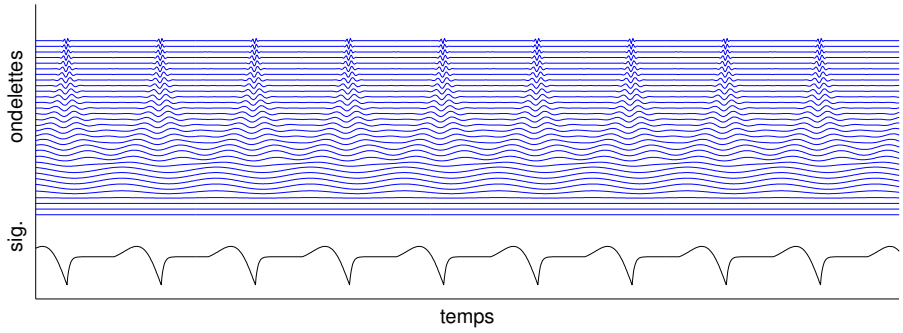


Figure 3.5 – Analyse par ondelettes d'une DODG, le filtrage du train d'impusions à l'origine de l'onde décale les maxima à travers les échelles.

Une étape supplémentaire dans l'analyse de l'effet du modèle de production linéaire sur l'analyse en ondelettes consiste en l'application d'un filtre vocalique au signal précédent. On choisit la voyelle /a/ de réponse impulsionnelle $f(t)$, présentée sur la figure 3.6.

$$s(t) = [f * g * \delta_T](t)$$

On remarque que le décalage des maxima est sensiblement identique à celui de la figure précédente et cela pour trois raisons :

- Pour le signal choisi, le formant glottique est nettement inférieur en fréquence et plus énergétique que le premier formant vocalique. Le décalage en basse fréquence est donc conditionné en majorité par la DODG.
- La fréquence fondamentale choisie permet une séparation complète des contributions du filtre et de la source : la réponse impulsionnelle du filtre pour un GCI est négligeable lors du GCI suivant. Lorsqu'il n'y a pas de repliement temporel de cette réponse impulsionnelle, les formants agissent peu sur le décalage des maxima lors des GCI. Des exemples seront donnés par la suite où ce repliement joue un rôle majeur dans le décalage des maxima.
- Le décalage temporel des maxima est fonction du retard de groupe et inversement proportionnel à la fréquence. Ainsi, les formants vocaliques placés dans le haut du spectre décalent peu les maxima en temps. En toute logique, le maximum de précision sur le GCI est obtenu en haute fréquence, en faisant abstraction des bandes où le bruit est trop présent.

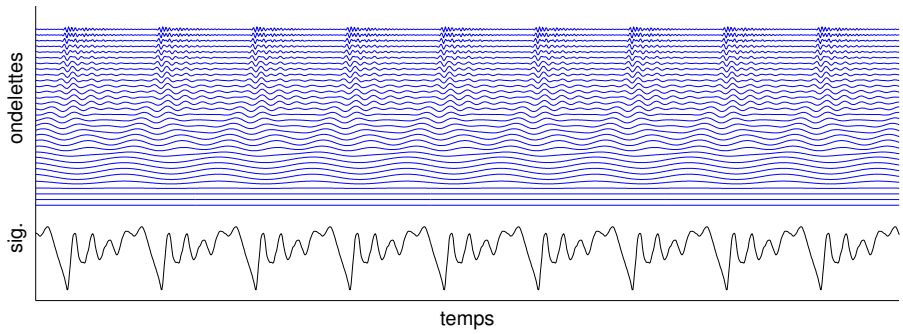


Figure 3.6 – Analyse par ondelettes d'un signal synthétique - signal de la figure 3.5 filtré par la fonction de transfert d'une voyelle /a/. Le filtre vocalique modifie d'avantage l'alignement des maxima.

3.2.3 Parcourir les lignes

Les travaux [Tuan et d'Alessandro, 1999] qui ont inspiré ceux présentés dans ce chapitre, ont montré que la combinaison des propriétés d'amplitude et d'alignement de phase des signaux temporels issus des bancs de filtre en ondelettes permet de retrouver les instants de fermeture glottique directement sur le signal de parole en descendant les maxima depuis les hautes fréquences. Cette méthode est appelée LoMA (Lines of Maximum Amplitude) car elle consiste essentiellement à tracer le chemin le plus court passant par les maxima de la représentation.

Il a été présenté précédemment qu'on retrouvait les propriétés d'un signal comportant des singularités par l'analyse en ondelettes d'un signal vocal, mais que l'alignement des maxima au travers des échelles d'analyse n'était pas maintenu suite aux différents filtrages (ODG + conduit vocal). Le parcours des lignes permet donc de compenser ce manque d'alignement (figure 3.7), nécessitant tout de même un algorithme approprié pour choisir convenablement le parcours à travers les échelles.

Soit le signal de TF $S(\omega) = F(\omega)G(\omega)\delta_{F_0}(\omega)$ et le décalage en phase $e^{\Phi(\omega)}$ détecté par le parcours des lignes. Dans le cas idéal on cherche :

$$e^{-\Phi(\omega)} = \frac{F(\omega)G(\omega)}{|F(\omega)G(\omega)|}$$

Afin de retrouver le signal compensé $\tilde{s}$ tel que :

$$\begin{aligned} \tilde{S}(\omega) &= S(\omega)e^{\Phi(\omega)} \\ &= |F(\omega)G(\omega)|\delta_{F_0}(\omega) \end{aligned} \tag{3.2}$$

Alors $\tilde{s}(t) \propto \delta_T(t)$ dans le cas où $TF^{-1}|F(\omega)G(\omega)|$ est principalement contenu autour de 0 (amortissement fort).

Le parcours des lignes en but de compenser le décalage de phase des filtres glottique et vocalique revient donc à une déconvolution partielle selon l'équation 3.2. L'effet précis de la compensation de phase vis à vis de l'utilisation d'un banc de filtre à facteur de qualité constant sera vu par la suite, lorsque nous chercherons à déterminer la forme exacte des lignes trouvées.

3.2.4 Forme des lignes

La forme des lignes est tout aussi intéressante que leur parcours (figure 3.8). En effet, le décalage de chaque maximum dans une échelle par rapport au GCI est fonction du retard de groupe

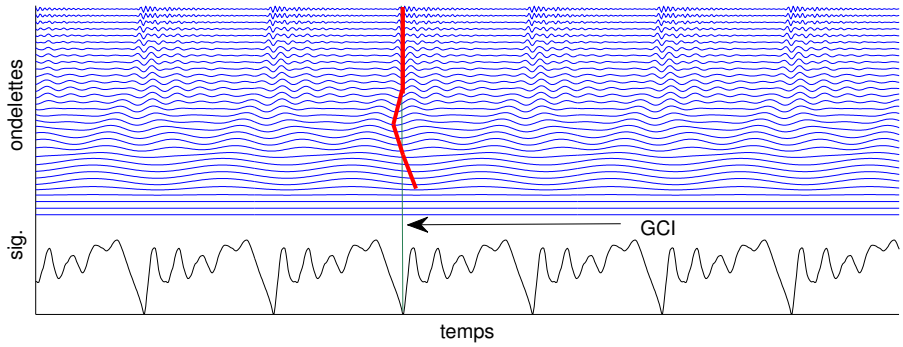


Figure 3.7 – Localisation du GCI par la position haute fréquence de la ligne (en rouge).

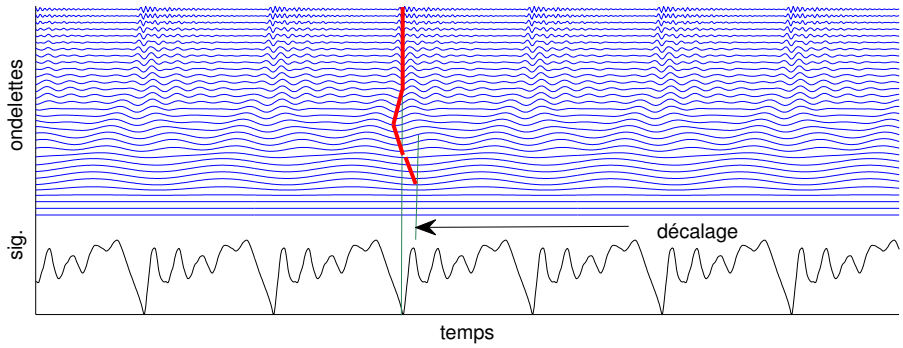


Figure 3.8 – Le premier harmonique présente un décalage par rapport au GCI.

autour de la fréquence centrale de cette échelle. Dans le cas où formant glottique et vocalique sont nettement séparables, on peut montrer que le décalage du maximum dans l'échelle la plus basse - celle comportant le fondamental - est une fonction du quotient ouvert. Un développement analytique et une série d'expériences seront réalisées à la section 3.6.

3.3 Méthode LoMA pour la détection de GCI

3.3.1 Méthode

Ce travail se situe dans la continuation directe de la méthode de détection des LoMA développée au LIMSI [Tuan et d'Alessandro, 1999]. Cependant, la méthode présentée dans ce chapitre n'utilise pas la même stratégie pour tracer les lignes d'amplitude maximum. Par rapport aux travaux précédents, nous utiliserons une connaissance *a priori* de la fréquence fondamentale et une programmation dynamique plus large trouvant la ligne d'amplitude maximum parmi toutes les possibilités offertes à une échelle donnée (pas de sélection des maxima, mais de la ligne dans sa totalité). Un coefficient de pondération sera introduit pour évaluer dans quelle mesure des résultats meilleurs sont obtenus en donnant une priorité plus ou moins importante au décalage entre maxima. Cette stratégie se révèle plus robuste et permet un contexte d'exploration des possibilités des LoMA. Après avoir exposé le principe de suivi des maxima à travers les échelles, l'algorithme sera détaillé.

Principe

Mallat a proposé d'utiliser les maxima du signal pour retrouver les singularités - discontinuités du signal. L'idée est apparue d'en tirer profit pour compenser la réponse du filtre vocalique dans le signal vocal [Tuan et d'Alessandro, 1999]. En effet, il a été observé précédemment que le filtre ajoute naturellement une avance de phase sur le signal, retard qui varie en fonction de la fréquence observée. Il est donc difficile de déterminer le GCI par une simple visualisation du signal. Les méthodes d'estimation les plus efficaces comportent une étape de compensation du filtre vocalique. Citons par exemple la méthode DYPSA qui procède à un filtrage inverse par LPC avant de commencer à estimer les GCI.

S'affranchir de la compensation du filtre vocalique, c'est aussi s'affranchir des erreurs de son estimation. Un *a priori* sur le modèle de production vocale n'est plus nécessaire pour traiter le problème. Mais cet *a priori* de la forme du filtre est remplacé par un *a priori* plus léger sur la forme du suivi des maxima. En effet, on limite l'excursion de la ligne de suivi sur l'échelle temporelle, c'est à dire qu'on rajoute un coût supplémentaire à la sélection d'un maximum plus important, mais plus éloigné. Le but est d'éviter d'accrocher les maxima provenant de résonances, ou de déphasages locaux des harmoniques causés par les apériodicités.

La figure 3.10 présente le principe de suivi des maxima sur un signal de parole réelle. On remarque que l'excursion temporelle (vers la gauche ou la droite) est fonction du retard de phase moyen induit sur la bande de fréquence considérée.Notons qu'un maximum local est défini comme la position de l'amplitude la plus importante entre deux passages par zéro.

Choix de l'ondelette

Afin d'assurer la synchronisation des réponses aux filtres par rapport au signal, il convient d'utiliser une ondelette à phase nulle. C'est une Gaussienne modulée en fréquence qui a été retenue, le suivi de maxima se faisant plus facilement lorsque la redondance d'informations est importante d'une bande à l'autre du banc de filtre [Delprat, 1992]. Comme nous recherchons les

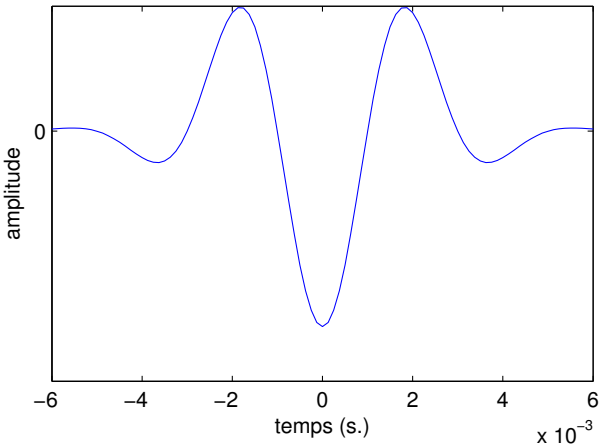


Figure 3.9 – Ondelette retenue, une gaussienne modulée en fréquence pour a=16 et Fe=8000.

instants de fermeture glottique, nous choisissons une ondelette négative. La fermeture brutale de la glotte entraine une dépression, donc un pic négatif dans le signal. Le choix d'une ondelette négative permet de visualiser un maximum sur les échelles à l'endroit supposé de l'extremum causé par le GCI, c'est un choix arbitraire qui rend la visualisation du suivi plus naturelle. L'équation de l'ondelette est donnée en 3.3.

$$h(t) = -\frac{1}{a}cos(2\pi\frac{F_e}{2a}t)e^{-2(\frac{F_e}{2a}t)^2} \tag{3.3}$$

Où F_e est la fréquence d'échantillonnage et a est l'échelle de l'ondelette (selon cette équation, l'échelle 1 correspond à un filtre de fréquence centrale $\frac{F_e}{4}$). Une illustration de la réponse temporelle de cette ondelette pour a=16 et Fe=8000 est donnée en figure 3.9.

Suivi des maxima

Le suivi des maxima se fait en cherchant, pour chaque passage de bande, à minimiser un chemin global. Soit $M(i, t_a)$ le maximum couramment sélectionné à l'échelle i et pour le temps t_a, on cherche donc le maximum à l'échelle $i + 1$ et au temps t_b selon l'équation 3.4.

$$t_b = \underset{t}{\operatorname{argmax}} \frac{M(i+1, t)}{1 + \alpha|t_a - t|} \tag{3.4}$$

Où α est un coefficient ajustable permettant de donner une priorité plus importante au décalage temporel $|t_a - t|$ ou à l'amplitude du maximum $M(i + 1, t)$.

Le suivi de maxima se révèle d'une grande efficacité sans connaissance *a priori* du schéma complet de leur répartition parmi les échelles. Afin de diminuer la quantité d'erreurs générées par l'algorithme de suivi des maxima, le critère est pondéré par une fonction de l'amplitude d'excursion en temps pour la recherche de maxima d'une échelle à une autre. Par exhaustivité, la totalité des lignes possibles est calculée sur l'équation 3.4, et un choix a posteriori détermine la ligne optimale en sélectionnant la ligne d'énergie maximum. L'algorithme est détaillé par la suite.

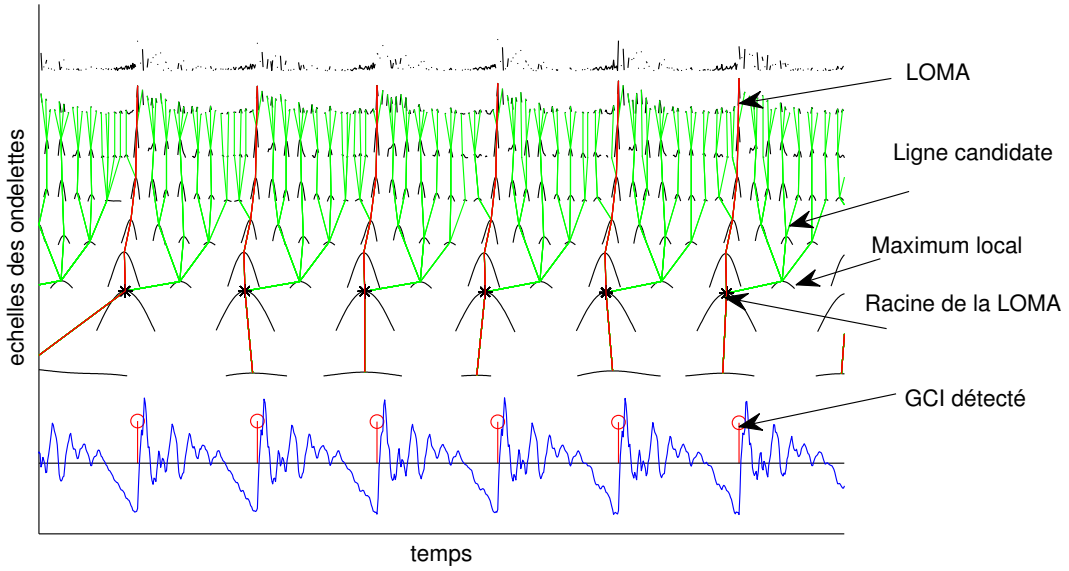


Figure 3.10 – Illustration de la détection des GCI par méthode multi-échelles : Lines Of Maximum Amplitudes - LoMA.

Principe de l'Algorithme

L'algorithme retenu est décrit ci-après et présenté sur la figure 3.10.

1. Dans un premier temps, il s'agit de calculer le banc de filtres. L'algorithme accepte comme paramètre la base logarithmique utilisée pour l'analyse multi-échelles. Pour des ondelettes dyadiques, ce nombre vaut 2, pour une infinité d'échelles il tend vers 1. Si l'expérience montre que choisir un nombre plus petit que 2 n'apporte rien en matière de suivi de maxima, l'avantage peut résider dans la représentation des signaux pour l'analyse visuelle, dans ce cas le coefficient optimal se situe entre 1.1 et 1.2. Cette différence de visualisation est présentée plus tard, pour des signaux différents, sur la figure 3.13 où le signal du haut est tracé pour $a = 2$ et celui du bas pour $a = 1.2$.
2. Les éléments nécessaires à l'algorithme, en plus de l'ondelette et du facteur d'échelle, sont les bornes supérieure et inférieure de la fréquence fondamentale, ainsi qu'une estimation approximative de fréquence fondamentale. Une simple utilisation de l'autocorrélation du signal permet d'obtenir une estimation suffisante pour l'algorithme à condition d'éviter les erreurs d'octave ; dans l'optique de ne pas biaiser l'expérimentation - basée sur la localisation du GCI plus que sur la détection de la fréquence fondamentale - les estimations de F_0 seront obtenues en utilisant l'algorithme Yin [de Cheveigné et Kawahara, 2002]. Cette connaissance de F_0 sert seulement au choix de l'échelle de départ pour la sélection des LoMA.
3. Dans un deuxième temps, la réponse de chaque filtre est calculée, les maxima y sont repérés (la partie positive de chaque réponse est présentée sur la figure 3.10). En partant de l'échelle la plus petite, les lignes - en vert sur la figure 3.10 - sont tracées en reliant chaque maximum au maximum le plus proche de l'échelle suivante. En partant des échelles haute fréquence on s'assure d'avoir toujours un nombre de maxima source supérieur au nombre de cibles

et donc d'avoir des lignes tracées en forme d'arbre qui convergent de haut en bas. Chaque maximum est repéré par son échelle i et son temps t par $M(i,t)$.

C'est en ce point que la méthode proposée ici diverge de la méthode proposée dans [Tuan et d'Alessandro, 1999]. Les maxima ne sont pas chaînés sur un critère mêlant amplitude et proximité, mais leur éloignement pénalise l'énergie cumulée sur tout la ligne (comme montré par la suite).

4. Dans un troisième temps, la valeur cumulée de chaque ligne est calculée depuis les basses jusqu'aux hautes fréquences en cumulant la quantité calculée d'amplitude $Q(n,i)$ pour la ligne n à l'échelle i :

$$\begin{aligned} Q(n, i_{max}) &= 0 \\ Q(n,i) &= Q(n,i+1) + \frac{M(i, t_{(n,i+1)})}{1 + \alpha |t_{(n,i+1)} - t_{(n,i)}|} \end{aligned} \tag{3.5}$$

Il est important de rappeler que les lignes sont numérotées de manière opposée à la croissance de la fréquence centrale du filtre auquel elles appartiennent. C'est à dire que $i = 1$ désigne l'échelle la plus petite - de plus haute fréquence - c'est pourquoi en remontant depuis les basses fréquences on calcule $Q(n,i)$ en fonction de $Q(n,i+1)$.

5. En fonction de la fréquence fondamentale moyenne, l'échelle correspondante est identifiée - étoiles noires sur la figure 3.10 -. Chaque maximum de cette échelle est considéré comme la racine de la LoMA. La ligne issue de cette racine possédant la plus grande quantité Q est sélectionnée. La position temporelle de cette ligne à l'échelle $i = 1$ désigne alors la position du GCI estimée pour la racine courante.

On désigne alors la n-ième LoMA L_n par une séquence de temps $t_{(n,i)}$ correspondant à l'endroit du maximum à chaque échelle i sur la ligne : $L_n(i) = t_{(n,i)}$.

On fait de même pour la séquence de l'énergie de la n-ième LoMA $E_n(i)$ associée à chaque maxima $M(i, t_{(n,i)})$ pour chaque échelle i : $E_n(i) = M(i, t_{(n,i)})$.

6. Dans le cas de signaux fortement bruités, il arrive que la présence de bruit en haute fréquence ne permette pas une localisation précise des GCI. Dans ce cas, l'algorithme dispose d'un paramètre qui est l'ordre du filtre le plus élevé à prendre en compte. Dans les études suivantes, l'échelle de fréquence maximale utilisée pour la détection des GCI est $i = 2$ en l'absence d'indication contraire. Soit la bande $2kHz - 4kHz$ pour des signaux échantillonnés à 16kHz. Sur la figure 3.10, on remarque que l'échelle de plus haute fréquence n'est pas prise en compte dans la construction des arbres.

3.3.2 Protocole de validation de l'algorithme

La validation de l'algorithme se base sur la mesure de la précision de l'estimation du GCI sur une base de données possédant un signal EGG enregistré de manière synchrone à la parole. Les signaux de parole en question sont extraits de la base de données [Tuan et d'Alessandro, 2000] dont deux locuteurs ont été choisis : un locuteur masculin (C) et un locuteur féminin (M) y lisent des articles de journaux selon une voix modale. L'EGG, manipulé avec précaution, se révèle être une référence de qualité pour les signaux naturels tant au niveau de sa précision que de sa simplicité d'analyse.

Afin de déterminer si la méthode présentée permet de se passer du filtrage inverse généralement utilisé pour l'estimation des instants de fermeture glottique, une méthode dénommée LoMA-LPC sera aussi testée. Cette méthode reprend l'intégralité de l'algorithme présenté ci-avant, mais sur le résidu d'une estimation par LPC. L'estimation est réalisée pour 18 pôles, sur une fenêtre de 20ms pondérée par fenêtre de Hann, avec une superposition de 10ms.

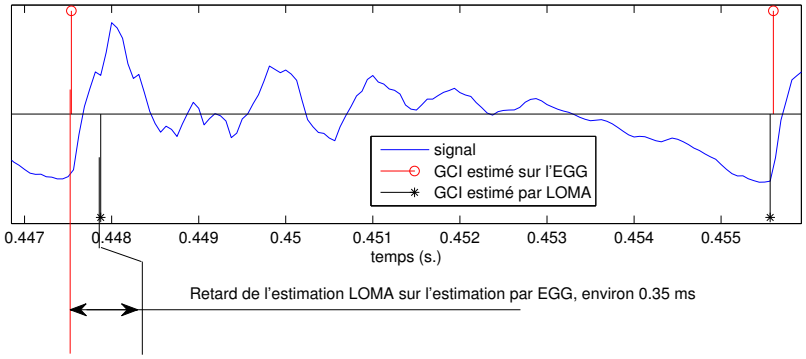


Figure 3.11 – Protocole de validation de l'estimation des GCI par la méthode des LoMA sur un cas très défavorable. Les GCI détectés sur l'EGG sont considérés comme une référence.

Le calcul de l'erreur est réalisé en cherchant l'instant estimé le plus proche d'un instant de référence sur une largeur temporelle d'une période (+/- $\frac{T_0}{2}$). Dans le cas où aucun instant d'estimation n'est trouvé, une omission est comptabilisée. Dans le cas où plusieurs instants d'estimation sont disponibles à proximité d'un GCI mesuré sur l'EGG, le plus proche est sélectionné et les autres sont comptabilisés comme des sur-détections. Ce principe est illustré sur la figure 3.11. Afin de simplifier l'écriture dans la suite du chapitre, nous désignerons la moyenne de cette erreur par la fonction $\tilde{E}()$ en référence à la notion d'espérance mathématique.

Pour évaluer les performances de l'algorithme nous l'avons comparé à la méthode DYPSA [Naylor *et al.*, 2007]. Dans un premier temps, les instants détectés $\hat{GCI}_{DYPSA}$, $\hat{GCI}_{LPC}$ (de la méthode LoMA sur le résidu de la LPC) et $\hat{GCI}_{LOMA}$ sont calculés avec leur méthode respective et la référence GCI_{egg} est extraite du signal.

Pour déterminer l'erreur, il convient dans un premier temps de corriger GCI_{egg} par le décalage naturel intervenant entre le signal acoustique capté au niveau du micro et le signal électrique capté par électroglottographie. Pour ne pas introduire de biais dans l'erreur, ce décalage est déterminé par une première passe du calcul d'erreur entre $\hat{GCI}$ de n'importe quelle méthode et GCI_{egg}. Le décalage est alors considéré comme la moyenne de cette erreur sur la première passe. Ce décalage D est calculé sur les jeux d'erreurs (par LoMA et par DYPSA) afin d'uniformiser la correction pour toutes les méthodes, l'expression de D est présentée en l'équation 3.6.

$$D = \frac{\tilde{E}[\hat{GCI}_{LOMA} - GCI_{egg}] + \tilde{E}[\hat{GCI}_{DYPSA} - GCI_{egg}]}{2} \tag{3.6}$$

En pratique, le décalage EGG-Parole est variable au cours du temps, en fonction de la position du larynx et de la position du locuteur par rapport au microphone, des données incontrôlables au stade de l'analyse. La correction du décalage sert donc principalement à éviter les erreurs de sur ou sous détection. Par la suite, l'attention sera portée principalement sur la forme et variance de la distribution de l'erreur. Mis à part la variation de ce décalage au court du temps, cette correction a principalement un effet sur la quantité de sous-détection ou sur-détection.

3.3.3 Résultats de l'évaluation

Les résultats de l'évaluation de l'algorithme de détection des LoMA avec et sans filtrage inverse par LPC et de la méthode DYPSA sont donnés sur la figure 3.12, les taux de fausses détections et de non détection sont présentés sur la table 3.1. Ils sont séparés par méthode et

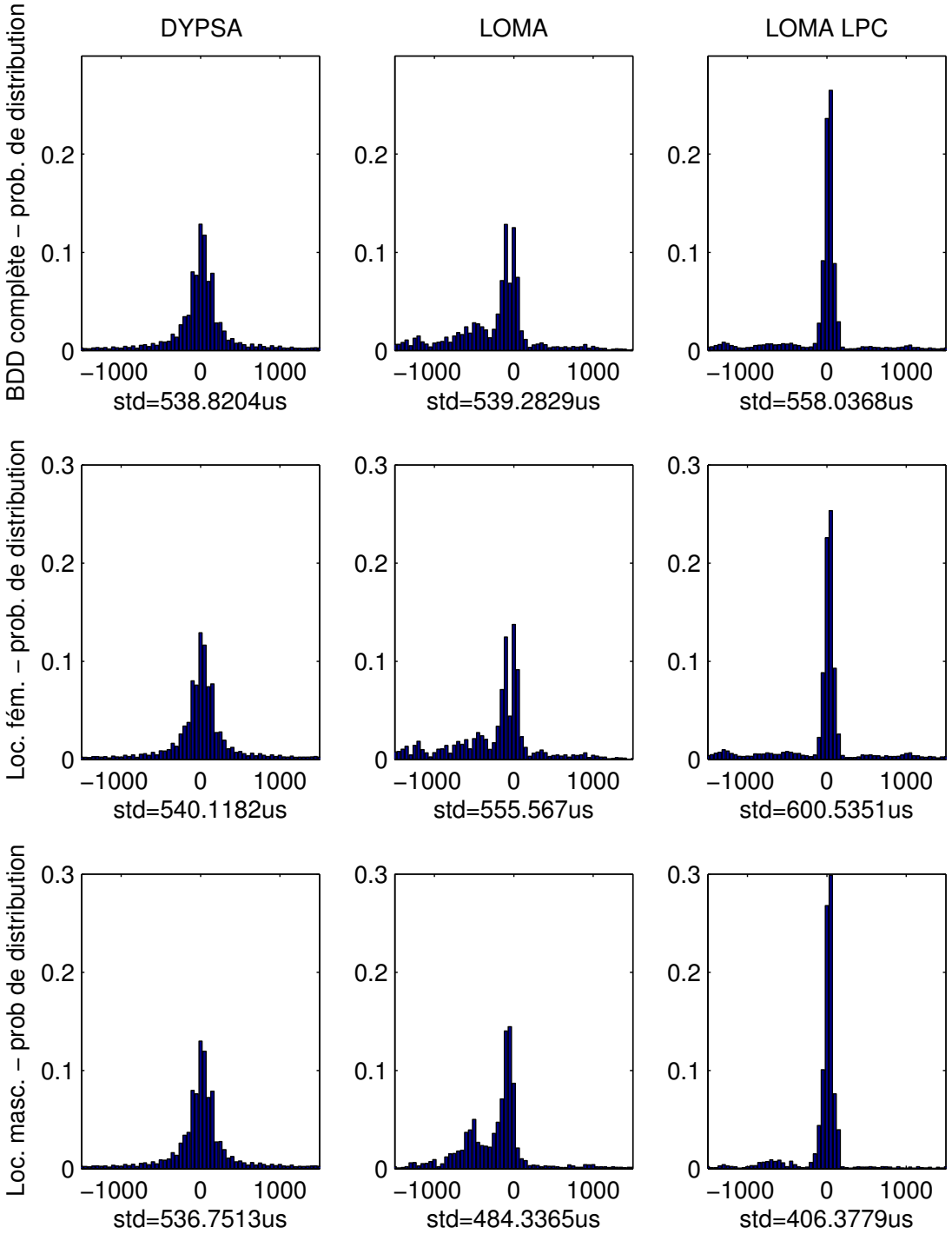


Figure 3.12 – Distribution de l'erreur de détection des GCI sur la base de données des signaux réels en micro-secondes.

Table 3.1 – Sous-détections (MR) et fausses alarmes (FA) pour le locuteur masculin (M), féminin (F) et la totalité de la base (T). EGG comparé à DYPSA (DYP), LoMA (LOM) et LPC-LoMA (LPC).

Methode	**MR T**	**MR M**	**MR F**	**FA T**	**FA M**	**FA F**
LPC	12.95%	10.25%	13.84%	0.53%	0.60%	0.50%
DYP	4.25%	1.33%	5.21%	0.52%	0.63%	0.48%
LOM	2.88%	3.03%	2.83%	0.50%	0.59%	0.47%

par locuteur. Il est utile de pouvoir évaluer les résultats en fonction du genre du locuteur car on peut ainsi les corréler avec la valeur moyenne de la fréquence fondamentale.

Pour l'analyse de l'ensemble des résultats (ligne supérieure de la figure 3.12) on remarque que les méthodes LoMA et DYPSA ne donnent pas les meilleurs résultats ; c'est la version utilisant un pré-traitement à base de LPC qui propose la distribution la plus resserrée pour un écart type seulement faiblement plus élevé. De même, les performances de LoMA sont comparables à celles de DYPSA. La distribution de la méthode LPC est très concentrée autour de l'erreur nulle. En ce qui concerne la variance, on remarque que les méthodes LoMA et DYPSA présentent le nombre le plus bas à $1\mu s$ près. Le taux important de sous-détections de la méthode LPC modère ses performances en terme de distribution. La méthode LoMA, qui présente une distribution à la dispersion proche de celle de DYPSA, présente un taux de ratés plus faible, pour un taux similaire de sur-détections.

Pour le locuteur masculin, la méthode à base de LPC donne la distribution la plus fine, mais un étalement large qui augmente la variance. On constate un pic secondaire sur les résultats de la méthode LoMA du locuteur masculin. Pour DYPSA, on remarque que la forme de la distribution est comparable aux résultats globaux. Dans le cas du locuteur masculin, c'est la méthode DYPSA qui présente la variance la plus faible, ainsi que le taux de ratés le plus faible.

Pour le locuteur féminin, la méthode à base de LPC présente la distribution la plus fine et possède la plus faible variance. Sur cette partie de l'analyse, les performances de DYPSA sont nettement moins bonnes, et le taux de ratés est pratiquement le double de celui des LoMA.

3.3.4 Discussion des résultats

Avant tout, les résultats nous montrent que la méthode LoMA présentée est compétitive face à une méthode comme DYPSA. Ces deux méthodes se différencient par leur conception : la méthode DYPSA utilise un grand nombre de traitements en amont et aval alors que ces étapes sont plus légères pour LoMA. Cependant, dans l'état actuel, la méthode LoMA utilise plus de temps de calcul que DYPSA, temps de calcul qui sera mis à profit par la suite pour extraire des informations supplémentaires sur les signaux vocaux (une optimisation de la méthode sera nécessaire avant sa diffusion sous forme de code, mais une telle optimisation sort du cadre de cette thèse).

Les résultats, qui donnent un écart type moyen de 0.6 ms montrent aussi la difficulté d'arriver à une estimation précise de l'instant de fermeture glottique. Le problème se situe probablement dans le protocole retenu pour l'évaluation des méthodes. Cependant, le choix s'est délibérément porté sur une analyse de l'erreur brute d'estimation par rapport à une référence et non sur le pourcentage de bonnes détections pour un certain seuil. Les travaux actuels font déjà état de nombreuses méthodes présentant des résultats proches de 95-100% de détections à un seuil de 0.25ms. Dans l'optique d'utiliser les GCI pour mesurer les apériodicités structurelles comme le

jitter, il est nécessaire de faire des mesures les plus précises possibles sans établir de seuil : c'est le parti pris par cette étude.

La méthode LoMA présente des extrema secondaires sur la figure 3.12. Ces extrêma sont gommés par l'application de l'algorithme sur le résidu de la LPC. On peut donc en conclure que sur le signal complet, l'algorithme a tendance à accrocher des maxima secondaires. Tout repose alors sur l'ajustement du coefficient de pénalisation d'éloignement du maximum lors du calcul de l'énergie de la ligne.

Le choix de la référence EGG est discutable, car le délai acoustique entre la glotte (où est mesuré le signal EGG) et le microphone varie en fonction non seulement des mouvements du sujet mais aussi de son articulation. Une demi milliseconde de délai représente environ 15cm à l'échelle de la vitesse du son, bien d'avantage que la plage de variation de la position de la glotte (environ 6 cm [Sundberg et Nordström, 1976]) ou la variabilité de la position du locuteur dans des conditions d'enregistrement rigoureuses.

Mais il faut prendre en compte la précision de l'estimation sur le signal. Dans la méthode LoMA, les signaux testés ont été échantillonnés à 16kHz. Dans une première approximation, et pour limiter l'impact du bruit de phonation sur la précision de l'estimation, la première bande du filtre par ondelettes (haute fréquence) est évincée du parcours de maxima. Par conséquent, la précision maximale d'estimation par LoMA sera légèrement supérieure à la fréquence de 4kHz soit une précision temporelle de l'ordre de 0,1 à 0,2 ms.

L'un dans l'autre, nous retrouvons donc une erreur globale sur l'estimation équivalente aux 0.6ms d'écart type observés sur les résultats de l'évaluation. Á ce stade, il serait difficile de décréter une méthode plus efficace qu'une autre pour une différence d'écart type de 40 μs. Cependant, la présence d'un lobe secondaire sur le graphique de résultats d'estimation de la méthode LoMA pour le locuteur masculin montre une propension de cette méthode à accrocher des harmoniques là où visiblement DYPSA reste plus stable. L'application de la détection des LoMA sur le résidu de la LPC donne des résultats intéressants, avec une grande précision de prédiction, mais un nombre important de ratés.

3.3.5 Relation entre GCI signal et GCI articulatoire

Se baser sur l'EGG est une bonne manière d'analyser les performances de l'algorithme LoMA de manière systématique. Mais la précision du signal EGG peut elle-même être mise en question, c'est pourquoi des signaux extraits de la base de données fournie par Nathalie Henrich [Karakozoglou *et al.*, 2010] ont aussi été analysés et observés. Cette base de données dispose, en sus du signal EGG, de la vidéo haute vitesse du mouvement des plis vocaux. La vidéo a été échantillonnée à une fréquence 4kHz, le signal audio à 44,1kHz. Il existe un délai entre les signaux EGG, vidéo et acoustique. Ces signaux ont été ajustés en amont de l'analyse de la base de données. L'équipement important nécessaire à l'enregistrement de ces données glottiques a produit un bruit important sur le signal, notamment une forte composante imputable à la fréquence secteur. Ce bruit est si important et étalé que même l'estimation de la fréquence fondamentale par Yin [de Cheveigné et Kawahara, 2002] en est trompée. Le filtrage nécessaire à l'analyse de ces signaux étant assez destructif, seule une poignée d'entre eux sera analysée.

Définition

La définition du moment articulatoire précis de l'instant de fermeture glottique peut être sujet à discussion, mais dans tous les cas il se situe entre 1) le moment où les plis vocaux commencent à se toucher et 2) le moment où ils sont fermés au maximum (pas toujours totalement). Loin d'être quantitative, cette analyse du rapport entre estimation du GCI par les LoMA et mouvement

Table 3.2 – Résultat de la détection des GCI par LOMA

Echantillon	67a	64	60a	57a	59a	61a
Bonne détection (%)	80	100	100	62	60	100
Bonne détection EGG (%)	100	100	91	87	100	100
TOTAL des détections	5	5	12	8	5	4

des plis vocaux se veut qualitative. On considère donc que tout GCI estimé entre ces deux points a une signification articulatoire acceptable. En effet, l'analyse LoMA se base sur le signal acoustique, qui n'a pas une forme linéairement dépendante de l'aire glottique (espace situé entre les plis vocaux). L'impulsion à l'origine de l'excitation du conduit vocal - et donc de l'apparition d'un GCI - peut survenir à tout moment de la fermeture (au sens propre) des plis vocaux.

Exemple

Deux GCI estimés par LoMA sont analysés sur la figure 3.13. La séquence d'images vidéo en proximité de ce GCI est donnée afin d'avoir des informations fidèles du mouvement des plis vocaux. Les instants d'échantillonnage de la vidéo et les GCI estimés par LoMA sont donnés sur le signal EGG synchronisé avec le signal acoustique. Sur le premier exemple (haute fréquence) l'EGG et les LoMA ont une à deux images de retard par rapport au mouvement des plis vocaux (0.5ms) sur la détection de l'instant de fermeture. Sur le deuxième exemple, les deux instants (EGG et LoMA) sont estimés dans le moment compris entre le début de fermeture et le maximum du contact des plis vocaux.

Analyse approfondie

Sur le tableau 3.2 on retrouve le pourcentage de bonnes estimations pour quelques échantillons de la base de données. Pour cette étude, on définit une bonne estimation comme la détection d'un GCI entre le moment de contact des plis vocaux et celui de fermeture maximum. Le nombre d'échantillons retenus étant petit (39), le comptage s'est fait à la main. Un petit nombre d'échantillons a été sélectionné, en prenant en compte la précision de l'estimation des GCI sur l'EGG, notamment.

Les résultats présentés sur le tableau 3.2 nous montrent que l'estimation des GCI par LoMA a une signification comparable à l'instant de fermeture glottique mesuré par EGG. Dans 85% des cas, l'instant de fermeture glottique détecté par les LoMA se situe dans le laps de temps correspondant effectivement au rapprochement des plis vocaux.

3.4 LOMA pour la mesure de l'énergie relative

La méthode LoMA ne donne pas que des informations temporelles sur le placement des GCI. L'utilisation de la quantité Q déterminée précédemment (équation 3.5) peut être utilisée pour donner une mesure relative de l'énergie d'une période à l'autre. En choisissant $\alpha = 0$ dans l'équation 3.5, on peut par exemple cumuler l'énergie de tous les maxima de la ligne, pour retrouver une image de l'amplitude de l'impulsion source. De la même manière, on peut tracer une évolution de l'énergie pour le GCI considéré à travers les échelles d'analyse et en déterminer une évolution de la richesse spectrale.

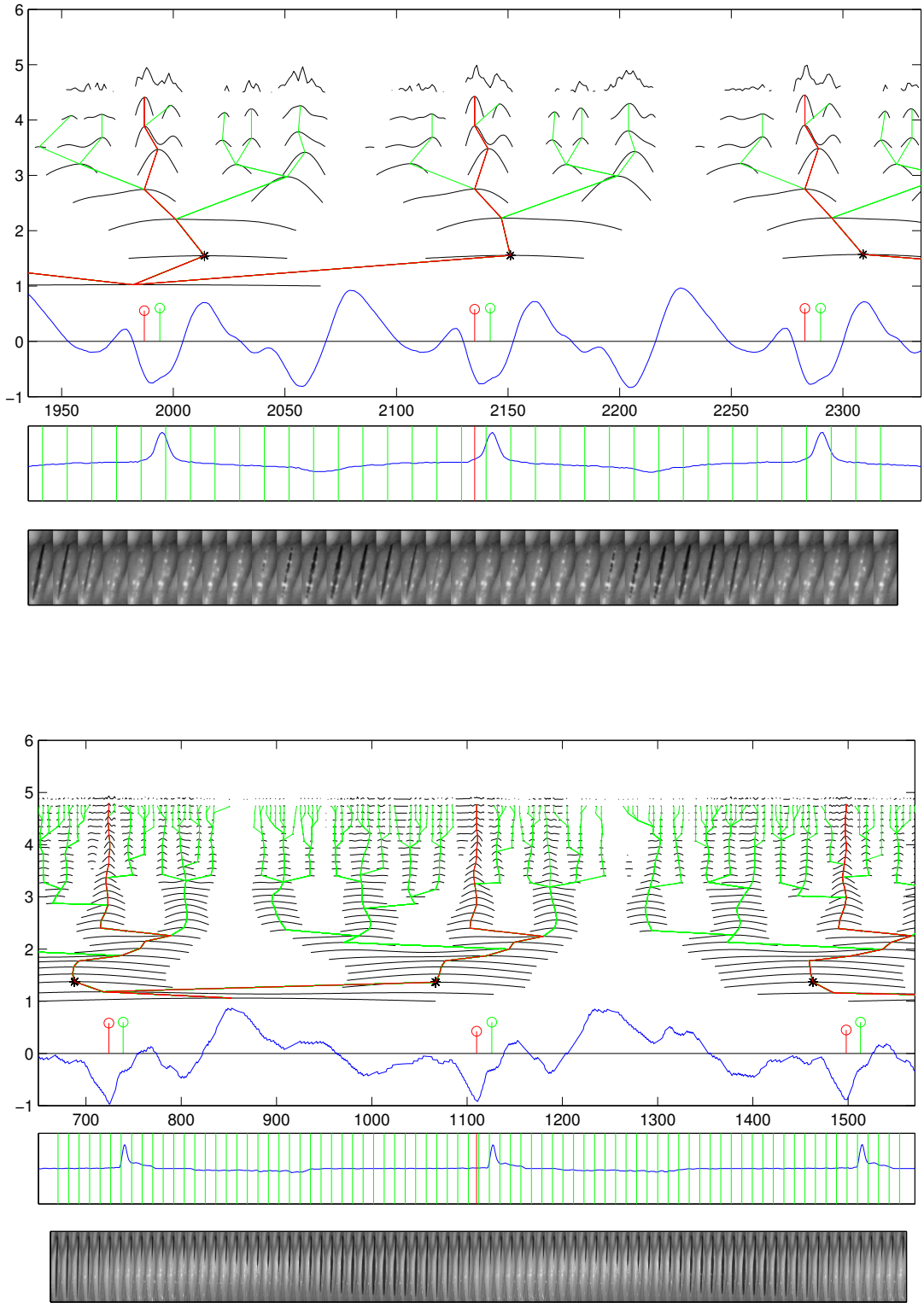


FIGURE 3.13 – Zoom sur un GCI estimé par les LoMA sur deux signaux produits par un locuteur masculin. Pour chaque figure, de bas en haut : ondelettes et signal acoustique (première case), DEGG puis Images de la vidéo, l'abscisse correspond aux échantillons du signal ($Fs = 44,1kHz$). Les épingles donnent les GCI estimés par LoMA (en rouge) et EGG (en vert). Les traits verts sur la DEGG donnent les instants d'échantillonnage de la vidéo, le trait rouge indique le GCI estimé par les LoMA.

De toute évidence, cette énergie sera liée à la forme du filtre vocalique. Mais après tout, la perception que nous avons de l'intensité d'un signal vocal est aussi fonction du filtre vocalique. Ce phénomène a été longuement étudié dans le cas du chant, notamment en ce qui concerne le formant du chanteur [Winckel, 1954].

Afin d'étayer l'utilité d'une telle mesure il convient de l'opposer à deux principes de détermination de l'intensité d'un signal perçu, généralement indépendants. D'une part, la richesse spectrale du signal et d'autre part la quantité d'énergie acoustique produite.

Dans un premier temps, une petite base de données de signaux tests a été formée, les mesures d'amplitude de ligne et de richesse spectrale ont été comparées aux résultats obtenus par les méthodes conventionnelles (puissance RMS, barycentre spectral [Mcadams, 1999]). Cette base de données est composée de 3 voyelles courtes et répétées avec une intensité évoluant dans le temps. Ces voyelles sont disponibles sur le site web [1] (fichiers `LOMA-Voyelle1.wav`, `LOMA-Voyelle2.wav`, `LOMA-Voyelle3.wav`).

3.4.1 Normalisation des signaux

Afin de pouvoir correctement déterminer si les lignes d'amplitude maximale permettent de déterminer convenablement les variations d'intensité perçues du signal sonore, il convient de procéder à des tests sur une base de données appropriée.

Dans un premier temps, des signaux tests composés de différentes voyelles prononcées avec un effort vocal plus ou moins important seront utilisées. Afin de limiter les biais de la mesure, les sons vont être normalisés sur deux niveaux :

- Normalisation en fréquence fondamentale, par la méthode overlap-add comme décrit dans [Allen et Rabiner, 1977]. Pour la normalisation en fréquence fondamentale, on choisit la valeur médiane de F_0 sur le fichier analysé. Cette normalisation a pour but de masquer un effet d'augmentation de la fréquence fondamentale liée à l'effort vocal.
- Normalisation en intensité sonore - RMS linéaire sur une fenêtre de 50ms - afin de masquer l'effet d'augmentation du niveau sonore avec l'augmentation de l'effort vocal.

Sur la figure 3.14, on retrouve les signaux et les spectrogrammes de tels signaux normalisés. On constate que malgré la normalisation, les spectrogrammes présentent une distribution différente, montrant qu'il reste des informations à extraire en plus de la notion de puissance RMS. Ces fichiers modifiés sont aussi disponibles sur le site web, les noms de fichiers sont `LOMA-NORMVoyelle1.wav`, `LOMA-NORMVoyelle2.wav`, `LOMA-NORMVoyelle3.wav` respectivement pour chacune des voyelles sus-mentionnées.

3.4.2 Energie cumulée sur la LoMA

L'énergie cumulée sur la LoMA peut servir à retrouver l'énergie de l'impulsion générant l'ODG. Comme la méthode retrace les maxima dans le temps, un alignement implicite est fait lors de l'ajout des maxima. Ce réarrangement des phases revient à procéder à un filtrage inverse en un point donné du signal sur lequel on possède *a priori* assez d'informations.

Cependant, il y a bien souvent un repliement des réponses impulsionnelles du filtre vocalique, plus longue que la période de l'ODG, ce qui rend difficile la mesure exacte de l'énergie de l'impulsion. Dans le cas de variations lentes de la fréquence fondamentale, ce repliement est quasi constant : on arrive donc, par le biais du gradient de l'énergie cumulée, à une image du gradient d'énergie de cette impulsion. Une des applications qui s'offre alors à l'exploitation de

1. http ://nicolas.sturmel.com/PHD/

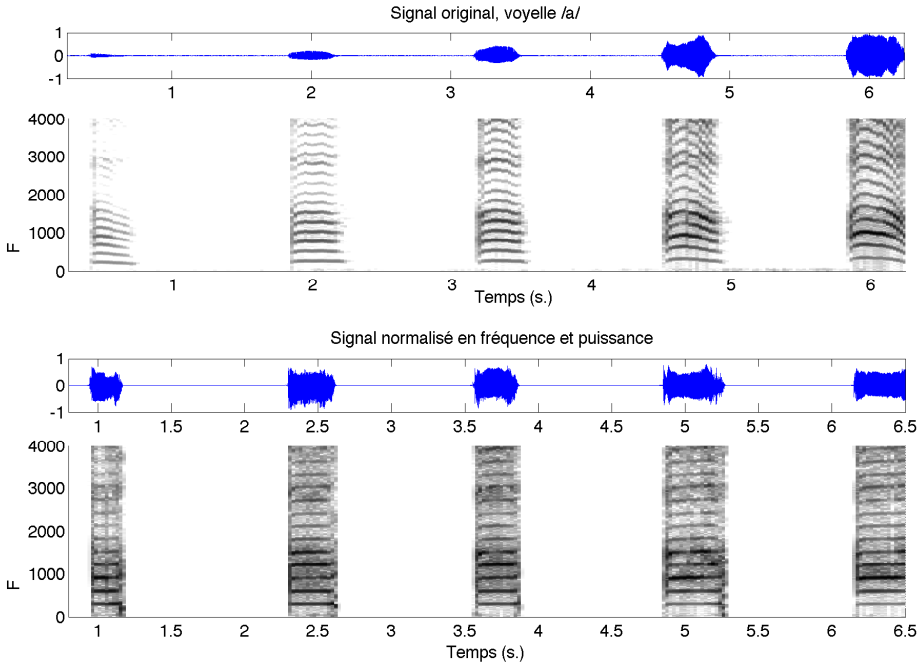


FIGURE 3.14 – Exemple d'un signal normalisé dans la base de données. De haut en bas : signal original et son spectrogramme, signal normalisé et son spectrogramme. Malgré la normalisation, il persiste des différences entre voyelles. Locuteur féminin, voyelle /a/ - fichier F8.wav.

cette facette de la méthode tient dans une mesure du shimmer. Seule cette application sera retenue dans cette étude. On calcule alors l'énergie totale E_T de la LoMA L_n :

$$E_T = \sum_{i=1}^{I} L_n(i)$$

Cette mesure sera appliquée à l'estimation du shimmer dans une prochaine section du chapitre.

3.4.3 Barycentre d'énergie de la LoMA

Une autre manière d'exploiter l'énergie de la ligne consiste à analyser sa répartition en fréquence. Un descripteur simple de la distribution d'énergie en fréquence est le barycentre, on peut chercher le barycentre B_n en Hz de la n-ième LoMA en fonction de l'indice i de la ligne - d'un total I, nombre de filtres dans le banc - de fréquence centrale f_i pour laquelle on mesure $E_n(i)$ l'énergie de la LoMA pour la ligne n, alors B_n a l'expression de l'équation 3.7.

$$B_n = \frac{\sum_{i=1}^{I} f_i E_n(i)}{\sum_{j=1}^{I} E_n(i)} \tag{3.7}$$

Ce barycentre va donner une idée de la pente spectrale du signal vocal. Plus le signal montre un spectre 'plat' sur le chemin reconstitué par la LoMA et plus le barycentre sera proche de $\frac{F_e}{2}$, la fréquence de Nyquist. A l'inverse, plus le signal sera pauvre en hautes fréquences, et donc avec une pente spectrale marquée, plus le barycentre sera proche de 0.

Un tel exemple est donné sur la figure 3.15. Sur un signal normalisé, la variation de $\tilde{B}_n$ est perceptible, donnant une image de la variation de la pente spectrale malgré la normalisation des signaux en puissance et fréquence.

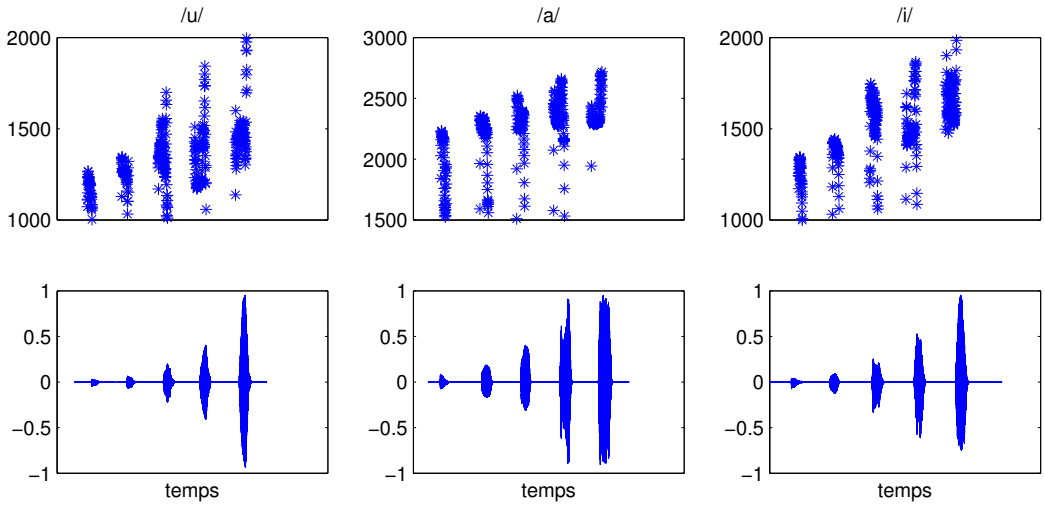


Figure 3.15 – Barycentre de la LoMA mesuré sur des signaux normalisés (en haut). Les signaux originaux (en bas) sont donné pour visualiser l'évolution de l'effort. Le barycentre de la LoMA est donné en Hz selon l'équation 3.7.

L'analyse de ces signaux nous montre la validité de l'utilisation du barycentre de la LoMA pour donner une image de la pente spectrale. Malgré la normalisation en fréquence fondamentale et en énergie, les signaux sont tout de même différentiables par ce barycentre comme en atteste la figure 3.15. Pour un certain niveau d'effort, il semble que la pente spectrale n'évolue plus. Une explication pourrait provenir des limites physiques du larynx à augmenter le rendement énergétique. Cette même augmentation de puissance non accompagnée d'une évolution de la pente spectrale se traduit par des apériodicités structurelles plus importantes : ceci montre ainsi encore les limites physiques du larynx.

On remarque aussi des valeurs différentes de barycentre en fonction des voyelles. Ce résultat est attendu dans la mesure où la LoMA est calculée sur le signal de parole naturel, et tient compte du conduit vocal. Il est donc naturel de trouver un barycentre plus haut sur les voyelles ouvertes (comme c'est le cas pour /a/) que sur les voyelles fermées (/i/ et /u/). Le deuxième formant des voyelles fermées étant placé plus haut en fréquence, il excitera une zone moins énergétique, alors que le premier formant placé relativement bas, aura tendance à ramener le barycentre vers 0Hz.

3.4.4 LoMA et distribution d'énergie : conclusion

L'analyse multi-échelles donne donc des informations sur la localisation temporelle d'un GCI, modélisé par une singularité du signal vocal, mais aussi sur son énergie, et sur sa distribution à travers les échelles de fréquences. La visualisation sur des signaux traités pour gommer un maximum de variations (figure 3.15) a notamment montré que le barycentre de chaque ligne d'amplitude maximum (LoMA) est un bon révélateur de la pente spectrale du signal. Cette méthode sera testée sur une grande base de données de voix expressive, à la fin du document.

Les LoMA donnent à la fois une position, mais aussi une amplitude pour chaque GCI. C'est donc une méthode appropriée à l'estimation des apériodicités structurelles sur des signaux vocaux. Dans la prochaine section, des expériences sur signaux synthétiques vont évaluer cette possibilité d'estimation.

3.5 Shimmer et jitter par les ondelettes

3.5.1 Base de données de signaux synthétiques

Afin de pouvoir contrôler la dose de jitter, de shimmer et de bruit additif influant sur la qualité de la détection des GCI et de l'énergie associée à chaque GCI par les LoMA, une base de données de signaux synthétiques a été générée selon une variation de paramètres présentée dans le tableau 3.3. La forme de la source glottique a été choisie constante et conforme aux observations pour des voix modales, à savoir $O_q = 0.5$ et $\alpha_m = 0.7$. Les voyelles utilisées pour la génération des signaux sont estimées à partir de signaux réels par LPC, une partie de la pente spectrale induite par le choix du paramètre Q_a de l'ODG est déjà présente dans l'expression du filtre. Ce paramètre est fixé à une valeur de $Q_a = 0.05$. La forme du bruit additif retenu est un bruit pulsé sur 30% de la période du signal généré, centré autour du GCI. Les signaux générés ont une durée de 1 seconde et une fréquence d'échantillonnage de 32kHz.

Table 3.3 – Récapitulatif des paramètres utilisés pour la génération de la base de données de signaux synthétiques testant l'estimation du Jitter et du Shimmer par la méthode des LoMA.

Voyelle	/a/, /i/, /u/
F_0 en Hz	$[90, 120, 150, 180, 210, 240, 270, 300]$
Jitter en % de période	$[0.5, 0.75, 1.13, 1.69, 2.53, 3.80, 5.70, 8.54, 12.81, 19.22, 28.83]$
Shimmer en % d'amplitude	$[0, 5, 10, 15, 20]$
RSB en dB	$[-\infty - 30 - 20]$

Les apériodicités structurelles sont générées par la fonction `randn()` de MATLAB© qui est un générateur de bruit Gaussien de variance 1. Cette variance est modifiée par la valeur du jitter et du shimmer donnée en % de variation (en fonction de la période ou de l'amplitude de l'ODG). Ainsi, les apériodicités générées ne sont pas exactement égales aux paramètres et on retrouve donc sur la figure 3.16 une distribution des valeurs relativement continue.

Il semble toutefois important de préciser que les notions de jitter et shimmer ne sont dissociables que du point du vue de l'analyse du signal. Au niveau de la production de la voix, il semble peu probable que les phénomènes soient totalement indépendants. Des micro-changements de la période du voisement entraînent probablement avec eux des changements de l'intensité.

3.5.2 Mesure du jitter

Pour estimer le jitter, on estime d'abord les instants de fermeture glottique sur le signal synthétique par la méthode des LoMA. Cette série de GCI est ensuite dérivée (au sens mathématique du terme) pour avoir $u(n)$ de taille N, regroupant les longueurs de périodes glottiques. On estime alors le jitter J à l'aide de l'estimateur décrit dans [Vasilakis et Stylianou, 2009] et donné en l'équation 3.8.

$$J = \frac{1}{N-1} \sum_{n=0}^{N-2} |u(n+1) - u(n)| \tag{3.8}$$

Les résultats de l'estimation du Jitter sur le corpus de signaux synthétiques sont présentés en figure 3.16 sur une échelle logarithmique. La plage de variations proposée étant importante, cette représentation semblait plus appropriée. Sur la gauche on retrouve les résultats pour des signaux sans shimmer, à droite des signaux avec shimmer. En haut on retrouve les signaux sans bruit additif pulsé, et en bas les signaux avec bruit additif pulsé.

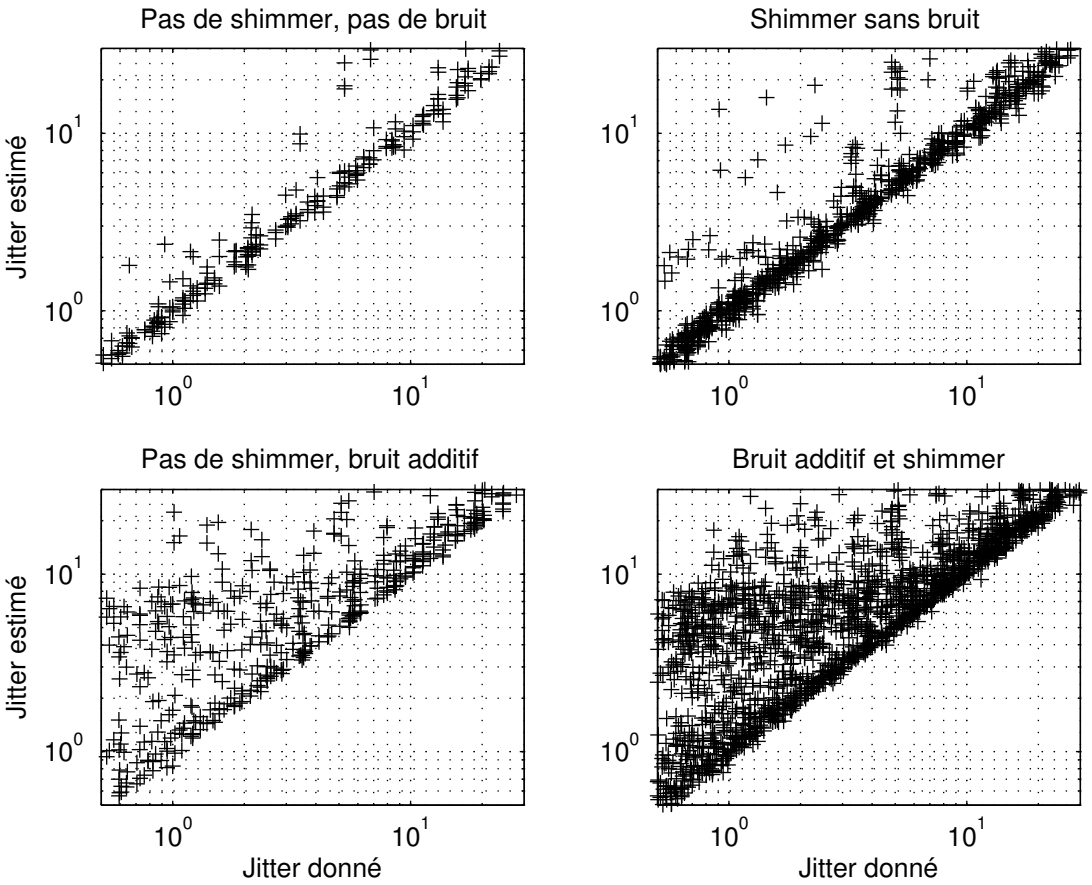


Figure 3.16 – Estimation du jitter par la méthode LoMA sur la base synthétique de signaux.

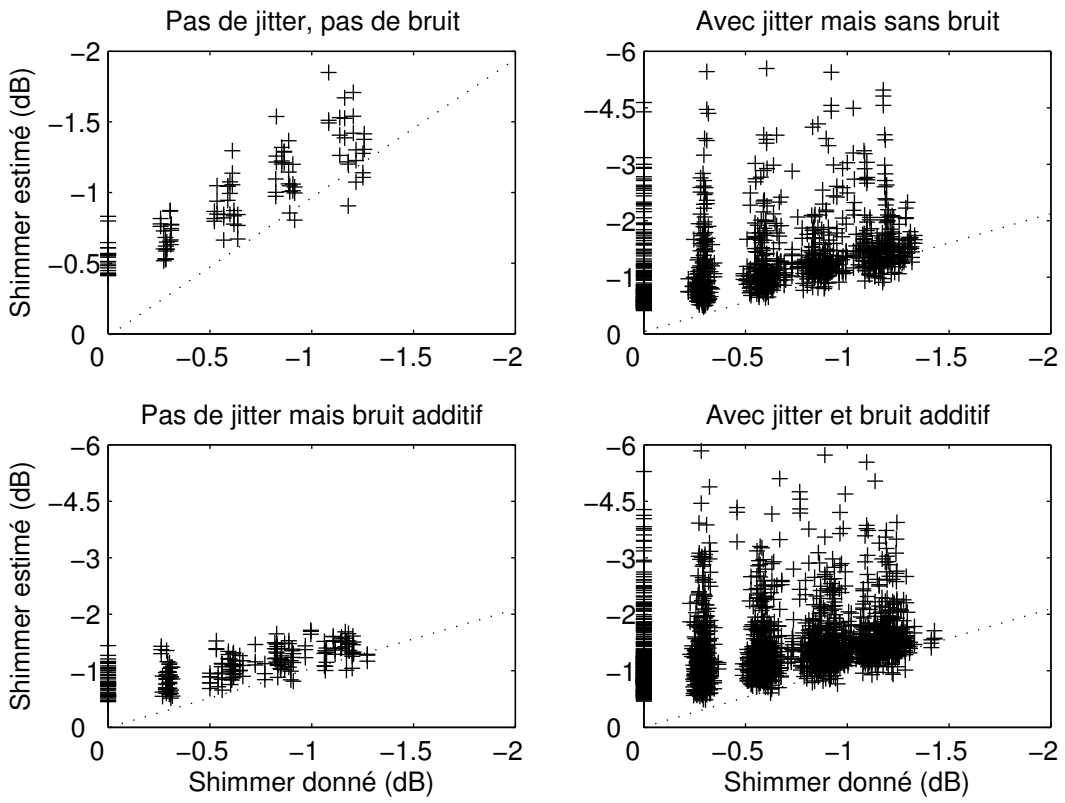


Figure 3.17 – Estimation du shimmer par la méthode LoMA sur la base synthétique de signaux.

Sur la figure en haut à gauche, sans shimmer ni bruit, la majorité des résultats se retrouve autour de l'identité entre jitter estimé et jitter donné. Ces résultats témoignent d'une immunité par rapport aux variations de voyelle ou de fréquence fondamentale.

On remarque que la présence de bruit est un facteur déterminant dans l'estimation correcte du Jitter sur un signal. La ligne basse de la figure 3.16 présente une distribution de résultats éparpillée avec une tendance à surestimer le jitter en présence de bruit. Ce comportement est attendu, étant donné que l'ajout de bruit va avoir tendance à diminuer la précision de localisation en temps des maxima sur les échelles des ondelettes. Un tel manque de précision induit donc une surestimation du jitter.

Quelle que soit la valeur du bruit, on remarque qu'il fait généralement dévier l'estimation du jitter. Plus le bruit est important et plus la déviation est importante. L'effet du Shimmer sur l'estimation du Jitter est négligeable. De droite à gauche les distributions varient peu et on retrouve un regroupement très fort des résultats autour de la ligne représentant l'identité entre jitter donné et jitter estimé. Au chapitre 4, un algorithme de séparation harmonique + bruit sera proposé, il sera vu alors dans quelle mesure cette décomposition améliore ou pas l'estimation des apériodicités structurelles.

3.5.3 Mesure du shimmer

Sur la même base que la mesure du jitter, les mesures du shimmer sont présentées sur la figure 3.17. Le shimmer est calculé en décibels avec l'estimateur suivant de l'équation 3.9 pour

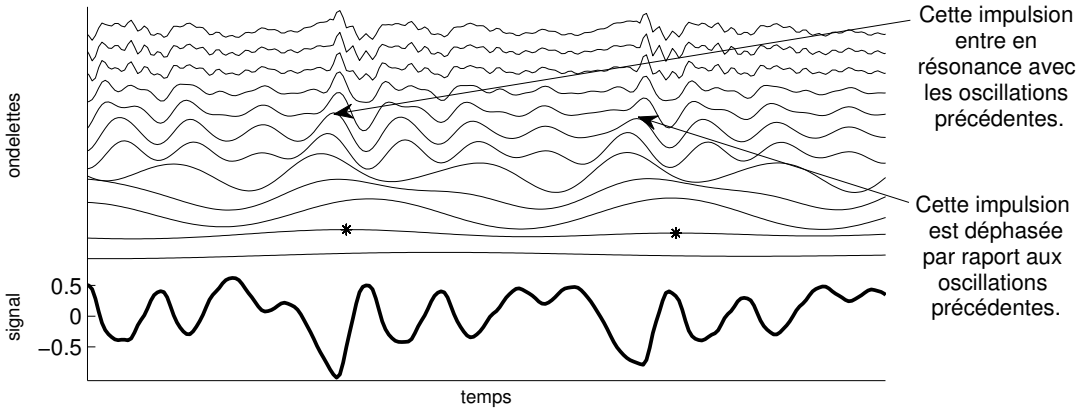


Figure 3.18 – Effet d'une variation locale de la période de voisement - jitter - sur l'amplitude de voisement. Le repliement de la réponse des filtres peut causer des variations de l'amplitude de voisement.

la suite d'amplitudes $a(n)$ de taille N regroupant les amplitudes glottiques,

$$S = 20log_{10}\left(1 - \frac{1}{N-1}\sum_{n=0}^{N-2}|a(n+1) - a(n)|\right) \tag{3.9}$$

Ainsi sur la figure 3.17, on retrouve $S_{dB} = 20log_{10}(1-S)$ comme la différence d'amplitude en dB par rapport à une amplitude de référence - l'unité. Afin d'aligner les amplitudes de référence, la suite d'amplitudes issue de la méthode LoMA est normalisée par sa moyenne.

On retrouve le même agencement des résultats en fonction des paramètres de synthèse sur la figure 3.17 que sur la figure 3.16 : sur la colonne de gauche pour l'absence de jitter (contre la présence de jitter pour la colonne de droite) et sur la ligne supérieure l'absence de bruit (contre la présence de bruit pour la ligne inférieure). Sur toutes les figures, les traits en pointillés donnent l'identité entre Shimmer donné et Shimmer estimé par les LoMA.

En l'absence d'autres perturbations que le Shimmer, un léger biais est visible sur la figure, empêchant une estimation précise dans le cas de signaux faiblement perturbés (Shimmer $<$ 1dB). De même, l'ajout de bruit perturbant la mesure n'a que très peu d'effet sur la précision de la mesure. Ce résultat est attendu, étant donné le niveau très faible du bruit par rapport à la variation d'amplitude (-20dB contre -2dB), même si le bruit provoque des perturbations dans le placement temporel des maxima fréquentiels (perturbant la localisation temporelle et donc l'estimation du Jitter) il ne perturbe que très peu l'amplitude de ces maxima, et donc la mesure de l'amplitude associée au GCI.

L'ajout de Jitter aux signaux de test montre une forte dispersion des résultats de l'analyse. La variation de la période glottique agit directement sur la manière dont les réponses impulsionnelles du filtre vont se superposer, et influe donc sur l'amplitude mesurée au GCI. On retrouve ici une forte interaction à sens unique entre jitter et shimmer : la présence de jitter empêche l'estimation convenable du shimmer. Un exemple de repli des réponses impulsionnelles (figure 3.18) montre clairement la variation d'énergie au moment du GCI.

On pourrait chercher à estimer le shimmer et le jitter sur le résidu de la LPC, pour tenter d'effacer l'effet du repliement des réponses impulsionnelles du conduit vocal. Les expériences en question n'ont pas donné de résultats pertinents et ne seront donc pas présentées ici.

3.5.4 Conclusion sur l'aptitude de la méthode LoMA à déterminer des apériodicités structurelles

Permettant d'avoir à la fois la position et une image de l'énergie de chaque instant de fermeture glottique, la méthode LoMA semble être un bon choix pour l'estimation des apériodicités structurelles. Les tests sur signaux synthétiques nous montrent que, à l'image de la réalité de la production vocale, les notions de jitter et shimmer interagissent entre elles et rendent difficile l'estimation d'une seule des composantes par les LoMA, la présence de shimmer a cependant beaucoup moins d'impact sur l'estimation du jitter.

Si la précision de l'estimation du jitter dépend principalement du bruit additif aux signaux de parole, le shimmer quant à lui est très sensible au jitter. L'effet de la réponse impulsionnelle du filtre n'est pas à négliger dans ce phénomène et l'utilisation d'un filtrage inverse en amont ne résout pas nécessairement le problème. Dans le cas favorable où le bruit additif pourrait être séparé du signal vocal, la méthode deviendrait alors extrêmement efficace pour la mesure du Jitter, ouvrant alors la possibilité d'y estimer le Shimmer.

3.6 Quotient ouvert et ondelettes

Les travaux récents de Bouzid et al. [Bouzid et Ellouze, 2007] ont mis en lumière la possibilité de détecter l'instant d'ouverture glottique (GOI - *Glottal Opening Instant*) par une analyse en ondelettes : le produit multi-échelles. Les résultats avancés par ce produit multi-échelles sont de l'ordre de 75% de bonnes détections pour une plage de détection de 0.5ms autour du GOI sur la base de données Keele [Plante *et al.*, 1994].

Une autre approche, proposée par [Thomas *et al.*, 2009] emprunte aussi les ondelettes, mais dans un contexte propre à l'algorithme de programmation dynamique Dypsa. Encore une fois, les résultats donnés sont satisfaisants, mais présentent la même plage d'erreur d'estimation. Un écart type de 1ms est donné sur le corpus d'analyse, ce qui est trop large dans le cas de la mesure du quotient ouvert. Cet écart type équivaut à une erreur de 50% pour un quotient ouvert de 0.5 et un signal de 250Hz de fréquence fondamentale.

Cette section explore les possibilités de mesurer le quotient ouvert à l'aide des lignes d'amplitude maximum. Après avoir analysé un modèle simplifié de production vocale pour en prévoir la forme des lignes, le principe de mesure sera présenté sur des signaux de tests. Les mesures sur signaux réels seront présentées au chapitre 5 dédié à l'extraction des paramètres de source.

3.6.1 Forme des lignes et forme du débit glottique

Jusqu'à présent, seule la position et la valeur des maximas ont été exploités. La forme de la LoMA, donnant une image du déphasage moyen des harmoniques pour chaque bande d'analyse, possède probablement des informations pertinentes sur la forme de l'onde de débit glottique.

Afin de déterminer la forme des lignes, il faut s'intéresser à la fois au retard de phase $\tau_\Phi = -\frac{\Phi(\omega)}{\omega}$ et au retard de groupe $\tau_g = -\frac{d\Phi(\omega)}{d\omega}$. Dans quelle mesure ces lignes, tracées sur le signal de la DODG, donnent-elle une image de ces paramètres ? Dans quelle mesure les modifications appliquées à ce signal (filtrage du conduit vocal, banc de filtres en ondelettes) modifient-elles la forme de ces lignes ?

Réponse en basse fréquence

Considérons tout d'abord la plus basse bande du filtre en ondelettes excitée par le signal. En vertu de l'invariance d'échelle, et par propriété des bancs de filtres en ondelettes dyadiques, on

considère en première approximation que seul le premier harmonique (fréquence fondamentale du signal) est présent dans cette bande. En réalité, l'ondelette choisie ne rejette pas intégralement les harmoniques suivants, mais suffisamment pour valider cette hypothèse.

Soit le signal $s_i(t)$ qui est issu de la bande i du filtre contenant le premier harmonique, alors :

$$s_i(t) = h_i(t) * f(t) * dg(t) * \delta_{T_0}(t)$$

Alors sa transformée de Fourier s'écrit (H_i étant à phase nulle) :

$$\begin{aligned} S_i(\nu) &= H_i(\nu)F(\nu)dG(\nu)\delta_{\nu_0}(\nu) \\ S_i(\nu) &= |F(\nu)dG(\nu)|e^{i(\Phi_F(\nu)+\Phi_{dG}(\nu))}\delta(\nu_0-\nu)|H_i(\nu_0)| \end{aligned}$$

Et par conséquent :

$$s_i(t) = |F(\nu)dG(\nu)||H_i(\nu_0)|cos(2\pi F_0(t-\tau_\Phi(F_0)))$$

La position des maxima sur l'échelle contenant le premier harmonique dépend donc du retard de phase τ_Φ. Le retard de groupe τ_g n'intervient pas dans cette expression car la contribution du filtre en basse fréquence est négligeable par rapport à la source. En pratique, la contribution du retard de groupe sera vue à la section suivante lors de simulations numériques.

Réponse en haute fréquence

Pour les plus hautes échelles d'analyse, on trouve de multiples harmoniques dans chaque bande. On fait alors l'hypothèse que ces harmoniques sont modulés par une enveloppe $A_i(t)$ de fréquence ω_i et de phase Φ_i :

$$s_i(t) = x(t) * A_i(t)cos(\omega_i t + \Phi_i)$$

Il vient alors :

$$s_i(t) = |X(\omega_i)|A_i(t-\tau_g(\omega_i))cos(\omega_i(t-\tau_\Phi(\omega_i))$$

Dans ce cas, le retard de groupe intervient de manière conséquente dans la position temporelle des maximas, mais comme τ_Φ et τ_g sont des fonctions qui varient avec l'inverse de la fréquence, leur contribution en haute fréquence est négligeable (variant en $\frac{1}{\omega}$). On conclut donc que la position haute fréquence de la LoMA pointe vers un déplacement temporel nul.

Ainsi, on définit le décalage du premier harmonique par rapport au GCI comme la valeur $\tau_\Phi(F_0)$. La prochaine section établira le lien entre ce décalage et la valeur des paramètres de la source glottique.

3.6.2 Principe de l'extraction de la configuration glottique par ondelettes

En introduction de ce chapitre, l'analyse en ondelettes a montré des formes caractéristiques liées à la configuration glottique. Nous allons montrer que le déplacement basse fréquence $\tau_\Phi(F_0)$ des maxima peut être lié à la valeur du quotient ouvert. Ce déplacement a été défini à la section précédente. En visualisant les profils de décalage temporel pour une période de DODG par le modèle LF comme sur la figure 3.19, on remarque que pour chaque fréquence du spectre, un décalage dépendant de la configuration glottique est induit. Sur cette figure, la variation de forme de la LoMA est donnée en haut pour un quotient ouvert variable et en bas pour une asymétrie variable. Intéressons-nous au décalage autour du premier harmonique, unique harmonique à être présent dans une bande d'analyse en ondelettes.

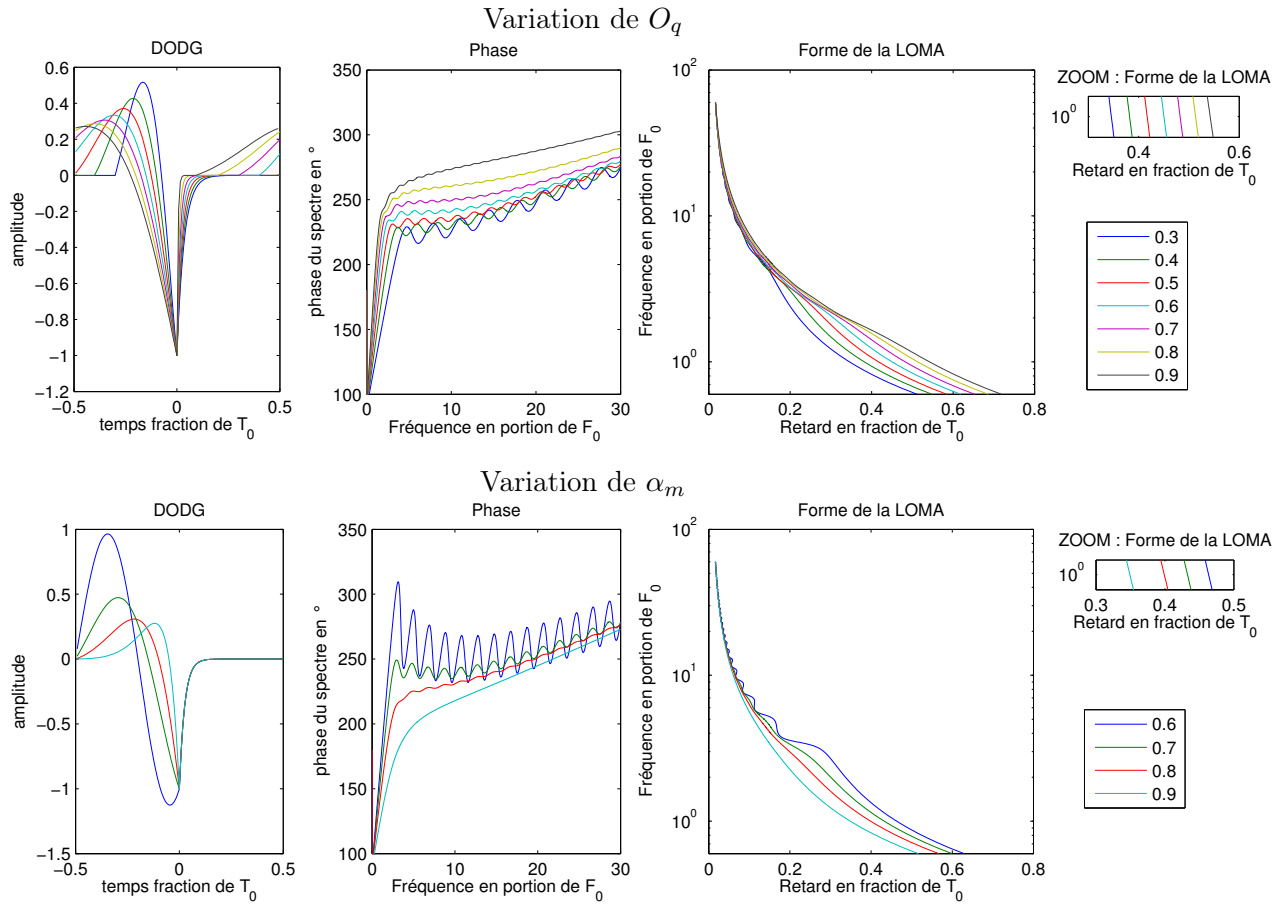


Figure 3.19 – Prédiction de la forme de la LoMA pour un débit glottique non filtré. Haut : variation du quotient ouvert, Bas : variation de l'asymétrie. De gauche à droite : forme de la DODG, phase du spectre en degrés, représentation temps fréquence du délai dû à la phase (prédiction de la forme de la LoMA) et zoom sur cette représentation au niveau de la fréquence fondamentale. Les temps et fréquences sont normalisés.

Ces simulations sont données pour une fréquence fondamentale de 133Hz (échantillonnage à 16kHz - donc période de 120 échantillons), mais en pratique les valeurs données sur la figure 3.19 sont indépendantes de et normalisées par la fréquence fondamentale. Pour faciliter la visualisation, un zoom est donné autour de la fréquence fondamentale (10^0). Ces simulations permettent de conclure qu'il faut s'attendre à trouver un décalage τ_Φ du premier harmonique compris entre $0.33T_0$ et $0.55T_0$ soit entre 40 et 66 échantillons pour un quotient ouvert variant respectivement entre 0.3 et 0.9. En ce qui concerne la variation de α_m, on constate qu'elle engendre une amplitude de décalage du premier harmonique nettement inférieure à celle due à O_q. Dans une plage de variation typique pour de la voix naturelle, l'asymétrie est comprise entre 0.7 et 0.8 ; le décalage varie alors de 0.4 à 0.44, soit 6 fois moins que pour O_q (décalage de 0.3 à 0.55). Le décalage du premier harmonique devrait donc permettre d'estimer assez précisément (avec une précision inférieure au seuil différentiel perceptif) les valeurs de O_q, mais avant de procéder à de tels tests il convient de déterminer la formule liant τ_Φ et O_q. Pour faciliter la lecture, définissons $\delta = \tau_\Phi(F_0)$. Il s'agit maintenant de simuler la production vocale dans son intégralité en ajoutant un filtre de conduit vocal.

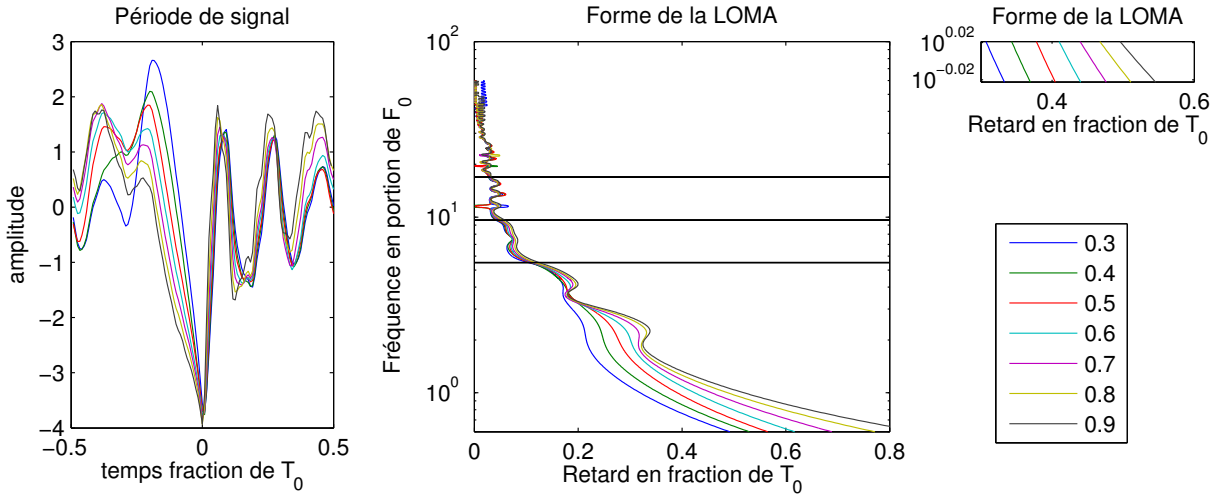


Figure 3.20 – Prévision de la forme de la LoMA sur un signal synthétique complet (DODG filtrée). La phase du spectre n'est pas montrée ici (illisible), mais on constate que les prévisions restent identiques, à quelques oscillations prêt autour des formants (traits noirs sur la figure du milieu). L'amplitude du déplacement ne change pas non plus. Variation du quotient ouvert de 0.3 à 0.9.

Ajoutons maintenant un filtre à cette dérivée d'onde de débit glottique, et visualisons la forme d'onde résultante. Une telle observation est donnée sur la figure 3.20, où le signal de la figure 3.19 est filtré par une voyelle /a/. On constate alors deux phénomènes :

- On prédit des oscillations supplémentaires sur les LoMA, dues aux formants du filtre vocalique.
- L'amplitude de déplacement du premier formant ne change pas.

Les prévisions précédentes sur la relation entre déplacement du premier formant et valeur du quotient ouvert restent donc identiques. On retrouve en outre une influence négligeable des basses fréquences du filtre vocalique. Il est cependant nécessaire de tenir compte d'un dernier phénomène : le retard de groupe $\tau_g = \frac{d\Phi}{d\omega}$. Si ce dernier est faible en basse fréquence, il est tout de même présent et peut nécessiter un ajustement supplémentaire. La figure 3.21 nous montre que le retard de groupe moyen pour l'amplitude de variation du quotient ouvert est de 0.1. Une constante de correction est donc nécessaire dans la fonction qui lie O_q à δ.

Pour réaliser une telle mesure il faut que le formant glottique soit séparé et inférieur à la fréquence du premier formant vocalique. Dans le cas contraire, le conflit entre ces deux formants, en particulier dans le cas où ils ne sont pas temporellement séparables (i.e. : repliement des réponses impulsionnelles), pourrait mener à une mesure erronée.

L'idée proposée est d'appliquer un filtrage inverse sans préaccentuation en amont de l'analyse par ondelettes pour bénéficier des propriétés suivantes :

- la précision sur la localisation des GCI est beaucoup plus importante avec une décomposition LPC en amont, au prix d'un taux plus important de GCI non trouvés.
- Le filtrage inverse par autocorrélation ne permet pas d'estimer correctement la phase de la partie anticausale de la DODG. Cette inaptitude est utilisée ici comme une propriété : l'absence de préaccentuation donne un spectre plat, mais déphasé. En conséquence, le décalage du premier harmonique devrait être plus facilement observable.

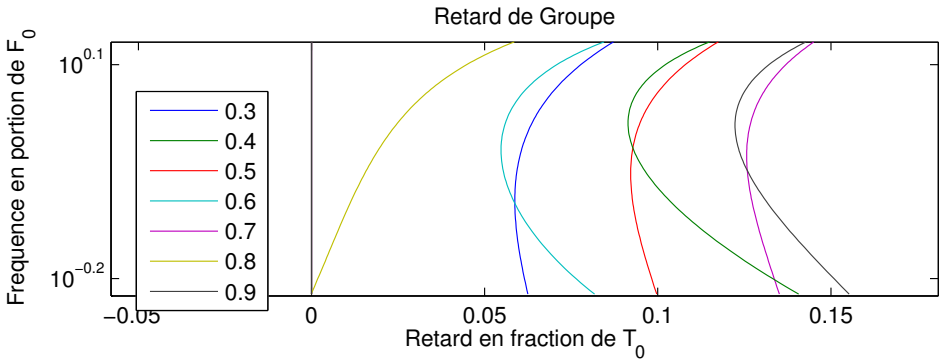


Figure 3.21 – Retard de groupe du signal filtré par rapport à l'instant de fermeture glottique autour de la fréquence fondamentale (unité). Variation du quotient ouvert de 0.3 à 0.9.

– l'absence de préaccentuation enlève un *a priori* fort sur l'estimation de la forme de l'onde de débit glottique tout en permettant de retrouver un de ses paramètres de forme.

Cette analyse directe du résidu de la LPC est utilisée sur la figure 3.22.

Au final, cette mesure revient à calculer la différence entre la position temporelle de la racine de chaque LoMA et le GCI déduit de cette LoMA, il s'agit du retard δ. Les tests sur des signaux synthétiques permettront de déduire la meilleure fonction $O_q = f(\delta)$.

3.6.3 Validation par un signal synthétique

Une fois le cadre de l'analyse défini, un test sur des signaux synthétiques s'impose. Dans un premier temps, on confirme par la pratique la linéarité entre le déplacement du premier harmonique et les paramètres de la configuration. La figure 3.22 détaille une telle analyse sur un signal synthétique (/a/ à 133Hz - 120 échantillons) pour lequel le quotient ouvert varie selon une loi quadratique. Le décalage mesuré n'est pas un retard, mais un avancement : le décalage est donc connu modulo T_0.

Modulo T_0, on retrouve les décalages prévus par l'analyse du spectre de la DODG sur la figure 3.19. Pour $O_q = 0.3$, on trouve un décalage de 80 échantillons, soit T_0 soustrait des 40 échantillons prédits précédemment pour cette valeur du quotient ouvert. Pour $O_q = 0.9$, on trouve un décalage d'environ 50 échantillons, soit un peu moins que les prévisions précédentes de $T_0 - 66 = 54$ échantillons. Dans l'ensemble, on constate que la courbe de décalage suit fidèlement la courbe de variation de O_q sur toute la plage.

Des tests approfondis sur des signaux synthétiques ont permis de déterminer la formule suivante pour lier ce décalage à la valeur de O_q :

$$O_q = 0.5 + \frac{\delta}{T_0}$$

La correction de 0.5 est un ajustement dû au retard de groupe visible sur la figure 3.21 combiné au fait que le retard soit mesuré modulo T_0. Cette formule n'est strictement vraie que pour une valeur typique de α_m (0.75), valeur qui a été utilisée pour générer le signal de la figure 3.22. Cependant, les excursions autour de cette valeur (pour de la parole naturelle) étant très faibles -environ 0.05 mesuré sur le bas de la figure 3.19 -, on s'expose à une erreur de mesure sur le décalage de 0.02 T_0, soit extrêmement faible vis à vis de la précision généralement attendue

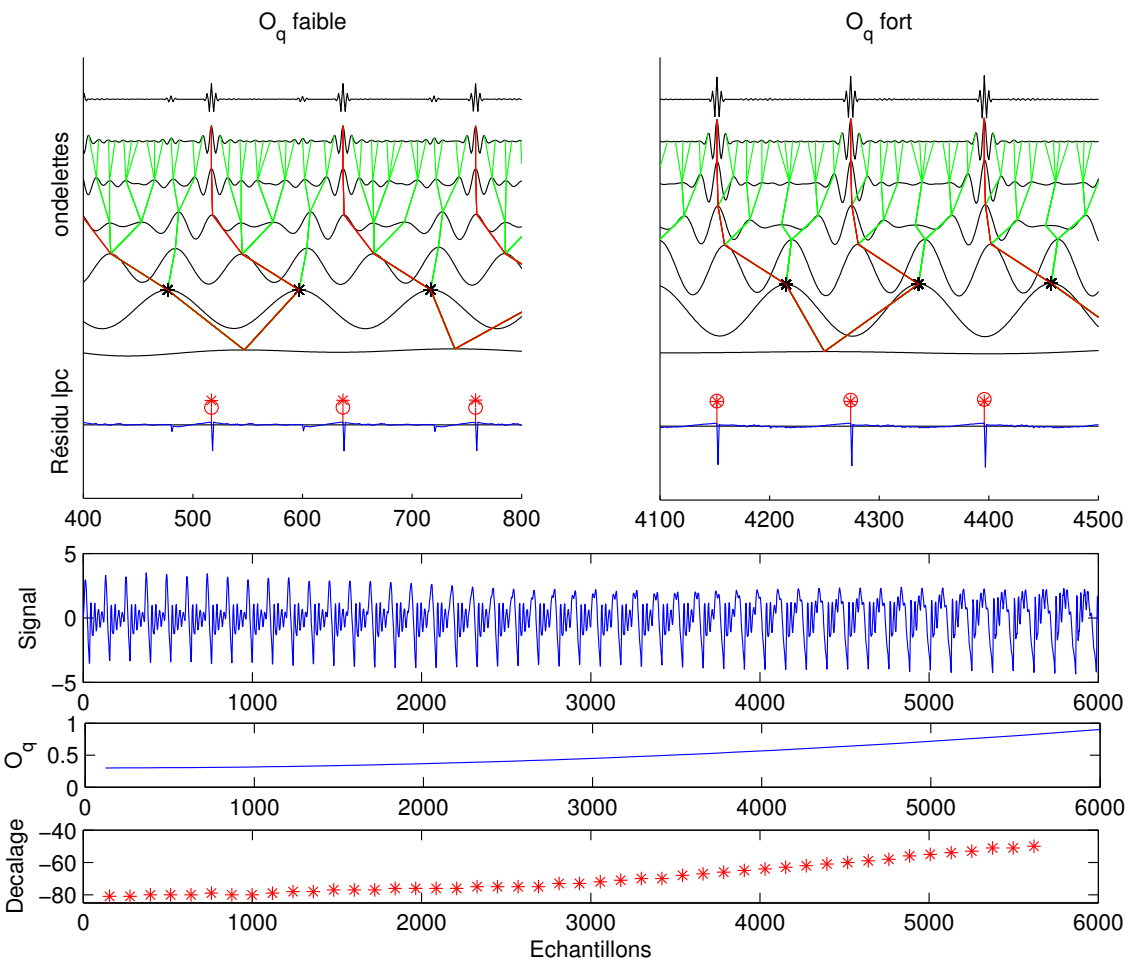


Figure 3.22 – Analyse d'un signal synthétique (voyelle /a/, 133Hz - 120 échantillons) avec O_q variant de 0.3 à 0.9 selon une loi quadratique. Deux zooms sur l'analyse en ondelettes sont donnés au début et à la fin du signal. Le décalage varie linéairement avec O_q.

sur une mesure de quotient ouvert. Précisons que le seuil différentiel perceptif mesuré pour O_q [Henrich *et al.*, 2003] est de 17%.

3.6.4 Tester l'algorithme plus en détail

La mise en oeuvre d'une telle mesure de quotient ouvert se révèle relativement simple dans la mesure où les points composant les LoMA (valeurs des maxima et positions temporelles) sont connus. Une évaluation de cette méthode de mesure du quotient ouvert sur une grande base de données de signaux de parole sera donc proposée au chapitre 5.

3.7 Parallèle avec Mean Square Phase

Degottex et al. ont aussi proposé une méthode d'estimation de GCI et de la forme de l'ODG par minimisation de la phase [Degottex *et al.*, 2010], dénommée MSP (*Mean Square Phase* -

phase quadratique moyenne) mais référencée ici par MSPh pour éviter toute confusion avec le produit multi-échelles MSP. Il s'agit de trouver le minimum de la fonction MSPh qui calcule la moyenne de la phase quadratique de $R_k^{(\theta,\Phi)}$ qui n'est autre que la déconvolution du signal original par sa version à phase minimum.

On peut voir plusieurs parallèles entre son approche analytique et notre approche pratique :

- Dans [Degottex *et al.*, 2010] il est clairement expliqué que le but du MSPh est de retrouver au plus près la forme d'un Dirac. C'est aussi le but des lignes d'amplitude maximum.
- La méthode cherche à minimiser la phase observée pour en déduire conjointement la position du GCI et la forme de la DODG. Notre algorithme détectant les LoMA fait la même chose, dans la mesure où il cherche le chemin le plus court du maximum du premier harmonique jusqu'aux hautes fréquences.
- La méthode MSPh se base uniquement sur les bins harmoniques. La représentation en ondelettes favorise des maxima placés en rapport avec les harmoniques.

La similarité des deux approches est révélatrice des pistes couramment explorées en matière d'estimation et d'analyse des signaux de parole. Ceci conforte aussi les méthodes développées en particulier sur des paramètres non accessibles directement par MSPh : le barycentre de la LoMA donnant une idée de l'effort vocal et l'amplitude de chaque GCI. L'approche décrite [Degottex *et al.*, 2010] a néanmoins l'avantage d'une formulation analytique très claire qui permet probablement une robustesse accrue de la détection des GCI.

3.8 Conclusion

Dans ce chapitre, l'accent a été mis sur la possibilité d'analyse des signaux vocaux par l'utilisation des bancs de filtres en ondelettes. Les travaux précédents [Tuan et d'Alessandro, 1999, Bouzid et Ellouze, 2007] donnent des indications fortes sur la possibilité de développer une méthode capable de détecter les instants de fermeture glottique et de mesurer les apériodicités structurelles du signal vocal, telles que le jitter et le shimmer.

Dans un premier temps, un nouvel algorithme de détection des LoMA a été proposé, utilisant une version plus robuste de la version précédemment proposée [Tuan et d'Alessandro, 1999] : toutes les lignes sont calculées et une bande racine est sélectionnée en fonction d'une valeur grossière de la fréquence fondamentale. L'algorithme proposé est alors dépendant d'un indicateur d'octave pour trouver la fréquence fondamentale : un simple estimateur de pitch par autocorrélation du signal peut suffire, mais une estimation précise comme celle fournie dans Yin [de Cheveigné et Kawahara, 2002] permet des résultats de meilleure qualité sur de la parole naturelle pour laquelle les variations de F_0 sont nombreuses.

Les tests de l'algorithme sur une base de données de voix parlée montrent une capacité à estimer précisément l'emplacement des instants de fermeture glottique. En comparaison avec une autre méthode réputée pour sa fiabilité (DYPSA [Naylor *et al.*, 2007]) et en tenant compte de la précision atteignable par l'utilisation de l'EGG comme référence, la méthode proposée présente des performances tout à fait équivalentes. DYPSA a toutefois recours à de nombreux traitements en amont et en aval auxquels LoMA ne fait pas appel, ce qui permet d'avancer que les lignes d'amplitude maximum ont un potentiel certain. L'application de l'algorithme d'estimation sur le résiduel de la LPC a montré une robustesse moindre pour des résultats plus précis.

Dans un deuxième temps, nous avons montré que les LoMA pouvaient être utilisées pour tirer des informations d'énergie au niveau du GCI sur le signal vocal. Le barycentre de la distribution des amplitudes en fréquence a notamment été exploré comme indicateur de l'effort vocal sur des signaux normalisés. Cet indicateur est cependant sensible au filtre vocalique.

La mesure des apériodicités étant difficile sur des signaux réels, cette partie a été traitée sur une grande base de signaux synthétiques. Les résultats montrent que cette méthode peut très bien s'appliquer à la mesure de la variation locale de période fondamentale : le Jitter. Cependant, cette mesure est extrêmement sensible à la présence de bruit stochastique dans le signal.

La variation locale d'amplitude de l'impulsion glottique peut, quant à elle, aussi être approchée par la méthode proposée. La mesure sur des signaux synthétiques, de moins bonne qualité par rapport à la mesure du jitter, reste exploitable. Cependant, il a été observé que cette mesure dépend fortement du Jitter.

La faisabilité d'une mesure du quotient ouvert par les ondelettes a aussi été discutée. En déterminant analytiquement ce qui causait la forme de la LOMA, nous avons lié le décalage entre le GCI mesuré et le premier harmonique à la valeur du quotient ouvert. Une relation empirique entre ce décalage et la valeur de O_q a été proposée. Une analyse complète sur signaux synthétiques et des exemples sur signaux réels ont été réalisés. L'application de cette relation à des signaux réels sera vue dans un prochain chapitre.

Finalement, l'analyse multi-échelles de la parole se révèle très efficace dans l'extraction d'informations sur la qualité vocale. Malgré la simplicité de l'analyse et de sa mise en œuvre, les résultats sont extrêmement satisfaisants. En exploitant simultanément des propriétés temporelles et spectrales, quatre paramètres différents sont extraits de l'analyse : l'amplitude du GCI, l'instant de fermeture glottique, le quotient ouvert et le barycentre de distribution de l'énergie du GCI. Les résultats de cet algorithme sur des signaux réels n'ont cependant été évalués que sur la position du GCI et des tests supplémentaires sont nécessaires.

Résumé

En bref...

Un nouvel algorithme d'analyse multi-échelles pour estimer les lignes d'amplitude maximum (LoMA) a été proposé. Les LoMA permettent de mesurer l'amplitude et la localisation de chaque instant de fermeture glottique. Ce travail s'inspire des travaux précédents [Tuan et d'Alessandro, 1999, Bouzid et Ellouze, 2007] proposant aussi l'analyse multi-échelles appliquée aux signaux vocaux. Le nouvel algorithme propose de nouvelles conditions pour la programmation dynamique et la connaissance *a priori* de la fréquence fondamentale pour une analyse plus robuste.
Cet algorithme est testé sur une base de signaux réels et sur des signaux synthétiques dans le cas de l'estimation des GCI et des apériodicités structurelles de la voix. Dans le cas de l'estimation des GCI, LoMA est comparé à la méthode DYPSA, [Naylor *et al.*, 2007] connue pour sa précision et sa stabilité.

Les résultats montrent

L'analyse du corpus de voix parlée montre que l'algorithme proposé (détection des LoMA) présente de bons résultats. Les résultats d'estimation du GCI sont, au regard de la précision de la référence choisie (le signal EGG), considérés comme comparables à ceux de DYPSA. Sur les signaux synthétiques, où les apériodicités structurelles sont connues et contrôlées, les résultats montrent qu'en l'absence de bruit stochastique, LoMA est un bon estimateur du jitter. Concernant le shimmer, le repliement de la réponse impulsionnelle du filtre vocalique cause quelques problèmes quand Jitter et Shimmer sont liés (ce qui est le plus probable sur des signaux réels). L'estimation du quotient ouvert, basée sur le principe simple du décalage du premier harmonique, donne des résultats satisfaisants qui nécessitent un approfondissement de cette approche. Cet approfondissement sera vu au chapitre 5.

En conclusion

La méthode LoMA se trouve être une alternative intéressante aux méthodes existantes pour la détection des instants de fermeture glottique. Encore pauvre en optimisations, cette méthode révèle déjà un grand potentiel. Elle présente des résultats comparables à une méthode au développement avancé comme DYPSA, sans pour autant avoir recours aux mêmes *a priori* sur la forme du signal.
Enfin, l'estimation du quotient ouvert par le décalage du premier harmonique est une voie intéressante pour extraire conjointement instant de fermeture glottique et quotient ouvert. Cette approche sera vue plus en détails au chapitre 5.

Chapitre 4

Décomposition Périodique/Apériodique

Sommaire

4.1 Amélioration de l'algorithme PAP 110
4.1.1 Choix de la méthode 110
4.1.2 Adaptation de la taille d'analyse à la fréquence fondamentale 111
Taille de la fenêtre idéale 112
Application aux conditions réelles d'observation 112
4.1.3 Fonction de coût 113
4.1.4 Algorithme PAP-A 114
4.2 Application à des signaux de tests 114
4.2.1 Critère d'évaluation 116
4.2.2 Résultats des signaux de tests 117
4.2.3 Discussion 117
Sans apériodicités structurelles 117
Avec un léger Shimmer ou un léger Jitter 118
Pour des apériodicités plus importantes 119
4.2.4 Conclusion sur l'analyse de signaux synthétiques 119
4.3 Application à des signaux réels 120
4.3.1 Voyelles tenues 120
4.3.2 Fricative voisée 122
4.3.3 Comparaisons perceptives et de spectrogrammes 123
4.4 Impact de la décomposition sur l'estimation des LoMA 123
4.5 Conclusion 127

La quantité de voisement est un facteur important de la qualité vocale, mais n'est pas facilement mesurable. Certaines méthodes d'estimation (notamment Yin [de Cheveigné et Kawahara, 2002]) permettent de faire une distinction binaire entre passages voisés et passages non voisés, mais cette distinction nécessite un seuil, et ne permet pas de séparer chaque composante. Lors d'un voisement dit mixte, les turbulences causées par le flux d'air à travers la glotte se superposent à la composante harmonique. Arriver à estimer chacune de ces composantes permettrait non seulement de connaitre leur contribution énergétique au signal final, mais aussi de permettre des transformations complexes du signal vocal.

Nous avons vu au chapitre 2 que la méthode PAP à base de reconstruction itérative du bruit [Yegnanarayana *et al.*, 1998] souffre d'artefacts lorsque les harmoniques ne sont pas assez espacés (lorsque les fréquences fondamentales sont basses) : en effet, cette méthode utilise une fenêtre d'analyse fixe : il est donc nécéssaire de faire un compromis entre durée d'observation et résolution fréquentielle. Les méthodes adaptatives existantes (dont PSHF [Jackson et Shadle, 2001]) proposent des tailles de fenêtre d'analyse variables en fonction de la fréquence fondamentale et permettent donc de s'affranchir de ce compromis.

Le travail présenté dans ce chapitre s'est concentré sur le développement d'une méthode de décomposition périodique/apériodique à reconstruction itérative s'adaptant en fonction de la fréquence fondamentale, mais aussi tenant compte de la dispersion en fréquence fondamentale sur la fenêtre d'observation. La longueur de la fenêtre de pondération utilisée pour l'analyse est alors dynamiquement ajustée. Pour ce faire, une fonction de coût est définie, afin de minimiser la contribution des lobes secondaires des harmoniques dans le bruit estimé au final. Un intérêt particulier sera accordé à une reconstruction fidèle des composantes, pour non seulement reproduire le spectre d'amplitude, mais aussi le spectre de phase jouant un rôle tout aussi important dans la perception des signaux de parole [Alteris et Paliwal, 2003] : aux mesures classiques de rapport Signal sur Bruit comparé au niveau injecté, une mesure de l'erreur RMS entre composantes synthétisées et composantes estimées sera utilisée. Dans un deuxième temps, nous explorerons dans quelle mesure cette décomposition est indépendante des apériodicités structurelles.

4.1 Amélioration de l'algorithme PAP

4.1.1 Choix de la méthode

Le premier intérêt poussant à utiliser la méthode PAP [Yegnanarayana *et al.*, 1998] comme point de départ de cette étude est de chercher à repousser les limites de cet algorithme en lui injectant des résultats et des observations plus récentes sur la manière de mettre en œuvre la décomposition périodique-apériodique.

Lors de la revue de l'état de l'art, de nombreuses méthodes ont été présentées pour permettre une décomposition périodique/apériodique des signaux vocaux. Parmi elles, les méthodes paramétriques (modèles sinusoïdaux) bénéficient d'une attention particulière dans les domaines du codage et de la compression des signaux. Mais la force de ces modèles (leur paramétrisation) est aussi une faiblesse majeure, dans la mesure où ces méthodes sont très sensibles à la variation de fréquence fondamentale. D'autres méthodes tiennent compte de la fréquence comme paramètre du modèle, mais s'appliquent mal à des signaux non périodiques, comme c'est généralement le cas pour de la parole naturelle qui contient du jitter et du shimmer.

Le deuxième intérêt de l'algorithme présenté [Yegnanarayana *et al.*, 1998] est donc sa robustesse relative vis-à-vis de la précision en fréquence. Comme seule la moitié du spectre est attribuée à la partie apériodique, un nombre de bins de part et d'autre de la fréquence supposée contenir l'harmonique est supprimé de cette partie et estimé par itérations successives. Supposons donc

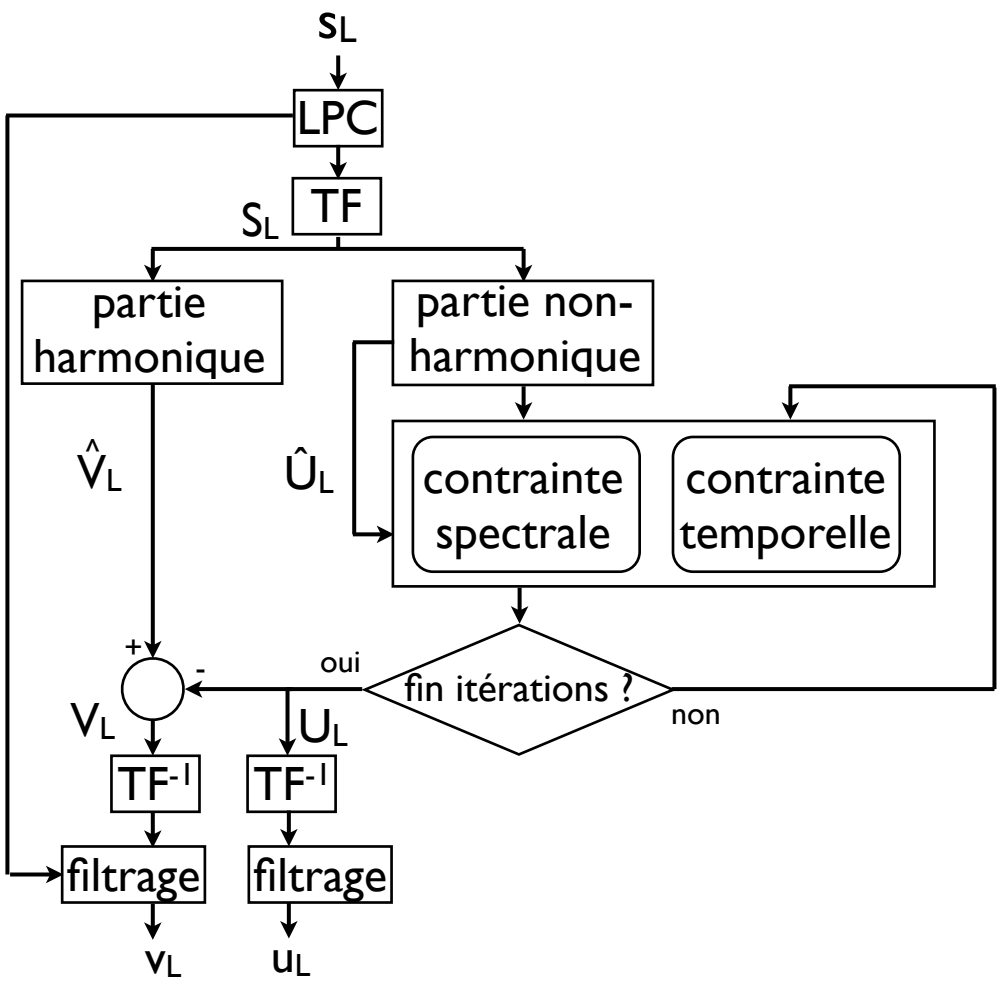


Figure 4.1 – Algorithme de la décomposition PAP (inspiré de [d'Alessandro *et al.*, 1998]).

des harmoniques séparés de 8 bins, et une précision sur la fréquence de 5%. Alors l'erreur de prédiction sur la position de l'harmonique sera de 2 bins (largeur de non affectation à la partie apériodique) au bout du sixième harmonique. En présence de Jitter constant, la variation de F_0 change peu ou pas en augmentant la taille de la fenêtre d'observation : une fenêtre d'observation plus longue sera donc bénéfique à la décomposition.

L'algorithme de la méthode PAP originale [Yegnanarayana *et al.*, 1998] est donné en figure 4.1. Une fenêtre d'observation du signal S_L de taille L est composée en deux parties : U_L et V_L (respectivement non voisée et voisée). Le problème est que la taille de cette fenêtre est fixée à L (512, 1024 ou 2048 préconisés). Nous allons donc agir en amont de la reconstruction, pour inclure une étape de choix de la taille de fenêtre L. Dans le reste de l'étude, la fréquence fondamentale est supposée connue. Pour les applications pratiques, on estime F_0 avec Yin [de Cheveigné et Kawahara, 2002].

4.1.2 Adaptation de la taille d'analyse à la fréquence fondamentale

Le principal désavantage d'une méthode à taille de fenêtre d'analyse constante est sa sensibilité à la fréquence fondamentale du signal. Si, pour des hautes fréquences fondamentales la décomposition est cohérente, lorsque les harmoniques se rapprochent dans le spectre (F_0 décroit) le risque de décomposition incomplète par le filtre harmonique apparaît. Cette décomposition incomplète se traduit aussi par une diminution artificielle du rapport signal/bruit décomposé par la méthode. Ceci explique en partie l'incapacité, lors de tests [d'Alessandro *et al.*, 1998], à arriver à estimer fidèlement des rapports son/bruit supérieurs à 20dB.

Ce principe a déjà été compris par Jackson [Jackson et Shadle, 2001], qui a adapté la méthode de décomposition périodique-apériodique de Serra [Serra et Julius Smith, 1990] pour tenir compte du F_0 courant.

Taille de la fenêtre idéale

Soit le spectre d'un signal périodique s de N échantillons et de période réduite $\frac{1}{\nu_0} = T_0 F_s$ pondéré par une fenêtre w et calculé sur un nombre N_{fft} de bins spectraux, alors les harmoniques sont espacés de :

$$n_0 = \nu_0 N_{fft}$$

Sur le spectre discret S de s, l'espacement des harmoniques est donc facteur de N_{fft}, mais la forme de la fenêtre d'analyse est dépendante principalement de la taille N de cette fenêtre. Ainsi, pour le cas d'une fenêtre de Von Hann ou de Hamming, la largeur du lobe principal est de $4\frac{N_{fft}}{N}$ en tenant compte de la complétion de zéros effectuée - il serait de $6\frac{N_{fft}}{N}$ pour une fenêtre de Blackman [Harris, 1978, Blackman et Tukey, 1959]. Cette largeur importante du lobe principal de cette fenêtre la rend de facto inadaptée à ce genre de décomposition pour garder la possibilité d'une durée d'observation minimale.

Or l'opération de fenêtrage par w revient à une convolution ($*$) spectrale $[S * W]$, et donc on retrouve la forme spectrale de la fenêtre de pondération autour de chaque harmonique.

Ainsi, si on cherche à détacher l'intégralité du lobe principal pour un signal purement harmonique lors du filtrage en peigne, il faut que les harmoniques soient espacés au minimum de la largeur du lobe principal de la fenêtre - dans ce cas extrême, aucun bin non nul ne serait présent dans la partie apériodique. On se retrouve donc avec l'inégalité suivante :

$$n_0 > 4\frac{N_{fft}}{N} \tag{4.1}$$

$$\nu_0 > \frac{4}{N} \tag{4.2}$$

$$N > \frac{4}{\nu_0} \tag{4.3}$$

$$\frac{N}{F_s} > 4T_0 \tag{4.4}$$

L'équation 4.4 obtenue nous indique le nombre minimum de périodes pour séparer deux lobes principaux. Dans le cas d'une fenêtre d'analyse de type Hann ou Hamming, ce nombre est de 4.

Cependant, pour reconstruire le bruit de manière convenable il convient aussi de conserver la moitié des bins lors du filtrage en peigne. Ce rapport de 50/50 entre bins périodiques et bins apériodiques est une condition nécessaire pour que la reconstruction itérative ne tende pas vers le spectre initial. Le nombre minimum de périodes à analyser passe donc à 8 afin de s'assurer que l'espace entre deux harmoniques soit supérieur à deux fois la largeur du lobe principal de la fenêtre de pondération.

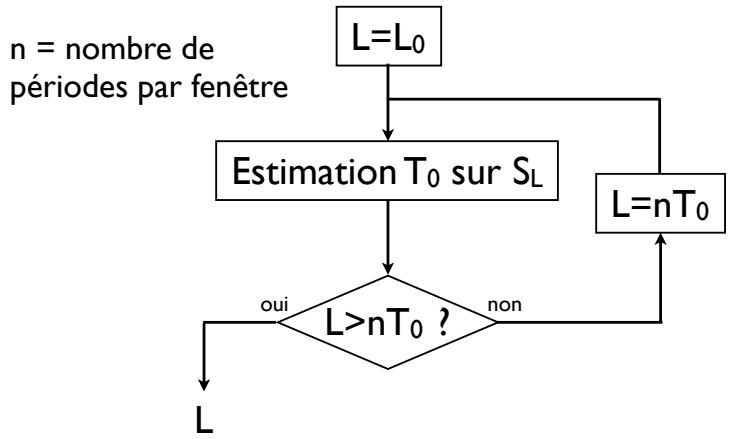


Figure 4.2 – Adaptation dynamique de la fenêtre d'observation en amont de la décomposition afin d'avoir toujours n périodes au minimum. S_L est le spectre de la fenêtre de longueur L observant le signal.

Application aux conditions réelles d'observation

Pour améliorer la décomposition, on peut augmenter le nombre de périodes à analyser. En montant à 12 périodes, il serait même possible de séparer le lobe secondaire induit par la fenêtre autour de chaque harmonique, diminuant d'autant plus l'erreur sur la décomposition. Cependant, augmenter le nombre de périodes à analyser sur un signal réel augmente aussi la dispersion de la fréquence fondamentale. Analyser 12 périodes d'un signal de parole à 200Hz, c'est observer le signal sur 60ms, bien au delà des 20ms utilisées généralement. Même le nombre minimal avancé au point précédent de 8 périodes représente un temps d'analyse très élevé. L'algorithme proposé effectue une évaluation dynamique de la fréquence fondamentale sur la période s'observation et vérifie l'équation 4.4 à chaque ajustement de la taille $L = NF_s$. Ce point de l'algorithme est illustré sur la figure 4.2. Afin de rester au plus près de la version originale de l'algorithme, une taille de fenêtre minimale est utilisée : en pratique sa valeur sera de 1024 échantillons, soit 64ms pour $F_s = 16kHz$. En principe donc, la correction de la taille de fenêtre se fera uniquement pour $F_0 > 125Hz$.

Une dispersion de la fréquence fondamentale se traduit par un étalement spectral des harmoniques, rendant caduque l'analyse précédente (l'étalement peut diminuer l'espace vide entre les harmoniques). Il est donc nécessaire d'adapter cette analyse théorique, ainsi que la forme de la fenêtre d'analyse, à la variation de fréquence fondamentale observée sur la largeur de la fenêtre d'analyse. Pour ce faire, une fonction de coût est utilisée et permettra de déterminer le meilleur compromis entre dispersion de F_0 et longueur d'observation.

4.1.3 Fonction de coût

Soit $F_0(n)$ la fréquence fondamentale instantanée du signal à la position n. Soit $\hat{F}_0$ la médiane de cette fréquence sur la durée d'observation L. Si la variance de F_0 sur l'intervalle observé est nulle, la suite est purement harmonique. Dans le cas contraire, on observe un étalement spectral sur le spectre discret autour des valeurs $k\hat{F}_0$ pour k entier.

Pour produire une séparation spectrale de qualité optimale, il convient de maximiser la taille de la fenêtre d'analyse tout en gardant un minimum de déviation de la fréquence fondamentale. On définit donc le coût comme étant la dispersion en fréquence de F_0 sur la durée d'observation par rapport à sa valeur médiane. On préfère la valeur médiane à la valeur moyenne comme

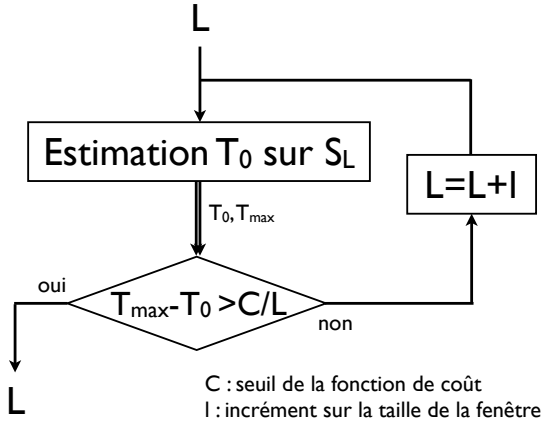


Figure 4.3 – Adaptation de la durée d'observation L en fonction de la dispersion en fréquence fondamentale du signal.

référence de fréquence fondamentale car elle est plus représentative de la position des harmoniques sur un signal réel. Soit $\hat{F}_0$ la valeur médiane de la distribution en fréquence et $F_0(n)$ la distribution en fréquence, ces deux vecteurs ont une taille N égale à la taille de la fenêtre.

$$C = max_n(|F_0(n) - \hat{F}_0|) \tag{4.5}$$

Le maximum C décrit par l'équation 4.5 est le décalage maximum en fréquence sur la distribution par rapport à la médiane. Ce décalage se retrouve amplifié par le rang des harmoniques. On cherche à minimiser la dispersion du pic harmonique sur le spectre et on choisit de manière arbitraire un décalage acceptable. La pratique montre qu'un décalage de $\frac{N_{fft}}{12N}$ (où N et N_{fft} sont respectivement la durée d'observation et la taille de la transformée de Fourier discrète) est acceptable dans la mesure où il garantit une décomposition au bin près jusqu'au 12ème harmonique.

Ce critère est ensuite adapté à la sélection de la taille de la fenêtre d'analyse dont le principe est présenté sur la figure 4.3. Une taille de fenêtre maximum (T_{max}) est fixée dans l'algorithme.

4.1.4 Algorithme PAP-A

Le code complet de la décomposition périodique / apériodique adaptative est développé dans l'algorithme 1. On considère que les scalaires N_{MAX}, N_{MIN} (taille minimum et maximum de la fenêtre), l (incrément), N_{period} (nombre de périodes minimum), F_S (fréquence d'échantillonnage) ainsi que les vecteurs S_{IN} (signal) et F_{IN} (fréquence fondamentale instantanée) sont connus. On récupère alors les vecteurs V_{out} et N_{out} respectivement des parties voisées et non-voisées. La procédure PAP réalise les étapes de la figure 4.1 alors que la procédure $[F_0, F_{max} \leftarrow EstimF0(s_L)$ est une procédure qui retourne la valeur médiane F_0 et la valeur maximum F_{max} de la fréquence fondamentale pour le signal s sur la durée d'observation L.

Cet algorithme est nommé PAP-A, pour Periodic - APeriodic Adaptive décomposition. Deux variantes sont proposées : PAP-A avec et sans utilisation de la fonction de coût.

4.2 Application à des signaux de tests

Il convient d'étudier la qualité de la décomposition opérée par PAP-A sur un certain jeu de signaux de tests à l'instar des évaluations présentées dans d'autres travaux

Algorithm 1 Décomposition PAP adaptative avec fonction de coût

```
soit N_MAX , N_pas , N_period , S_IN , F_IN , F_S
soit K , L
i ← 0
tant que i < longueur(s)
    L = N_MIN
    s_L ← s(i à i + L)
    [F_0, F_max] ← EstimF0(s_L)
    tant que N_period / F_0 < L % adaptation de la fenêtre à F_0
        L = L + N_pas
        s_L ← s(i à i + L)
        [F_0, F_max] ← EstimF0(s_L)
    fin tant que
    tant que F_max − F_0 < C/L % Augmentation de la durée d'observation
        L = L + N_pas
        s_L ← s(i à i + L)
        [F_0, F_max] ← EstimF0(s_L)
    fin tant que
     [n,_L v_L] = PAP(s_L, F_0) % Décomposition avec reconstruction itérative
    ajouter v_L à V_out(i à i + N)
    ajouter n_L à N_out(i à i + N)
    i ← i + N_pas
fin tant que
retourner V_out et N_out
```

Table 4.1 – Ensemble des paramètres utilisés pour la base de données de signaux synthétiques pour le test des cinq méthodes de décomposition.

Paramètre	Valeurs	Unité
Voyelle	/a/, /i/, /u/	
F_0	$[90, 150, 300]$ ou 90 à 300 en 0.75 s.	Hz
RSB	$[0, 5, 10, 20, 30]$	dB
jitter	$[0, 1, 3, 5]$	%
shimmer	$[0, 6, 12, 19]$ ou encore $[0; 0,5; 1; 1.5]$	% ou dB
largeur du bruit	$[30, 95]$	% de T_0

[d'Alessandro *et al.*, 1998, Jackson et Shadle, 2001]. Dans tous les cas, la configuration glottique utilisée pour générer les signaux sera considérée constante, à des valeurs de paramètres fréquemment mesurées ($O_q = 0.5$ et $\alpha_m = 0.7$). La fréquence fondamentale sera d'une part fixe, et d'autre part variable afin de tester les algorithmes dans des conditions stationnaires ou non : deux jeux de résultats seront donc obtenus et présentés séparément. Tous les signaux générés auront une longueur de 100 périodes à FS=16kHz, pour chacune des 3 voyelles cardinales du triangle vocalique : /a/, /i/ et /u/.

Le bruit sera pulsé autour du GCI, de 30 ou 90% de la période du signal. Son énergie par rapport au signal (RSB) sera variable de 0 à 30dB ; le bruit est ajouté avant filtrage autorégressif. Des perturbations structurelles seront ajoutées au signal : le jitter sera variable de 0 à 5% de la période (écart type sur la durée du signal), tout comme le shimmer qui sera variable de 0 à 19% de l'amplitude glottique (toujours en écart type). L'ensemble des paramètres utilisés pour la base de synthèse sont présentés dans la table 4.1.

Afin de tester en détail les améliorations portées à l'algorithme original PAP et en comparaison de la méthode PSHF [Jackson et Shadle, 2001], 5 méthodes seront testées :

- L'algorithme PAP original [Yegnanarayana *et al.*, 1998] pour une taille de fenêtre fixe de 1024 échantillons. (fréquence fondamentale estimée avec Yin)
- L"algorithme PAP original où la fréquence instantanée est parfaitement connue.
- L'algorithme PAP-A proposé ici, mais sans utilisation de la fonction de coût ($C = 0$). (fréquence fondamentale estimée avec Yin)
- L'algorithme PAP-A proposé ici, conforme à l'algorithme 1. (fréquence fondamentale estimée avec Yin)
- L'algorithme PAP-A proposé ici, dont la fréquence fondamentale est parfaitement connue.

Pour toute l'étude, les paramètres utilisés pour les algorithmes PAP-A : 8 périodes minimum, 1024 points pour la taille de fenêtre minimum, 8192 points la taille de fenêtre maximum, déviation maximum de $\frac{N}{8F_S}$. Comme précisé, la fréquence fondamentale est estimée avec Yin [de Cheveigné et Kawahara, 2002] et est donc considérée imparfaitement connue sauf indication contraire.

4.2.1 Critère d'évaluation

Le premier critère d'évaluation qui vient à l'idée est de comparer la puissance du bruit estimée par chaque méthode à la puissance de bruit injectée dans le signal synthétique. C'est notamment la mesure proposée précédemment [d'Alessandro *et al.*, 1998]. Cette mesure sous entend que la méthode n'estime pas la partie apériodique du signal, mais un signal aléatoire de variance similaire à celle de la partie apériodique du signal source. Avoir un RSB similaire est une

condition nécessaire mais non suffisante à la décomposition. Cependant, pour faire le parallèle avec les tests précédents sur PAP, de tels résultats seront donnés dans un premier temps.

Dans un deuxième temps, et selon les travaux de Jackson, les mesures d'erreur de prédiction sont utilisées. Elle permettent de quantifier la capacité de la décomposition à estimer une réalisation d'un processus aléatoire (donc un signal déterminé). Les résultats seront évalués en fonction de l'erreur RMS en dB (*Root Mean Square*) e_{dB} entre le bruit estimé $\tilde{b}$ et le bruit injecté b normalisé par l'énergie moyenne du bruit injecté telle que définie à l'équation 4.6.

$$e_{dB} = 10log10\left(\frac{\sum(b(i) - \tilde{b}(i))^2}{\sum b(j)^2}\right) \tag{4.6}$$

Ce critère de puissance permet de quantifier l'erreur par rapport au bruit injecté : une erreur de 0dB signifie donc que la puissance de l'erreur d'estimation est de l'ordre de la puissance du signal (une décomposition peu ressemblante), plus l'erreur est basse et meilleure est la décomposition. Par propriété du théorème de Parceval, minimiser e c'est aussi minimiser l'erreur spectrale de l'estimation du bruit.

Pour éviter les effets de bords et initialisation des algorithmes, notamment sur l'estimation de la fréquence fondamentale, l'erreur est calculée sur le centre des signaux, de 20% à 80% de leur longueur.

4.2.2 Résultats des signaux de tests

Les résultats sont donnés en comparant la puissance du bruit estimé par rapport à la puissance du bruit généré. Afin de visualiser convenablement l'influence des perturbations structurelles sur la qualité de l'estimation du bruit, les résultats sont divisés en 4 groupes :

1. Un premier groupe ne comportant pas de perturbation structurelle.
2. Un deuxième groupe comportant de légères perturbations structurelles : 0.5dB de shimmer et 1% de jitter.
3. Un troisième groupe comportant de moyennes perturbations structurelles : 1dB de shimmer et 3% de jitter.
4. Un quatrième groupe comportant de fortes perturbations structurelles : 1.5dB de shimmer et 5% de jitter.

Les résultats de comparaison entre SNR injecté et SNR mesuré de ces signaux synthétiques sont présentés à la figure 4.5 et 4.7 respectivement pour les tests à fréquence fondamentale constante et fréquence fondamentale variable. Les résultats sur l'érreur d'estimation de la partie voisée sont donnés sur les figure 4.4 et 4.6. Sur chaque figure, quatre graphiques y sont présentés, un pour chaque groupe de résultats donné ci-dessus. Les résultats sont moyennés pour chaque voyelle, pourcentage de périodes bruitées et fréquence fondamentale. Ce regroupement est motivé par les résultats obtenus précédemment sur l'algorithme PAP [d'Alessandro *et al.*, 1998].

4.2.3 Discussion

Sans apériodicités structurelles

Ce qui frappe de prime abord sur le premier graphique (en haut à gauche des figures 4.5 et 4.4), c'est la linéarité des méthodes PAP-A au niveau de la constance des résultats par rapport à PAP. Comme prévu par les résultats obtenus dans les expériences précédentes

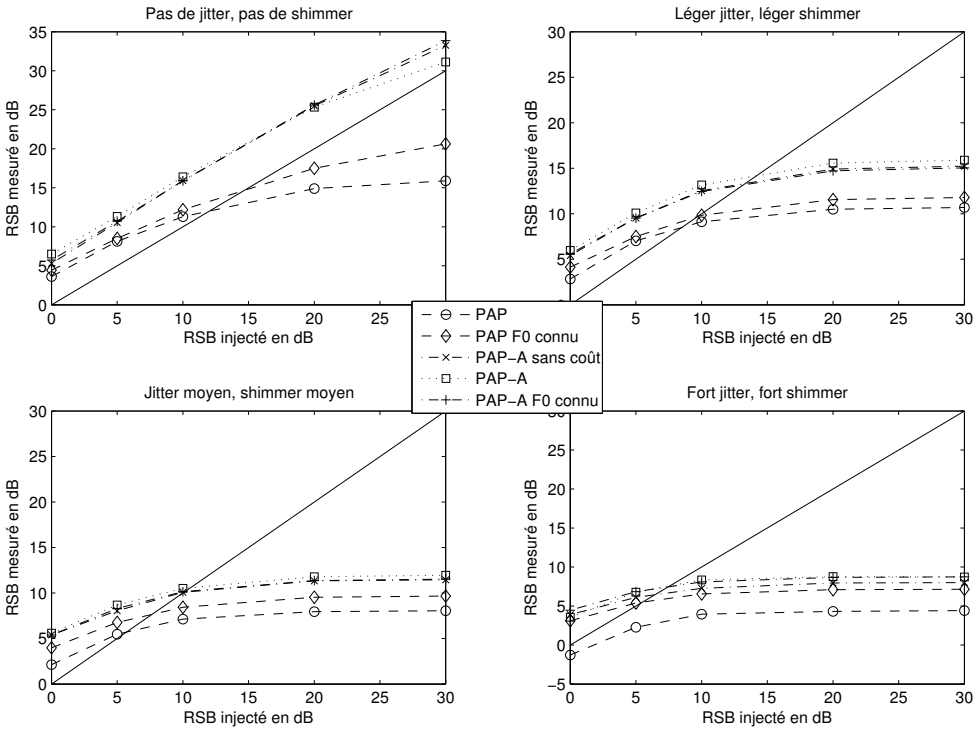


Figure 4.4 – Résultats d'estimation du RSB pour les 5 méthodes sur des signaux de tests à fréquence fondamentale fixe, en fonction de la quantité de jitter, de shimmer.

[d'Alessandro *et al.*, 1998], il y a un léger déficit entre le SNR mesuré et le SNR injecté sur la figure 4.4 : ce déficit est constant à 6dB pour PAP-A pour toutes les valeurs de SNR. Concernant l'erreur d'estimation (figure 4.5) pour une reconstruction adaptative les résultats sont meilleurs et constants (à l'image du SNR), à une valeur constante de -3dB soit une erreur de reconstruction importante, mais plus proche de l'original que pour PAP. Les courbes de PAP-A sont toutes superposées, alors que PAP ayant F_0 connu donne de meilleurs résultats que PAP utilisant Yin : ceci montre déjà une sensibilité plus importante à la précision sur F_0 de l'algorithme PAP que de PAP-A. Les résultats de PAP-A sont aussi comparables à ceux obtenus avec l'algorithme PSHF [Jackson et Shadle, 2001] non testé ici.

Concernant l'analyse de signaux à fréquence fondamentale variable sur les figures 4.7 et 4.6 les résultats sont moins marqués. La méthode PAP ayant connaissance exacte de la fréquence fondamentale propose des résultats comparables à PAP-A dont l'erreur d'estimation reste acceptable jusqu'à un SNR de 5 ou 10dB (i.e. : erreur inférieure à 0dB). Dans ce cas de signaux à fréquence fondamentale variable, l'erreur donne des résultats plus intéressants que la comparaison des SNR qui est uniquement fidèle autour de -10dB probablement du fait de certains artefacts (l'erreur étant plus élevée pour que les valeurs plus faibles de SNR).

Avec un léger Shimmer ou un léger Jitter

Les résultats de ces analyses sont présentés sur les graphiques en haut à droite des figures 4.5, 4.4, 4.7 et 4.6. Pour ces faibles apériodicités on retrouve la même hiérarchie qu'en l'absence de jitter et de shimmer, mais les performances des algorithmes PAP-A sont moins bonnes pour des fortes valeurs de SNR (au dessus de 10dB). Somme toute, on trouve une puissance d'erreur

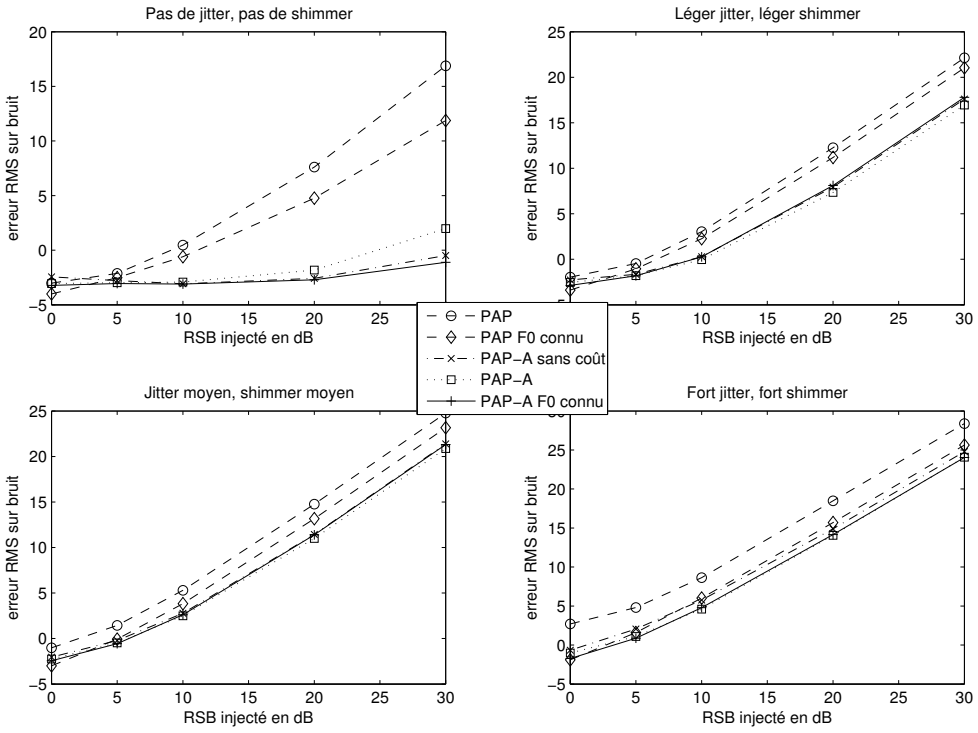


Figure 4.5 – Erreur d'estimation de la partie apériodique pour 5 méthodes sur des signaux de tests à fréquence fondamentale fixe, en fonction de la quantité de jitter, de shimmer.

toujours inférieure à la puissance de bruit injectée jusqu'à SNR=10dB ce qui montre une capacité à estimer plus efficacement la partie bruitée.

Alors que l'utilisation de la fonction de coût ne donnait pas de meilleur résultat pour F_0 fixe, on trouve une légère tendance d'amélioration pour de faibles perturbations structurelles.

Pour des apériodicités plus importantes

Les résultats de l'erreur d'estimation pour les valeurs moyennes et fortes de jitter et shimmer sont données sur les lignes inférieures des figures 4.4 et 4.6 sur les graphiques du bas. Les valeurs mesurées du RSB sont données sur les graphiques du bas des figures 4.5 et 4.7. Ces derniers ne sont pas très pertinents, et l'analyse de l'erreur d'estimation est plus appropriée.

Dans le cas où la fréquence fondamentale est fixe, les méthodes ont des résultats qui se dégradent progressivement. Cependant, la méthode PAP a une erreur évoluant bien plus vite avec la quantité d'apériodicités. Pour un fort jitter et un fort shimmer, la puissance de l'erreur est vite supérieure à la puissance de bruit injectée sur le signal. Dans le cas où la fréquence fondamentale est variable, l'écart entre les trois variantes de PAP-A se marque progressivement pour donner l'avantage à la connaissance parfaite de la fréquence fondamentale.

4.2.4 Conclusion sur l'analyse de signaux synthétiques

Les résultats de l'analyse de signaux synthétiques donnent un certain nombre d'informations sur l'approche qui a été adoptée pour améliorer la décomposition PAP.

Toutes les variantes proposées donnent des résultats équivalents ou meilleurs que l'algorithme

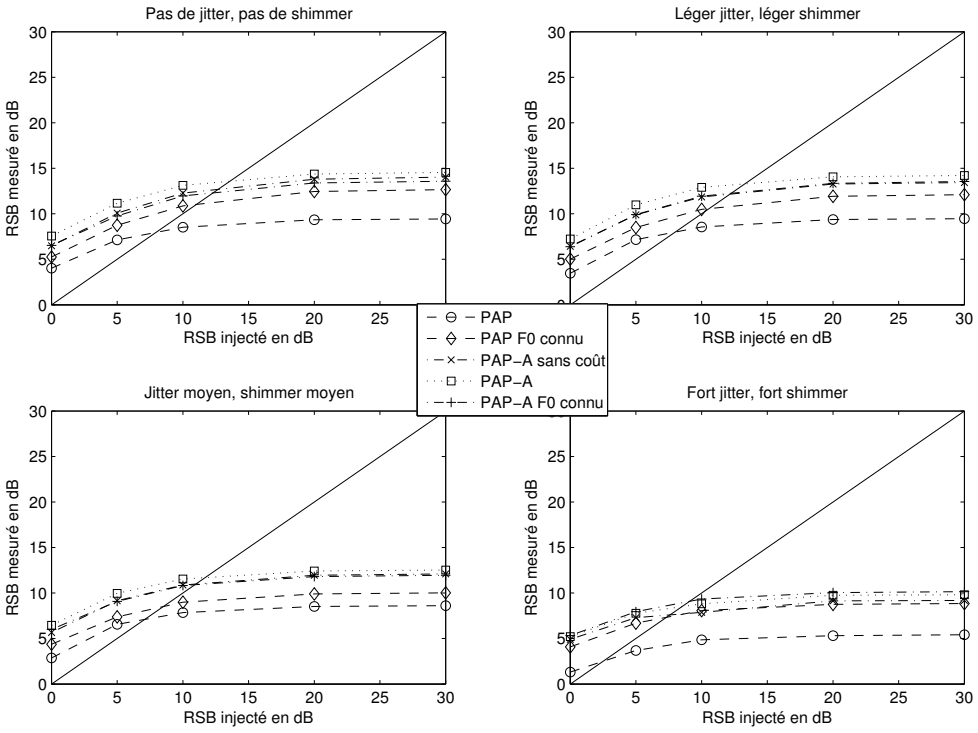


Figure 4.6 – Résultats d'estimation du RSB pour les 5 méthodes sur des signaux de tests à fréquence fondamentale variable, en fonction de la quantité de jitter, de shimmer.

original, en particulier dans des conditions où les apériodictés structurelles sont très présentes. La connaissance parfaite de la fréquence fondamentale permet principalement d'améliorer la décomposition pour l'algorithme PAP jusqu'à des niveaux d'erreur comparables à la méthode adaptative.

La contribution majeure est sans conteste l'utilisation du fenêtrage adaptatif, la fonction de coût intervenant principalement dans le cas d'un faible jitter : où la variation de F_0 (alors inférieure à la limite C) permet d'augmenter la taille de la fenêtre et de produire des décompositions de meilleure qualité.

On remarque de plus que la méthode adaptative PAP-A est considérablement moins sensible à l'erreur d'estimation de la fréquence fondamentale que la méthode PAP originale. L'utilisation d'une fenêtre adaptée à la fréquence fondamentale moyenne permet de maintenir un écart constant entre les harmoniques et de limiter un effet de superposition intervenant lors d'une modulation de F_0.

4.3 Application à des signaux réels

La validation d'une telle méthode par une analyse systématique de signaux réels se révèle problématique dans le sens où il n'existe pas de référence pour l'évaluation de la quantité de bruit dans un signal acoustique complexe. On peut cependant donner quelques exemples probants, en particulier sur des points spécifiques de la décomposition : d'une part un signal stationnaire et d'autre part un signal au voisement mixte.

Par la suite, une écoute comparative sera proposée au lecteur pour juger des avantages per-

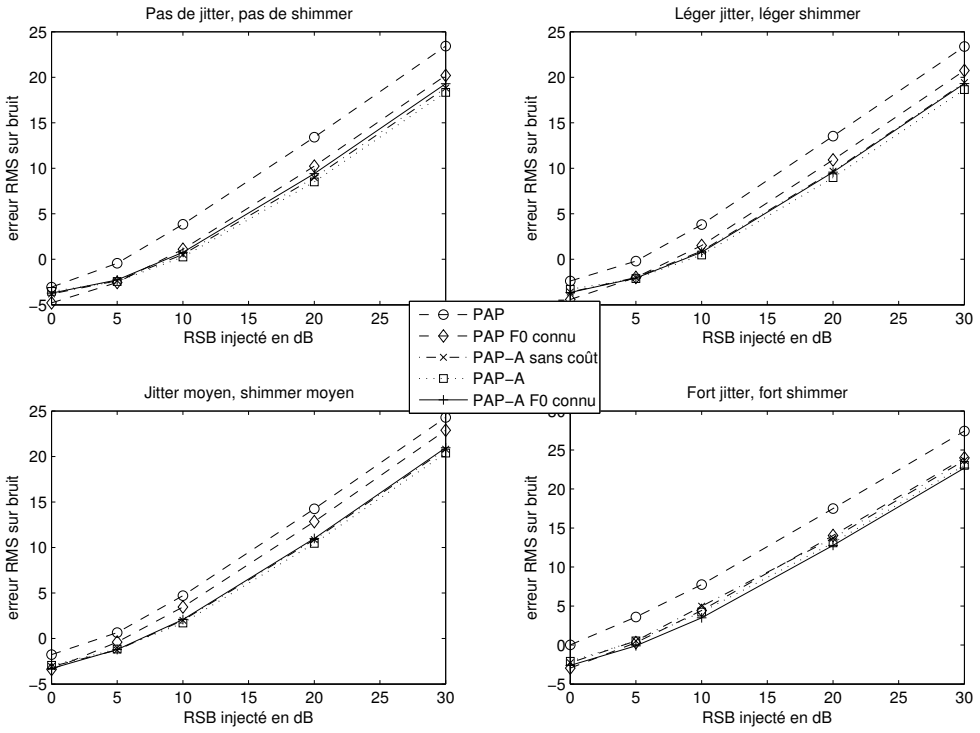


Figure 4.7 – Erreur d'estimation de la partie apériodique pour 5 méthodes sur des signaux de tests à fréquence fondamentale variable, en fonction de la quantité de jitter, de shimmer.

ceptifs de l'utilisation d'une fenêtre adaptative.

4.3.1 Voyelles tenues

Le premier exemple de décomposition est celui d'une voyelle tenue. Sur la figure 4.8 sont présentés le spectre et la forme d'onde du signal original, de la partie périodique et de la partie apériodique de deux voyelles /a/ prononcées par une locutrice féminine. Les signaux sont suffisamment stationnaires pour y visualiser clairement les harmoniques. La ligne du haut montre l'analyse d'un voisement doux, presque soufflé tandis que la ligne du bas montre l'analyse d'un voisement très clair. Ces signaux sont disponibles sur le site web [1] : ce sont les fichiers PAP_Voyelle1_XXX.wav, où XXX désigne 'PERI' pour la partie périodique, 'ORIG' pour le signal original et 'NOIS' pour la partie apériodique.

Dans le cas du premier signal, la voyelle 'soufflée', le bruit estimé est de l'ordre de grandeur des composantes du signal entre 2000 et 4000Hz. Le spectre du signal montre des harmoniques noyés dans le bruit pour cette plage de fréquence alors que le spectre de la partie périodique - en vert - permet clairement de distinguer des harmoniques. Sur tout le spectre, le rapport signal/bruit est d'environ +15 à +20dB.

Dans le cas du deuxième signal, la voyelle prononcée de manière claire, le bruit estimé est nettement en dessous des harmoniques, on constate un RSB d'environ +30 à +40dB. De 0 à 4000Hz, aucun harmonique n'est noyé dans le bruit, la difficulté de décomposition n'est donc pas très importante.

1. http ://nicolas.sturmel.com/PHD/

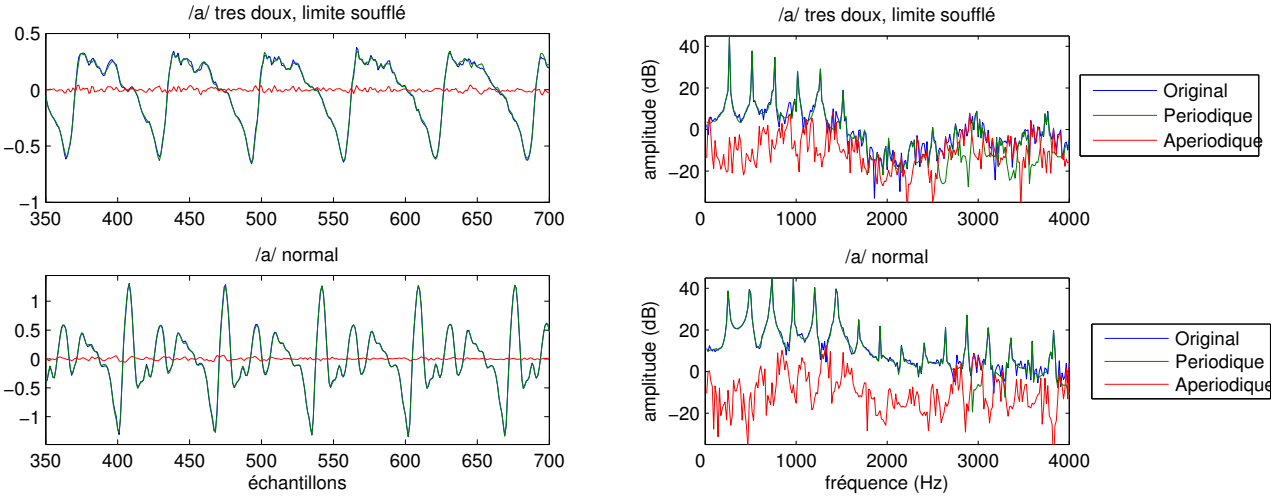


Figure 4.8 – Décomposition PAP-A d'une voyelle /a/ produite par une locutrice féminine. $F_0 \approx 250Hz$. En haut, phonation très douce, O_q élevé, bruit important. En bas, phonation modale et claire. Sur la gauche, les formes d'onde ; sur la droite les spectres du signal, de la partie périodique et de la partie apériodique.

4.3.2 Fricative voisée

L'analyse d'une fricative voisée est un bon déterminant de la qualité d'une décomposition périodique-apériodique. Ces sons comportent naturellement un début de voisement et une grande quantité de bruit de phonation dû à l'articulation du locuteur. L'exemple du son [ʒ ə] prononcé dans 'je' est présenté sur la figure 4.9. On y visualise à la fois les formes d'onde et les spectres pour le son original ainsi que pour chacune des deux parties : périodique et apériodique obtenues par la méthode PAP-A. Dans ce cas il n'est pas pertinent de visualiser le spectrogramme, les signaux n'étant pas suffisamment longs pour que la visualisation soit pertinente. Ces signaux sont disponibles sur le site web[2], ce sont les fichiers PAP_Fricative1_XXX.wav, où XXX désigne 'PERI' pour la partie périodique, 'ORIG' pour la signal original et 'NOIS' pour la partie apériodique.

La décomposition périodique/apériodique est produite avec qualité. La partie apériodique (milieu de la figure 4.9 est vidée de quasiment toute périodicité mais contient la quasi-totalité du bruit présent dans le signal, y compris les variations basses fréquences. On notera cependant une périodicité résiduelle autour de 1.2 secondes, ce qui n'est pas forcément signe d'une mauvaise décomposition dans la mesure où un bruit peut apparaitre plus important à l'instant de fermeture glottique, et donc être pulsé avec régularité. La partie voisée (bas de la figure 4.9) présente un spectre plus harmonique que le signal original (haut de la figure 4.9). Cet aspect de spectre harmonique semble aussi présent sur la composante apériodique, mais principalement par une incapacité de la méthode à reconstruire précisément le bruit sur les bins harmoniques : on visualise donc des creux au niveau des harmoniques, en particulier dans les hautes fréquences. On remarque tout de même quelques artefacts notamment au début de la phonation, où l'effet de la fenêtre marque moins bien l'attaque sur les parties périodiques et apériodiques que sur le signal original.

2. http ://nicolas.sturmel.com/PHD/

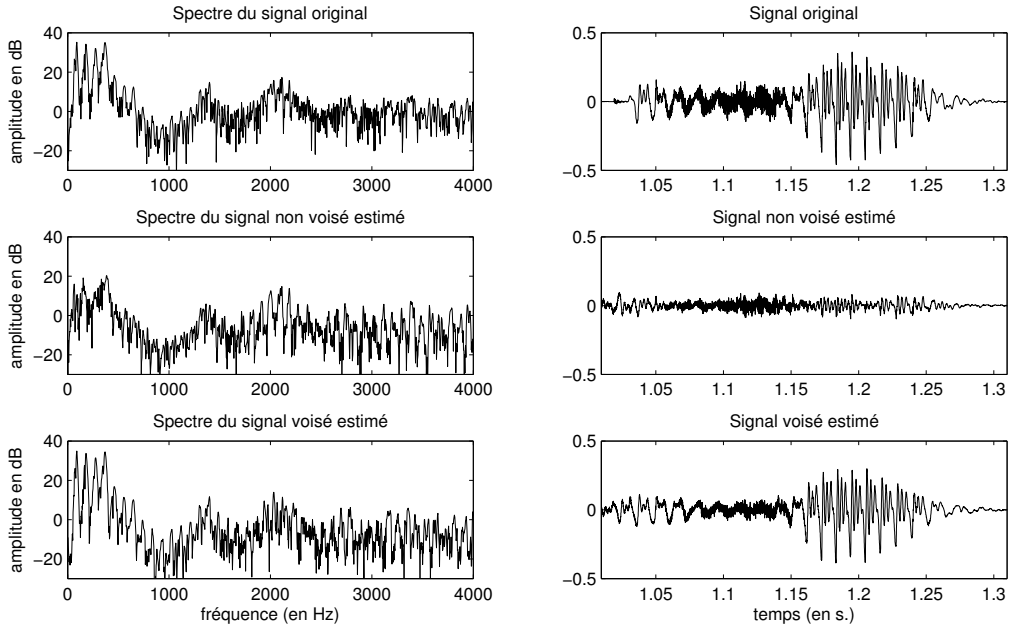


Figure 4.9 – Exemple de décomposition sur le son [ʒə] : représentation des signaux temporels et de leurs spectres.

4.3.3 Comparaisons perceptives et de spectrogrammes

Sur le site web [3] sont proposés 15 fichiers sonores composés de la version originale, les versions voisé et non voisé estimées par PAP, les versions voisé et non voisé estimées par PAP-A pour trois phrases différentes prononcées par deux locuteurs. Ces phrases sont tirées de la base de données utilisée pour la validation de la détection des LoMA au chapitre 3. Les spectrogrammes en bande étroite (fenêtre de hann de 512 échantillons, déplacement de 32 échantillons) de décomposition de ces fichiers sont donnés sur les figures 4.10, 4.11 et 4.12 pour les fichiers C202, C203 et M202 respectivement.

Les deux premières phrases (C202 et C203) permettent de tester les différences au niveau de l'algorithme adaptatif en PAP et PAP-A. En effet, la voix du locuteur est inférieure à 125Hz, fréquence limite pour avoir 8 périodes sur une fenêtre de 1024 points à une fréquence d'échantillonnage de 16kHz. Les différences sont assez nettes entre les fichiers `N-PAP` et `N-PAPA` donnant la partie voisée respectivement estimée par PAP et PAP-A. Sur les parties voisées (`V-PAP` et `V-PAPA`) on entend une cohérence plus marquée pour PAP-A, avec un effet "phaser" moins prononcé.

Pour la troisième phrase, prononcée par un locuteur féminin, la taille de fenêtre minimum de 1024 points assure quasiment toujours les 8 périodes présentent sur la durée d'observation. On entend donc principalement des différences au niveau de la fonction de coût et de la taille de fenêtre ajustée en fonction de la dispersion en fréquence. Les différences sont notamment présentes sur les transitoires comme sur le mot "commun" entre les secondes 5 et 6 de l'extrait. PAP-A présente aussi moins d'artefacts au niveau de la reconstruction.

3. http ://www.sturmel.com/nicolas/PHD

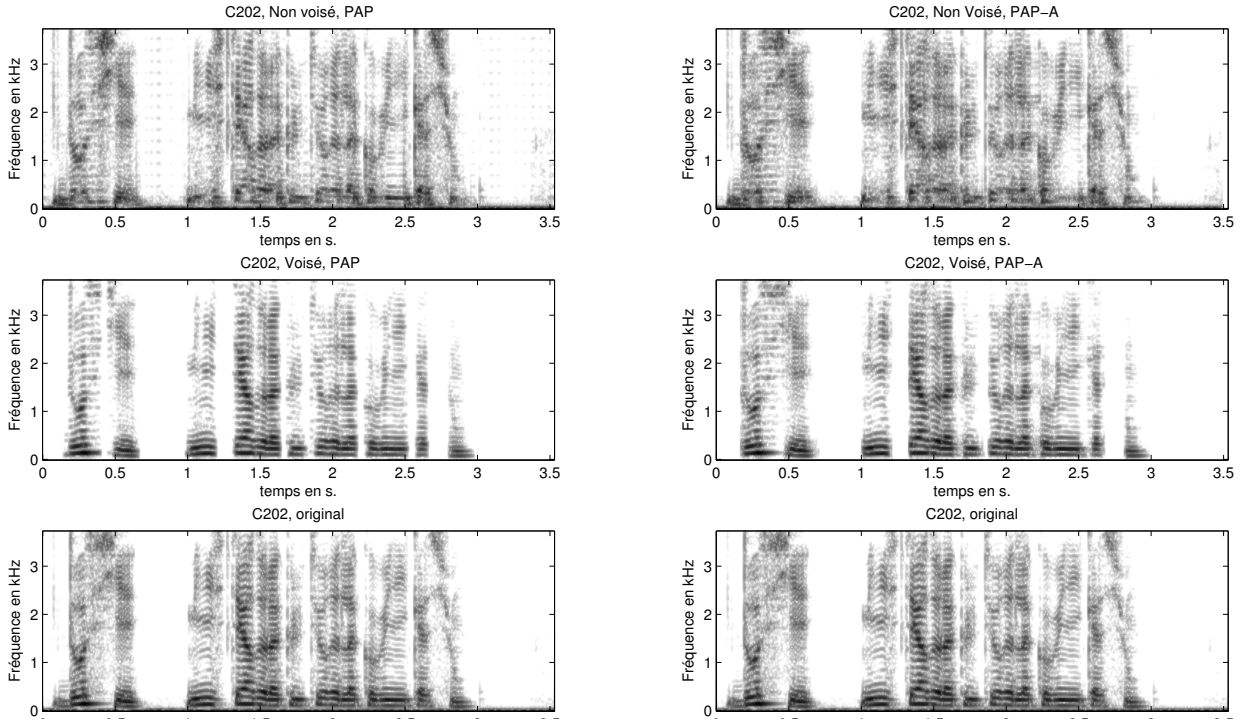


Figure 4.10 – Spectrogrammes en bande étroite de décomposition périodique/apériodique par PAP et PAP-A du fichier C202.

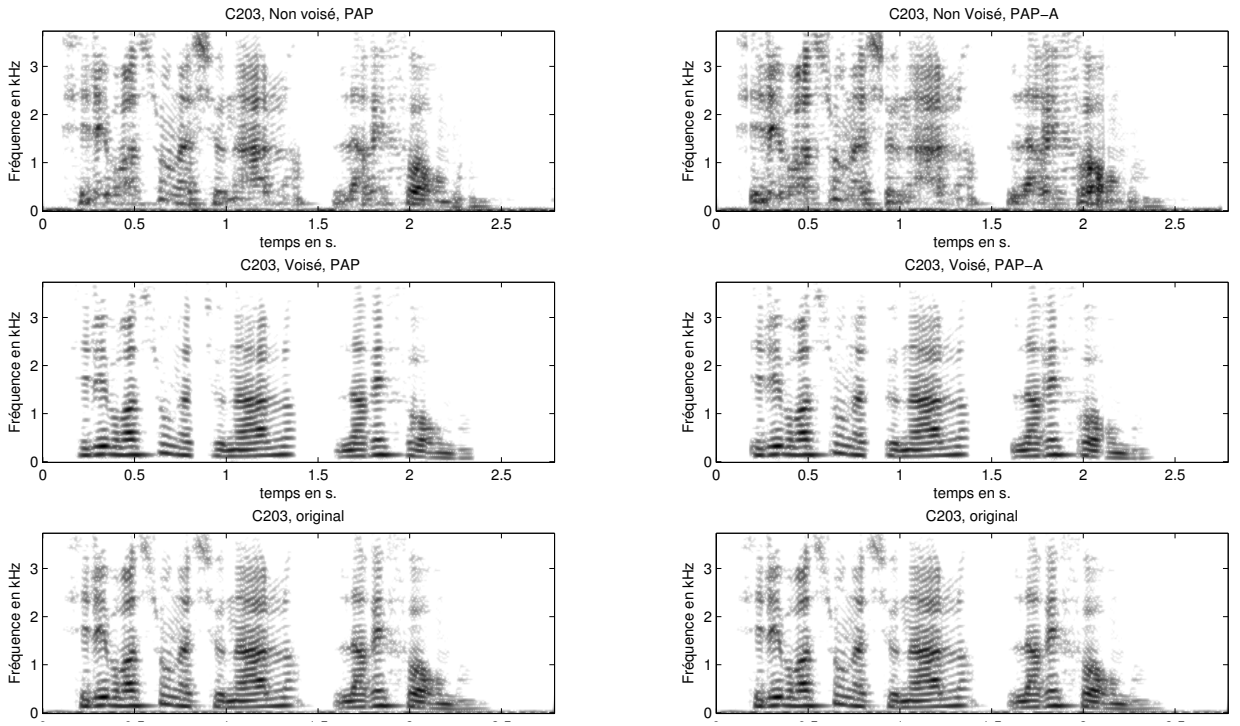


Figure 4.11 – Spectrogrammes en bande étroite de décomposition périodique/apériodique par PAP et PAP-A du fichier C203.

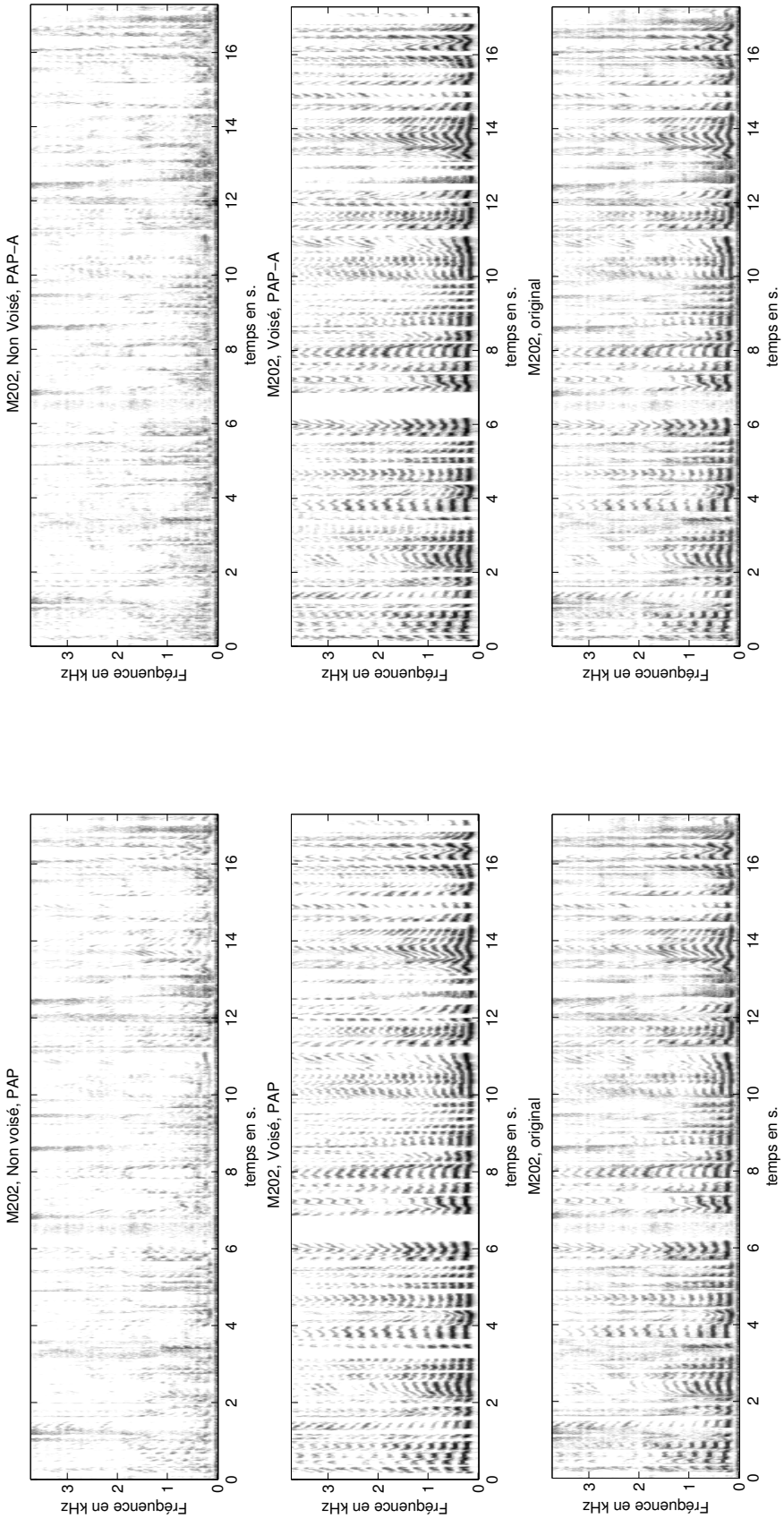


Figure 4.12 – Spectrogrammes en bande étroite de décomposition périodique/apériodique par PAP et PAP-A du fichier M202.

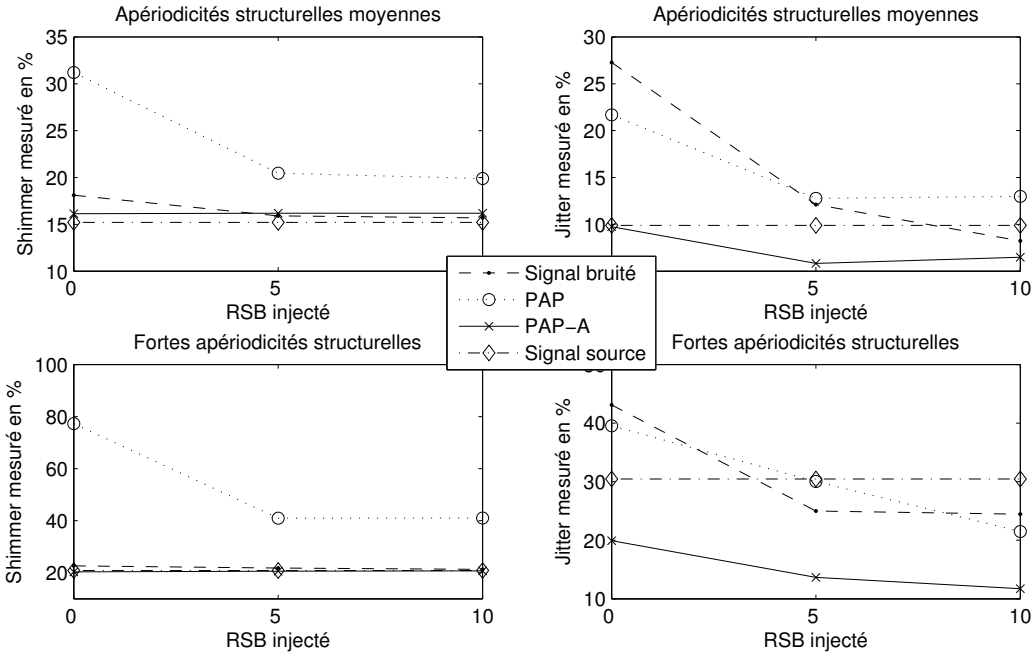


Figure 4.13 – Estimation du jitter et du shimmer avant et après décomposition périodique apériodique par méthode PAP (rond pointillés) et PAP-A (croix trait continu), l'estimation sur la source non bruitée est donnée par le trait mixte et l'estimation sur le signal bruité par le trait discontinu.

4.4 Impact de la décomposition sur l'estimation des LoMA

Dans le chapitre précédent (chapitre 3) nous avons montré que le bruit avait un impact sur la capacité de la méthode LoMA à fournir des données exploitables pour estimer jitter et shimmer. En reprenant les signaux générés selon les paramètres du tableau 4.1 pour une fréquence fondamentale fixe de 90Hz, le jitter et le shimmer ont été estimés par les LoMA sur les parties voisées extraites de ces signaux. Ces résultats sont présentés sur la figure 4.13, ou sont présentés jitter et shimmer estimés sur le signal bruité, sur le signal périodique obtenu par PAP-A et PAP et sur le signal original non bruité et non filtré par le conduit vocal. Les résultats sont donnés pour trois niveaux de RSB : 0, 5 et 10dB. Un bruit moins présent n'a qu'un impact faible sur l'estimation du jitter et du shimmer.

On observe sur la figure 4.13 une tendance à estimer un shimmer plus fort que l'original pour PAP alors que PAP-A produit un niveau estimé conforme à celui estimé à la fois sur le signal de source et sur le signal bruité. Cependant, le jitter est lui estimé par PAP bien plus proche de la valeur trouvée sur le signal de source. Comme montré au chapitre 3, la mesure du jitter est plus fiable.

4.5 Conclusion

Dans ce chapitre, deux améliorations pour la méthode de décomposition des parties périodiques et apériodiques des signaux vocaux ont été présentées. En tenant compte des travaux précédents [Yegnanarayana *et al.*, 1998] et [Jackson et Shadle, 2001], un algorithme de reconstruction itérative de la partie apériodique a été modifié pour adapter automatiquement la taille de la fenêtre d'analyse à la fréquence fondamentale courante. Afin de détecter les parties stationnaires du signal, nous avons proposé une fonction de coût estimant la dispersion du contour de fréquence fondamentale.

Les deux méthodes dérivées des modifications successives apportées à l'algorithme original ont été testées sur une batterie de signaux synthétiques, à l'instar des protocoles de test précédents [d'Alessandro *et al.*, 1998, Jackson et Shadle, 2001]. Afin d'observer la sensibilité de la décomposition à la précision de l'estimation *a priori* de la fréquence fondamentale, certaines méthodes ont été adaptées pour tenir compte de la valeur réelle de F_0, parfaitement connue pour l'analyse des signaux synthétiques. Ces deux méthodes sont :

1. PAP-A : décomposition PAP adaptant la fenêtre d'observation à la fréquence fondamentale courante et utilisant la fonction de coût pour augmenter la fenêtre d'observation si possible.
2. PAP-A n'utilisant pas la fonction de coût.

Un protocole de test sur signaux synthétiques a comparé les résultats des méthodes PAP avec ou sans connaissance parfaite de F_0 avec les méthode PAP-A. Les résultats montrent que l'amélioration la plus importante se situe dans l'adaptation de la taille de la fenêtre à la fréquence courante. La fonction de coût a une utilité marginale dans le cas d'un jitter assez faible pour satisfaire la condition de limitation de dispersion en fréquence. Ces résultats montrent aussi que PAP-A est nettement moins sensible aux erreur d'estimations de la fréquence fondamentale que PAP.

Des exemples de décompositions sur des signaux réels ont été donnés et discutés. La méthode PAP-A se comporte bien à la fois sur des signaux stationnaires, mais aussi sur de la parole au voisement mixte, comme c'est le cas pour les consonnes voisées. Des exemples sonores ont été commentés et permettent au lecteur de se faire une idée sur la qualité subjective de la décomposition, Il serait toutefois nécessaire de mettre en œuvre une expérience perceptive afin de quantifier la qualité de cette décomposition sur des signaux réels de manière plus poussée.

Le choix de l'estimateur de F_0 est une étape déterminante dans la qualité de la décomposition, les résultats obtenus dans ce chapitre montrent toutefois que l'estimateur Yin [de Cheveigné et Kawahara, 2002] produit une estimation de F_0 à la qualité suffisante. Vis à vis de la conservation des apériodictés structurelles, une courte étude a montré que PAP-A donnait des résultats légèrement meilleurs que PAP, notamment du point de vue du shimmer.

Cet algorithme produit cependant des résultats perfectibles et il serait notamment judicieux de chercher à résoudre le problème des fenêtres d'observation où la variation de fréquence fondamentale est trop importante pour permettre l'apparition d'harmoniques sur le spectre.

Résumé

Problème

Une dimension importante de la qualité vocale est le rapport entre la partie voisée et la partie non voisée du signal. La première présente une structure harmonique alors que la seconde a généralement une structure aléatoire qui se traduit par une impression de bruit à l'écoute. Les travaux précédents menés [Yegnanarayana *et al.*, 1998, Jackson et Shadle, 2001] se sont penchés sur cette décomposition des parties harmoniques et bruitées des signaux. Le choix d'utiliser un estimateur à base de transformée de Fourier permet une immunité plus importante vis-à-vis de l'estimateur de fréquence fondamentale, tout en s'affranchissant d'un modèle purement périodique de la voix. Le principal inconvénient d'une telle approche étant une décomposition de moins bonne qualité même dans les cas les plus favorables.
Malheureusement, ces méthodes ne produisent pas des signaux de bonne qualité. Il semblerait aussi que la décomposition PAP gagne à être modifiée pour lui appliquer les dernières avancées en matière de décomposition harmonique + bruit.

Apport Scientifique

Une nouvelle méthode a été proposée : PAP-A. La taille de la fenêtre minimum est précisée en fonction de la fréquence fondamentale afin d'optimiser la qualité de la décomposition. Une fonction de coût permet, tant que le coût ne dépasse pas un certain seuil, d'augmenter la taille de cette fenêtre pour augmenter la résolution de la décomposition. Deux méthodes sont alors testées, chacune d'entre elle incluant une modification supplémentaire par rapport à l'algorithme PAP originel : PAP-A et PAP-A sans fonction de coût. L'impact de la précision de l'estimation de F_0 est aussi évalué sur les signaux synthétiques.
Une écoute comparative entre décompositions est proposée au lecteur, afin de juger la qualité subjective de la décomposition pour chaque méthode.

En conclusion

Les résultats sur signaux synthétiques, suivant le même protocole que les méthodes précédemment proposées, montrent que la méthode PAP-A donne des résultats de meilleure qualité et moins sensibles à la précision d'estimation sur F_0. Sur les exemples sonores, on remarque notamment une tendance à mieux restituer les passages transitoires, même sur des signaux où la fréquence fondamentale élevée ne fait pas nécessairement appel à l'adaptation de la fenêtre d'observation.

Chapitre 5

Estimation des paramètres de la source glottique

Sommaire

5.1 Validation des Zéros de la Transformée en Z comme technique de séparation source/filtre . . . 131

5.1.1 Expériences sur signaux synthétiques . . . 132

Choix des méthodes de filtrage inverse . . . 132

Choix des paramètres pour l'onde de débit glottique . . . 132

Choix de la distance spectrale appliquée à l'estimation . . . 133

Protocole expérimental . . . 133

5.1.2 Résultats . . . 134

5.1.3 Discussion . . . 136

Discussion des résultats du corpus . . . 136

Paramètre fixe sur le corpus de paramètres standard . . . 136

Paramètre fixe sur la totalité des conditions . . . 136

Particularité du quotient de retour . . . 136

Conclusion pour les résultats sur signaux synthétiques . . . 137

5.1.4 Application aux signaux de parole naturelle . . . 137

Cas d'école . . . 138

Voix modale pour voyelles fermées . . . 139

Voix d'homme non modale . . . 140

Voix féminine modale . . . 141

5.1.5 Conclusion . . . 142

5.2 Précision nécessaire pour l'estimation de O_q et α_m . . . 143

5.3 Formalisation du modèle pour l'extraction des paramètres . . . 143

5.3.1 Equation liant O_q et α_m . . . 143

5.3.2 Estimation du formant glottique . . . 144

5.3.3 GCI et maximum de l'ODG . . . 144

5.3.4 Précautions . . . 146

Forme de la fenêtre d'oubli 146
Plage d'estimation et critère de forme 146
Précision de l'estimation 147
5.3.5 Algorithme proposé 147
5.4 Mesures préliminaires 149
5.4.1 Voyelles tenues 149
5.4.2 Résultats de l'estimation 149
5.4.3 Discussion 150
Locuteur masculin 150
Locuteur féminin 150
Asymétrie 151
5.5 Protocole d'analyse sur signaux naturels 152
5.5.1 Adaptation de l'algorithme aux signaux naturels 152
5.5.2 Résultat de l'algorithme sur de la voix parlée 154
5.5.3 Discussion 154
5.5.4 Conclusion 155
5.6 Méthode hybride combinant ZZT et LoMA pour l'estimation du quotient ouvert 155
5.6.1 Principe 155
5.6.2 Résultats de chaque méthode 156
5.6.3 Pondération 156
5.6.4 Facteur de certitude 158
5.6.5 Discussion et conclusion 158
5.7 Conclusion 160

Un dernier aspect de l'analyse des signaux vocaux est la caractérisation du débit glottique. Avant de chercher à estimer les paramètres d'une expression modélisant ce débit, il s'agit d'arriver à l'estimer convenablement. Pour ce faire, de nombreuses méthodes reposent aujourd'hui sur une modélisation auto-régressive du conduit vocal mais une nouvelle modélisation du débit glottique (modèle CALM[Doval *et al.*, 2003]) a permis de mettre en place une approche originale de décomposition source/filtre : la décomposition par ZZT [Bozkurt *et al.*, 2005] qui, il a été vu au chapitre 2, peut être considérée elle aussi comme un filtrage inverse.

Une fois obtenu le débit glottique, il s'agit alors d'estimer les paramètres de forme de l'onde de débit glottique. Contrairement à d'autres approches [Degottex *et al.*, 2010], nous ne rendrons pas la forme de cette onde à un seul paramètre, mais nous nous concentrerons sur la phase ouverte de l'onde, décrite par les deux paramètres que sont le quotient ouvert O_q et l'asymétrie α_m. Les travaux préliminaires à cette thèse [Sturmel *et al.*, 2006] ont montré une dépendance forte entre les valeurs de ces deux paramètres, imposant *de facto* une estimation jointe de ces deux paramètres.

Ce chapitre s'articulera en trois points. Dans un premier temps, la ZZT sera comparée à des méthodes de filtrage inverse basées sur la modélisation autorégressive du conduit vocal. Les résultats permettront de déterminer dans quelle mesure la ZZT est un choix intéressant pour estimer l'onde de débit glottique sur des signaux réels. Dans un deuxième temps, une méthode d'estimation conjointe du quotient ouvert de l'asymétrie sera présentée, basée à la fois sur les propriétés spectrales et temporelles de l'onde de débit glottique. Cette méthode sera optimisée au regard des résultats obtenus sur des signaux réels. Dans un troisième temps, il sera vu dans quelle mesure les résultats de cette méthode d'estimation peuvent être liés aux résultats obtenus par les LoMA (chapitre 3) pour l'estimation du quotient ouvert.

5.1 Validation des Zéros de la Transformée en Z comme technique de séparation source/filtre

L'approche sur laquelle se base la décomposition par ZZT propose une autre modélisation de la production vocale que la prédiction linéaire. Ce modèle causal/anti-causal propose une décomposition source/filtre par différence de l'évolution de la phase entre chaque étape (ouverte/fermée). La modélisation auto-régressive, plus classique, repose sur un *a priori* de forme de filtre vocalique et de la nécessité de compenser la pente spectrale du signal par une pré-accentuation.

Il est important, avant même d'utiliser la ZZT sur des signaux de parole naturelle pour en extraire les paramètres du modèle de production vocale, de vérifier si la modélisation causale/anticausale est capable de réaliser des filtrages inverses de qualité suffisante par rapport à la prédiction linéaire. Nous cherchons donc à opposer la ZZT à la prédiction linéaire sur ces deux plans : d'une part la modélisation auto-régressive du conduit vocal et de l'autre l'utilisation d'une pré-accentuation des signaux.

Dans une recherche d'objectivité, des analyses comparatives seront tout d'abord faites sur un corpus de signaux synthétiques regroupant une excursion complète des paramètres du débit glottique, des configurations du filtre vocalique et de la quantité de voisement. Dans un second temps, des résultats sur parole naturelle seront commentés.

On pourrait légitimement se poser la question de la comparaison de la ZZT aux récents (par rapport à la prédiction linéaire) avancements en la matière, et notamment les méthodes non linéaires par contraintes de type ARX comme ARX-LF[Vincent *et al.*, 2007]. Le but de cette étude n'est pas de prouver que la ZZT en elle même peut rivaliser avec ces méthodes contraintes,

mais qu'elle peut servir de base à des méthodes non linéaires et contraintes par un modèle en lieu et place de la modélisation auto-régressive. Ces travaux ont d'ailleurs servi à la mise en place d'estimations non linéaires notamment par Drugman et al. [Drugman *et al.*, 2008], ou pour une estimation directe [Sturmel et d'Alessandro, 2010].

5.1.1 Expériences sur signaux synthétiques

Choix des méthodes de filtrage inverse

Deux méthodes de filtrage inverse ont été retenues. Elle ont été choisies pour leur notoriété et leur simplicité similaire au principe de la ZZT : elles utilisent des informations sur le modèle de production vocale (GCI ou filtre autorégressif et/ou coefficient de pré-accentuation) sans contraintes supplémentaires.

La première méthode [Makhoul, 1975] réalise un filtrage inverse par estimation du filtre à partir de la méthode d'autocorrélation. C'est une méthode simple qui a fait ses preuves et reste une référence en matière de filtrage inverse. Elle est basée sur un estimateur à base d'autocorrélation. Cette méthode est utilisée pour l'estimation d'un filtre auto-régressif de 18 pôles (suivant la règle de Fe/1000+2 pôles, où Fe est la fréquence d'échantillonnage du signal, fixée à 16kHz) ainsi que d'une préaccentuation (filtre d'expression $1 - 0.98z^{-1}$). Cette méthode est utilisée dans une implémentation fournie par le paquet de fonctions VoiceBox [Brookes, 2000] pour MATLAB© .

La seconde méthode, IAIF (Alku, [Alku, 1992]), est basée sur une prédiction linéaire utilisant un ajustement itératif du coefficient de préaccentuation. Tout comme pour l'autocorrélation, on utilise un nombre de pôles égal à 18 pour l'estimation. L'implémentation utilisée fait partie du paquet "aparat" [Airas *et al.*, 2005]. Elle utilise aussi un estimateur à base d'autocorrélation, mais l'ajustement adaptatif de la préaccentuation la rend particulière.

Choix des paramètres pour l'onde de débit glottique

L'onde de débit glottique retenue est celle du modèle LF, nous utilisons 4 paramètres pour contrôler les signaux synthétisés : F_0 (fréquence fondamentale), O_q (quotient ouvert), α_m (asymétrie), Q_a (phase de retour). Á ces 4 paramètres vient s'ajouter une composante importante de la phonation dans le cas de la qualité vocale : le bruit de voisement. Ce bruit sera rajouté directement à la source (par conséquent, filtré par la suite) ; il est contrôlé en énergie par rapport à l'énergie de l'onde de débit glottique. Il sera abrégé RVB pour Rapport Voisement sur Bruit (en dB). Il a été montré que le bruit est un facteur important vis à vis de la stabilité de la décomposition ZZT [Bozkurt, 2005]. Un paramètre est absent : le paramètre E de l'amplitude de la dérivée de l'onde de débit glottique. Toutes les synthèses seront faites à E constant. Les valeurs retenues sont présentées dans le tableau 5.1. Étant donné le nombre important de conditions de voisement qui seront testées, le corpus est décomposé en trois sous-corpus afin de faciliter le dépouillement et la discussion des résultats :

- un sous-corpus de voisement sans bruit regroupant toutes les variations des paramètres suscités à l'exception du bruit qui reste tout le temps à son niveau minimum. Ce sous-corpus regroupe un total de 13720 signaux de test.
- un sous-corpus de voisement standard, regroupant uniquement une variation de voyelle et des valeurs fixes pour les autres paramètres : $F_0 = 150Hz$, $O_q = 0.5$, $\alpha_m = 0.8$, $Q_a = 0.05$ et RVB=60dB pour la totalité des 10 voyelles testées ; soit un total de 10 signaux tests.
- le corpus complet, intégrant donc la variation du bruit de phonation au premier sous-corpus, pour un total de 54880 signaux testés.

Table 5.1 – Table des valeurs choisies pour la variation des paramètres lors de la création de la base de données de signaux synthétiques pour un total de 54880 conditions de tests.

F_0	$90, 120, 150, 180, 210, 250, 300$
O_q	$0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$
α_m	$0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9$
Q_a	$0, 0.05, 0.1, 0.15$
Voyelle	[a], [æ], [ʌ], [ɛ], [o], [e], [u], [ɯ], [y], [i]
RVB	$120dB, 60dB, 30dB, 15dB$

Choix de la distance spectrale appliquée à l'estimation

Dans le cas des signaux synthétiques, le signal original est connu et il est donc possible d'utiliser des critères objectifs pour évaluer la qualité de la décomposition source/filtre. Dans ce chapitre, l'intérêt se porte sur la qualité de restitution du signal de débit glottique effectué par la ZZT. Juger les signaux sur leur aspect temporel peut être problématique : de grosses distorsions peuvent passer inaperçues, notamment en basse fréquence. Or, dans la lignée de l'analyse spectrale des ondes de débit glottique [Doval *et al.*, 2006], une attention particulière se porte sur la forme du spectre estimé plutôt que sur la forme temporelle du signal. Ainsi, le critère retenu est une différence spectrale dont l'équation est présentée en 5.1.

$$e = \sqrt{\int_{f=0}^{4000} (Est_{dB}(f) - Ref_{dB}(f))^2 df} = \sqrt{\int_{f=0}^{4000} \left(20log\frac{|Est(f)|}{|Ref(f)|}\right)^2 df} \tag{5.1}$$

Où $Est(f)$ et $Ref(f)$ sont respectivement le spectre de la source estimée et de celle de référence. Est_{dB} et Ref_{dB} sont leurs spectres de puissance en dB.

Ce type de distance présente deux problèmes :

- En premier lieu, il faut observer que le signal glottique ne présente que très peu d'informations au-delà de 4kHz. Cette fréquence est donc choisie comme limite maximum du calcul de la distance sur le spectre.
- Une différence de gain entre les spectres peut engendrer une erreur trop importante. Ce problème est contourné en normalisant tous les spectres par l'amplitude du premier harmonique (le fondamental). Ainsi, et à l'instar des mesures de Hanson et al. [Hanson, 1994], c'est l'estimation de la forme du spectre par rapport au premier harmonique qui va prévaloir sur le résultat de la distance obtenue.
- Cette distance ne tenant pas compte de la phase des composantes du spectre, une distance de zéro décibel ne signifie pas une erreur de prédiction nulle, mais bien des spectres d'amplitude identiques. Il existe cependant une infinité de signaux différents ayant le même spectre d'amplitude.

Protocole expérimental

Le protocole de création de la base de données est le suivant :

1. synthèse d'une forme d'onde dérivée de débit glottique suivant une des 54880 combinaisons de paramètres décrites dans le tableau 5.1.
2. filtrage de cette onde de débit glottique par un des 10 filtres de conduit vocal.
3. application des méthodes de filtrage inverse (ZZT, IAIF et LP par autocorrélation).
4. mesure de la distance spectrale.

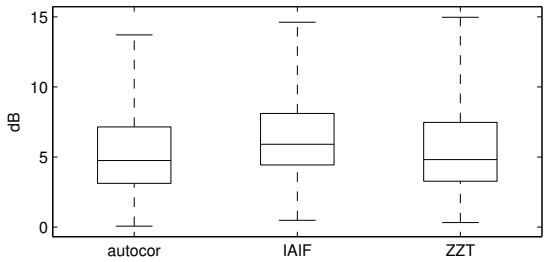


Figure 5.1 – Boîtes à moustaches (voir texte) représentant la distance spectrale pour chaque méthode d'estimation pour le sous-corpus de voisement sans bruit.

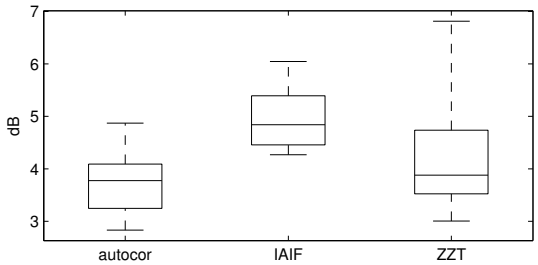


Figure 5.2 – Boîtes à moustaches (voir texte) représentant la distance spectrale pour chaque méthode d'estimation pour le sous-corpus de voisement standard.

5.1.2 Résultats

Dans un premier temps, les résultats regroupés par sous-corpus sont proposés sous forme de boîte à moustaches sur les figures 5.1, 5.2 et 5.3. La boite à moustaches est utilisée car les distributions ne sont pas Gaussiennes : l'utilisation d'une représentation de type moyenne + écart type serait erronée. La boite à moustaches présente la distribution par quartiles. La médiane est située au milieu de la boite et les quartiles supérieurs (médiane des valeurs entre la médiane et la valeur maximum) et inférieurs forment les extrêmes de la boîte.

On remarque d'emblée que les résultats sont très serrés avec un léger avantage pour l'autocorrélation dans tous les cas de figure. La différence entre le corpus de voisement sans bruit et le corpus complet est négligeable, montrant que le bruit a peu ou pas d'effet sur la qualité de la décomposition. Sur le sous-corpus des conditions standard, la ZZT donne des résultats sensiblement meilleurs que les autres sous-corpus.

Dans un deuxième temps, analysons les résultats sur l'excursion d'un paramètre à la fois. C'est-à-dire : la moyenne des distances sur le corpus pour toutes les variations des paramètres sauf celui considéré. Notons que le corpus standard ne permet pas ce genre d'analyse de par sa composition (une seule valeur par paramètre glottique). L'excursion est alors faite pour tous les paramètres fixés aux valeurs standard à l'exception de celui qui est analysé. Le sous-corpus de voisement sans bruit et le corpus complet présentent des distributions très similaires. L'analyse des variations d'un paramètre dans les visualisations éclatées (figure 5.4) sera donc limitée au corpus complet et au sous-corpus de voisement standard présentés sur la figure 5.4.

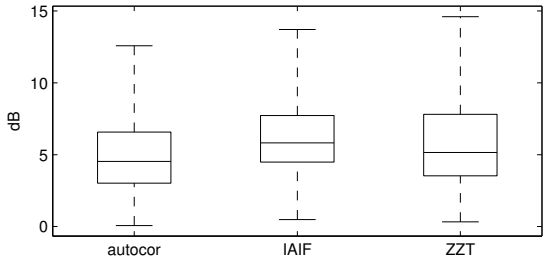


Figure 5.3 – Boîtes à moustaches représentant la distance spectrale pour chaque méthode d'estimation pour le corpus complet.

Paramètres Standard

Totalité des conditions

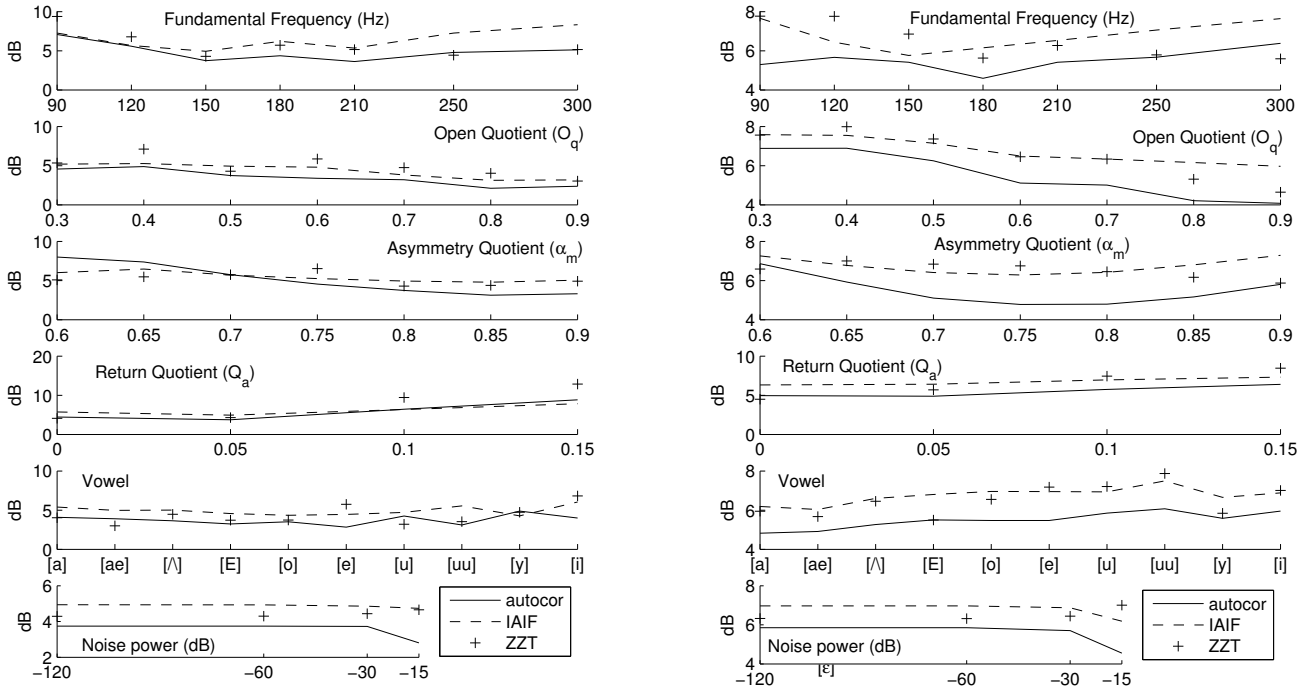


Figure 5.4 – Représentation détaillée des distances spectrales pour le sous-corpus standard et le corpus complet. Moyenne des distances pour les conditions correspondant à la valeur du paramètre fixe.

5.1.3 Discussion

Discussion des résultats du corpus

Les boîtes à moustaches sur les figures 5.1, 5.2, 5.3 présentent des distributions comparables pour toutes les méthodes, à 2-3dB d'erreur près. Dans les trois cas, les médianes peuvent se classer de la plus petite à la plus grande : Autocor, ZZT, IAIF ; plaçant le filtrage inverse par ZZT dans des performances comparables aux deux méthodes de filtrage inverse.

L'amplitude de la distribution peut se classer de la plus petite à la plus grande comme suit : Autocor, IAIF, ZZT. Les résultats par ZZT ne possèdent pas encore la stabilité des résultats par prédiction linéaire ; les analyses "divergentes" poussent donc la limite supérieure de la distribution de la ZZT plus loin que celle des méthodes à base de prédiction linéaire. On peut toutefois noter que la limite supérieure de la boite (premier quartile au dessus de la médiane) est toujours inférieure à celle de IAIF. Ainsi, plus de 75% des analyses par ZZT donnent des résultats comparables aux deux méthodes de prédiction linéaire.

L'analyse des résultats par ANOVA donne des moyennes statistiquement indépendantes ($p<0.01$) pour tous les sous-corpus sauf pour le set de conditions standard contenant par ailleurs un nombre d'items très faibles (10). Ce qui justifie d'autant plus les visualisations présentées sur la figure 5.4.

Paramètre fixe sur le corpus de paramètres standard

L'étude détaillée des résultats sur le corpus de paramètres standard montre des courbes très proches l'une de l'autre. Les résultats de la ZZT se situent globalement entre ceux de l'autocorrélation (meilleurs sauf pour les faibles asymétries) et ceux de IAIF. Le bruit additif n'a que peu d'influence sur la qualité de la décomposition par ZZT et IAIF, mais un niveau de bruit important (-15dB) augmente les performances de l'autocorrélation. Ce comportement est attendu, l'estimation du filtre autorégressif supposant une excitation du conduit vocal par un bruit blanc.

La ZZT se distingue par des résultats moins bons sur les voyelles antérieures non arrondies : [e] et [i] ayant un premier formant bas et un deuxième formant haut. La ZZT montre ainsi des distances plus importantes pour des valeurs élevées du quotient de retour. Le quotient de retour ajoute un filtrage passe bas supplémentaire sur l'onde de débit glottique ce qui diminue les hautes fréquences et rend la forme de l'onde au GCI moins abrupte.

Paramètre fixe sur la totalité des conditions

Sur la totalité des conditions, on retrouve les même évolutions que précédemment, mais la différence entre autocorrélation et IAIF est globalement plus marquée. Les résultats de la ZZT se situent au niveau de ceux de IAIF. Pour l'asymétrie, l'évolution est différente de celle du corpus de paramètres standard et on visualise une courbure de l'erreur pour l'autocorrélation, ce creux se situe autour de $\alpha_m = 0.75$. On ne retrouve pas cette forme sur la courbe de IAIF, ce qui pousse à penser que ce creux est le résultat du choix du coefficient de préaccentuation.

Particularité du quotient de retour

Les formes d'ondes synthétisées possèdent une phase de retour variable, avec Q_a variant de 0 à 0.15. La ZZT, *de facto* permet d'estimer uniquement la phase ouverte du débit glottique, et donc devrait donner les résultats les plus mauvais pour les quatrièmes graphiques de la figure 5.4. Or, les méthodes à base de filtrage inverse ne produisent pas des résultats de bien meilleure qualité, et on observe une tendance systématique d'augmentation de l'erreur quand Q_a augmente, quelle

que soit la méthode ou le corpus. Pour les paramètres standard, les performances de la ZZT sont comparables à celle des méthodes de prédiction linéaire pour $Q_a = \{0; 0,5\}$.

Ces résultats peuvent s'expliquer par deux hypothèses :

- De prime abord, comme la ZZT ne reconstitue pas du tout la phase de retour, on peut déduire que les méthodes à base de prédiction linéaire, elles aussi, ne sont pas capable de restituer correctement cette composante du débit glottique. Á défaut commun, les trois méthodes offrent donc des performances similaires.
- Cependant, on constate aussi que les DODG obtenues par ZZT possèdent une composante continue (i.e. : l'équation implicite $\int_T \frac{dg}{dt}(t)dt = 0$ n'est pas vérifiée). Ceci ayant tendance à confirmer que ces formes d'ondes sont bien des DODG tronquées de leur phase de retour (la phase de retour manquante produisant une air supplémentaire pour vérifier l'équation implicite). Dans ce cas, le spectre de la phase ouverte sera alors variable en fonction de la valeur de Q_a. Cette hypothèse là n'expliquant que partiellement les résultats obtenus.

L'exploration des pistes ouvertes par ces hypothèses n'a malheureusement pas permis de répondre de manière tranchée à cette question, et il est fort probable que ces deux phénomènes jouent un rôle concomitant. Les résultats de la figure 5.4 montrent que l'inaptitude de la ZZT à estimer une phase de retour n'est pas un facteur limitant par rapport à la prédiction linéaire. Au contraire, il semblerait que la phase ouverte estimée par ZZT comporte déjà des informations sur la phase de retour.

Conclusion pour les résultats sur signaux synthétiques

De prime abord, la qualité des résultats obtenus par la prédiction linéaire à base d'autocorrélation peuvent surprendre. En effet, cette méthode est non seulement une méthode antérieure aux deux autres, mais aussi la plus simple. La qualité des résultats trahit probablement le protocole simplifié pour synthétiser les signaux de tests : l'utilisation d'un modèle de synthèse très proche (filtre AR + source) du principe de décomposition par LPC pèse probablement dans la balance. On remarquera toutefois que si IAIF présente des résultats moins bons, ils sont tout de même plus stables vis à vis de la variation des paramètres sur la figure 5.4, à l'exception de la variation de fréquence fondamentale.

Ce test sur signaux synthétiques a montré la viabilité de la méthode ZZT comme méthode de décomposition source/filtre. Sur une large plage de variation des paramètres de la configuration, la ZZT donne des résultats comparables aux méthodes à base de modélisation autorégressive du conduit vocal. Il faut cependant noter que la génération des signaux synthétiques utilise cette même modélisation autorégressive : c'est le point faible de cette étude. Afin d'être non biaisée, un synthétiseur à modèle physique aurait dû être utilisé mais aurait posé des difficulté de contrôle de paramètres de synthèse.

5.1.4 Application aux signaux de parole naturelle

Dans le cas des signaux de parole naturelle il n'existe pas de moyen d'obtenir la forme d'onde précise du débit glottique. Utiliser une distance spectrale est donc impossible. Restent alors deux moyens de discuter la qualité de la forme d'onde obtenue par séparation source / filtre :

- la discussion par critère de forme de l'onde de débit glottique obtenue. Sujette à de forts *a priori* mais efficace pour qualifier une décomposition réaliste ou non, à l'instar des mesures proposées ailleurs [Moores et Torres, 2008, Alku *et al.*, 2005],
- la discussion basée sur la détection des instants d'ouverture en référence à des signaux EGG. Car comme pour les tests sur signaux synthétiques, les GCI sont connus pour permettre l'analyse par ZZT.

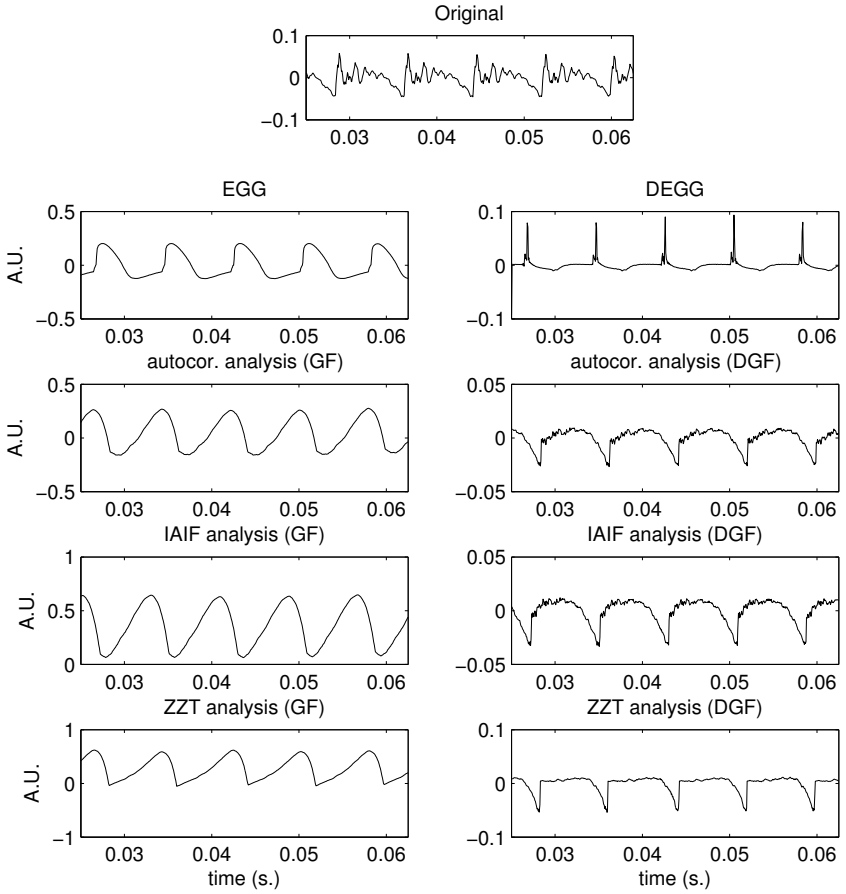


Figure 5.5 – Analyse d'un signal de voix produit par un locuteur, voyelle /a/ de fréquence fondamentale proche de 120Hz, voix modale. De haut en bas : signal original, signal EGG, analyse par autocorrélation, analyse itérative (IAIF) et analyse ZZT. Les 3 analyses donnent des résultats similaires.

Les signaux analysés sont disponibles pour écoute sur le site web [1]. Dans l'ordre d'analyse, ils correspondent aux fichiers : `ZZT-C1.wav`, `ZZT-C3.wav`, `ZZT-C5.wav` et `ZZT-M1.wav`.

Cas d'école

On se place dans un premier temps dans le cas d'une phonation classique : voix modale d'un homme (environ 120Hz) prononcée clairement sans expressivité vocale particulière (voix neutre) et pour la voyelle /a/. Les résultats, présentés sur la figure 5.5 sont très bons pour toutes les méthodes, sans surprise vis-à-vis de la simplicité du signal. La faible valeur de la fréquence fondamentale associée à une valeur moyenne de Oq placent le formant glottique largement en dessous de la valeur du premier formant de la voyelle (environ 800Hz). Les conditions sont telles qu'une préaccentuation normale avec un coefficient de 0.98 suffit à compenser la pente spectrale de l'asymétrie du débit glottique. Aucune distorsion perceptible de l'onde de débit glottique n'est

1. http ://nicolas.sturmel.com/PHD/

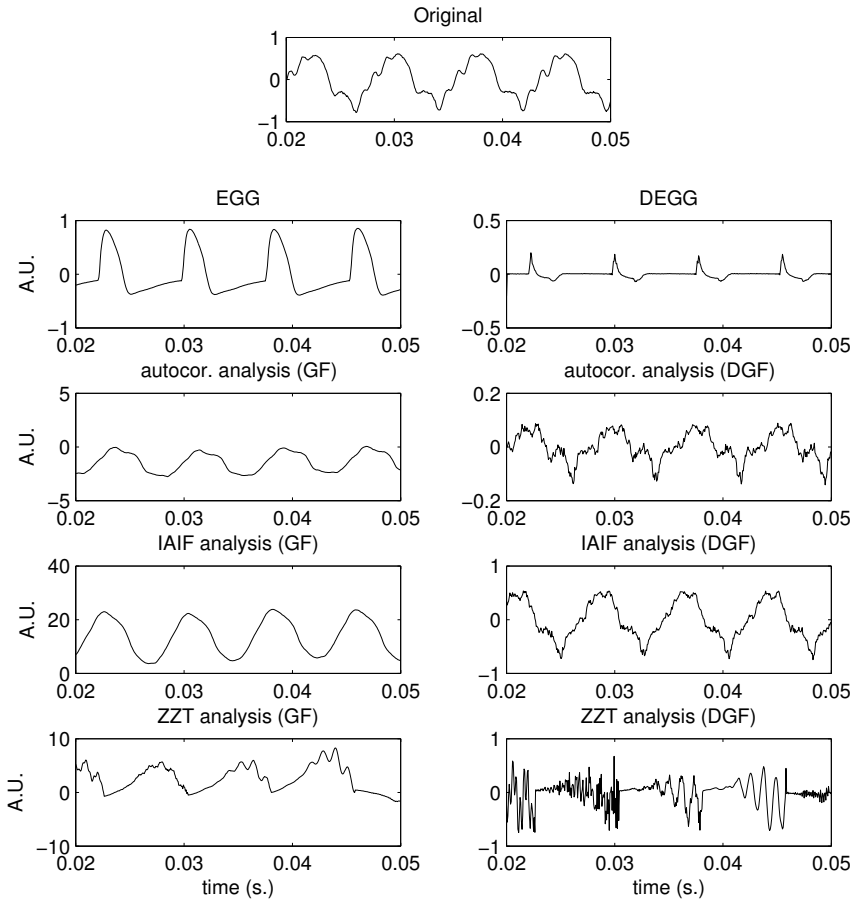


Figure 5.6 – Analyse d'un signal de voix produit par un locuteur, voyelle /u/ de fréquence fondamentale proche de 120Hz, voix modale. De haut en bas : signal original, signal EGG, analyse par autocorrélation, analyse itérative (IAIF) et analyse ZZT.

présente pour aucune des analyses présentées. Les ondes obtenues par LPC semblent toutefois plus bruitées sur la partie fermée du cycle glottique.

Voix modale pour voyelles fermées

Dans le cas des voyelles fermées dont un exemple est présenté sur la figure 5.6, le premier formant vocalique se situe bien plus bas en fréquence, ce qui peut mener à de mauvaises estimations de la forme du filtre vocalique et du spectre du débit glottique. On remarque notamment que la forme de la DODG pour les méthodes à base de prédiction linéaire est très proche de la forme originale du signal, ce qui empêche une estimation convenable du quotient ouvert (on mesure sur l'EGG un quotient ouvert d'environ 0.7) et indique une séparation incomplète du premier formant.

La DODG obtenue par ZZT présente des oscillations importantes, mais elles sont nettement moins visibles sur l'onde de débit glottique. Sur cette même onde, on retrouve une forme permettant de déduire un quotient ouvert plus proche de celui ouvert mesuré sur l'EGG. Ainsi, ce

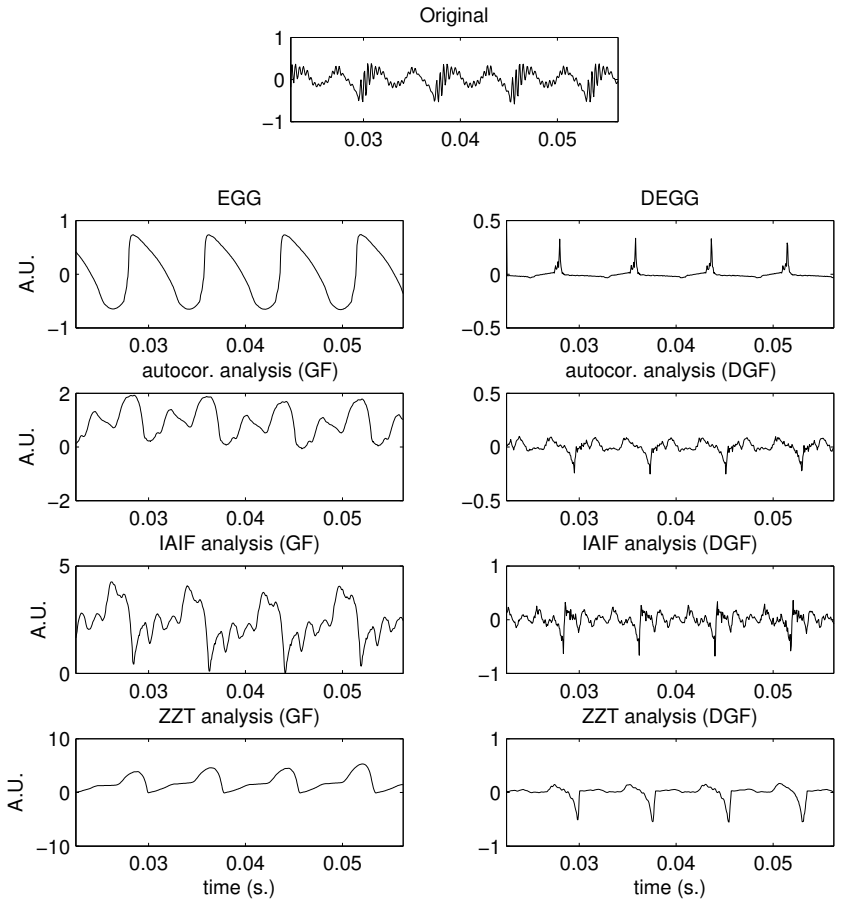


Figure 5.7 – Analyse d'un signal de voix produit par un locuteur, voyelle /i/ de fréquence fondamentale proche de 120Hz, voix serrée. De haut en bas : signal original, signal EGG, analyse par autocorrélation, analyse itérative (IAIF) et analyse ZZT.

signal est caractéristique des comparaisons mises en avant sur le test synthétique : si d'emblée la DODG obtenue par ZZT semble très mauvaise, elle permet d'obtenir un débit glottique beaucoup plus asymétrique et d'y visualiser un quotient ouvert plus proche de la valeur de 0.7 mesurée sur l'EGG.

Voix d'homme non modale

Lorsque la qualité vocale varie, le formant glottique se déplace autour de F_0 et peut rentrer en conflit avec le premier formant vocalique, en particulier lors de voisements serrés pour lesquels la valeur de O_q s'approche de 0.3 et donc où la fréquence du formant glottique augmente.

Dans l'exemple présenté sur la figure 5.7, une voyelle /i/ est prononcée avec un quotient ouvert proche de 0.3 mesuré sur l'EGG. Dans ces conditions là, la méthode ZZT présente des résultats remarquables en comparaison de la prédiction linéaire. C'est aussi la seule méthode grâce à laquelle on peut retrouver la valeur du quotient ouvert à la fois sur l'ODG et sur la DODG.

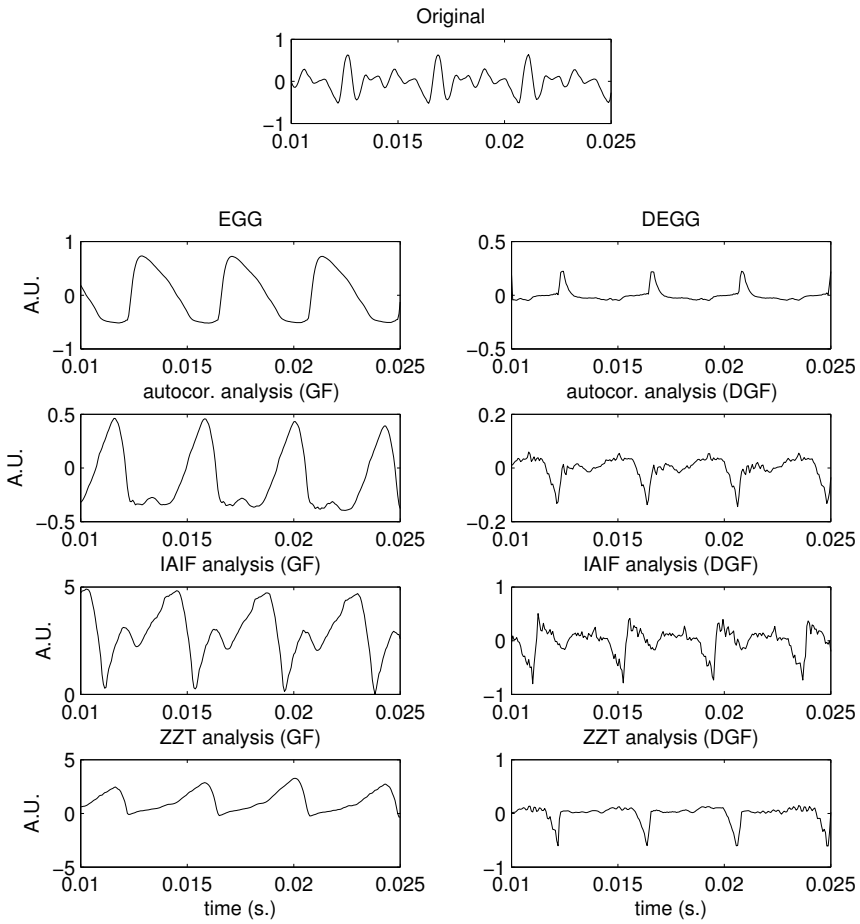


Figure 5.8 – Analyse d'un signal de voix produit par une locutrice, voyelle /a/ de fréquence fondamentale proche de 240Hz, voix modale. De haut en bas : signal original, signal EGG, analyse par autocorrélation, analyse itérative (IAIF) et analyse ZZT.

Le facteur déterminant dans la qualité de la décomposition vient probablement de la phase fermée assez longue qui permet de séparer distinctement la réponse impulsionnelle du filtre. On retrouve une forme de phase ouverte assez proche de celle obtenue par la ZZT sur la DODG estimée par autocorrélation. La méthode IAIF a nettement supprimé la contribution du premier formant vocalique en ajustant la pré-accentuation.

Voix féminine modale

Le dernier exemple présenté sur la figure 5.8 concerne un /a/ produit par un sujet féminin. L'autocorrélation semble donner les résultats de meilleure qualité, notamment du point de vue de la DODG, mais encore une fois la forme de la dérivée obtenue par ZZT est trompeuse et le flux glottique sur la gauche est parfaitement exploitable pour y mesurer le quotient ouvert.

5.1.5 Conclusion

Cette section a traité de la comparaison de la méthode de déconvolution par ZZT à deux méthodes de filtrage inverse : IAIF [Alku, 1992] et la prédiction linéaire par autocorrélation [Makhoul, 1975]. Ces deux méthodes sont différentes dans leur manière de modéliser la production vocale : d'un coté la LPC considère le filtre vocalique comme auto-régressif alors que de l'autre la ZZT a besoin des instants de fermeture glottique pour la décomposition causale/anticausale. Comme il est impossible de mesurer objectivement les qualités d'un filtrage inverse sur une base de parole naturelle, une base de données synthétiques obtenue par pavage sur le domaine des paramètres du modèle a été créée. Ce principe de test a depuis été repris dans d'autres travaux [Drugman *et al.*, 2008, Drugman *et al.*, 2009b]. L'étude de cette grande base de données a permis d'automatiser l'évaluation des méthodes. Le dépouillement des résultats de la base de données synthétiques a été divisé en 3 groupes : des cas les plus courants aux cas les moins courants de configuration du modèle de production vocale. L'évaluation objective s'est faite au travers d'un critère de distance spectrale entre le signal de source orignal et celui estimé par les méthodes retenues.

Le protocole utilisé dans cette section est contestable sur plusieurs points :

- Le choix restreint de deux méthodes de filtrage inverse qui ne reflètent pas les derniers travaux en la matière est motivé par la volonté d'opposer le principe de modélisation du débit glottique et non la mise en œuvre de la décomposition. Les méthodes retenues injectent un minimum de connaissance sur le signal pour en estimer chaque composante. Tout comme la méthode ARX-LF [Vincent *et al.*, 2007], nous espérons qu'un tel travail servir a comme preuve de la viabilité du modèle causal/anticausal, en motivant l'utilisation de la ZZT comme décomposition de base pour une estimation contrainte.
- Le choix de la distance spectrale est motivé par la tendance actuelle à estimer les paramètres et à décrire les ondes de débit glottique par leurs propriétés spectrales [Doval *et al.*, 2006]. Les exemples sur signaux naturels ont par ailleurs montré qu'une forme de DODG semblant mauvaise pouvait tout de même donner une ODG exploitable.
- Le choix de la méthode de génération des signaux synthétiques est contestable, dans la mesure où il utilise le paradigme même de la modélisation autorégressive pour générer les signaux. Un synthétiseur à modèle physique aurait été plus approprié, mais n'aurait pas permis un contrôle aussi précis sur les paramètres de l'onde de débit glottique.

Enfin, les tests sur des signaux de parole naturelle nous montrent que l'analyse par ZZT a tendance à restituer plus fidèlement les phases du débit glottique vis à vis des connaissances que nous en avons par EGG. Bien sur, la prédiction linéaire est utilisée avec raison depuis des années pour le filtrage inverse, mais dans des conditions d'analyse automatique il n'existe pas de méthode capable de mesurer le quotient ouvert sur son résidu (par exemple). Le développement d'indicateurs "annexes" (comme NAQ [Alku et Bäckström, 2002]) indique clairement une incapacité à restituer la forme de l'ODG dans son intégralité, en particulier dans les cas limites de configurations glottiques : quotient ouvert très élevé ou très bas.

Nous avons donc montré que la ZZT était une méthode compétitive, basée sur un modèle original et viable de la production vocale. Seule, elle ne permet pas de surpasser largement les techniques à base de prédiction linéaire, mais les débits glottiques obtenus semblent plus facilement exploitables pour une extraction des paramètres. Ceci sera exploré au court de ce chapitre où un principe d'estimation conjointe du quotient ouvert et de l'asymétrie sera présenté et évalué.

5.2 Précision nécessaire pour l'estimation de O_q et α_m

Il a été vu au chapitre 1 que les rôles du quotient ouvert et de l'asymétrie sont prépondérants dans la forme du signal vocal. Afin d'étudier la voix et le rapport entre configuration glottique et qualité vocale il est nécessaire d'arriver à estimer ces paramètres sur le signal vocal. Lorsque le signal EGG est disponible, il peut servir de référence pour évaluer la méthode proposée. C'est le quotient ouvert qui sera utilisé comme critère principal de l'évaluation de l'algorithme. Pour l'instant, cette mesure est la seule présentant une fiabilité suffisante pour être considérée comme une référence.

Les travaux concernant le seuil différentiel perceptif (JND - *Just Noticeable Difference*) [Henrich *et al.*, 2003] sur la valeur du quotient ouvert nous permettent de mettre une limite inférieure à la précision de l'estimation de O_q. En l'occurrence, il s'agit de 17%. Ainsi, le but principal de l'estimation de ce paramètre sur des signaux naturels sera d'arriver à 100% d'estimation sous ce seuil différenciel.

Dans un second temps, on peut séparer la plage de variation de O_q en trois, autour des valeurs centrales : 0.3, 0.5 et 0.8, correspondant généralement à des voix tendues, normales, lâchées respectivement. La précision minimum pour pouvoir discriminer deux valeurs voisines est alors de 25%. Un critère secondaire va donc consister en la quantité de détections réalisées par l'algorithme à l'intérieur de cette plage d'erreur de 25%. Dans le cas où une estimation précise (sous le JND) n'est pas possible, cet indicateur grossier permettra tout de même de caractériser en partie les signaux vocaux.

5.3 Formalisation du modèle pour l'extraction des paramètres

Le forme de la source glottique (donnée au chapitre 1 à la figure 1.7) possède deux jeux de paramétrisation : un jeu temporel (T_e, T_p, T_a) et un jeu relatif à la fréquence fondamentale F_0 (O_q, α_m, Q_a). Le quotient ouvert O_q est défini comme le rapport $\frac{T_e}{T_0}$, α_m, l'asymétrie du débit glottique est définie comme le rapport $\frac{T_p}{T_e}$. La grandeur centrale autour de ces paramètres est donc la durée de la phase ouverte T_e, qui dépend aussi de la fréquence du formant glottique ω_g, pulsation d'oscillation de la dérivée du flux glottique visible sur la figure 1.7. Il a été noté que le maximum du spectre, ayant pour fréquence le formant glottique, n'est pas nécessairement égal à $\frac{\omega_g}{2\pi}$. En première approximation, la pulsation de modulation du modèle LF et la fréquence du formant glottique seront confondues. Ceci sera d'autant moins vrai que α_m sera proche de 1. La durée d'une demi période de F_g est alors située entre deux passages par zéro du flux glottique, c'est la grandeur T_p. Il vient donc des travaux de [Fant *et al.*, 1985] et de l'approximation précédente, l'équation 5.2 définissant le temps T_p en fonction de F_g.

$$T_p = \frac{1}{2F_g} \tag{5.2}$$

Afin de connaître T_e il est donc nécessaire de déterminer le temps situé entre l'instant de fermeture glottique (le GCI) et le maximum du débit glottique, soit A ce temps donné selon l'équation 5.3.

$$A = T_e - T_p \tag{5.3}$$

5.3.1 Equation liant O_q et α_m

On peut donc déterminer les expressions des deux coefficients comme suit :

$$\alpha_m = \frac{T_p}{T_e} \tag{5.4}$$

$$= \frac{1}{1+2AF_g} \tag{5.5}$$

$$O_q = \frac{T_e}{T_0} \tag{5.6}$$

$$= \frac{1}{T_0}(\frac{1}{2F_g} + A) \tag{5.7}$$

En considérant donc F_g, T_0 et A connus, on peut déduire O_q et α_m grâce aux équations 5.5 et 5.7. Pour étudier les interdépendances entre ces paramètres, il peut être judicieux de définir F_g et A en fraction de T_0. Sur la figure 5.9 sont présentés des abaques illustrant les équations ci-avant pour des valeurs habituelles de O_q et α_m. On remarque que contrairement à un raccourci utilisé dans d'autres études et notamment [Bozkurt *et al.*, 2004b], la fréquence du formant glottique dépend de l'asymétrie et la seule connaissance de F_g ne permet donc pas de déterminer O_q avec une précision suffisante. En particulier dans le cas de la parole expressive, où la valeur de α_m peut varier grandement (effort plus marqué).

5.3.2 Estimation du formant glottique

Le formant glottique F_g est estimé sur le retard de groupe τ_g ([Murthy *et al.*, 1989, Murthy et Yegnanarayana, 1991]) de la partie anti-causale de la décomposition par ZZT en y cherchant le minimum comme indiqué sur l'équation 5.8. Ce retard de groupe est calculé d'après la phase $\Phi[G_\rho]$ du spectre G_ρ calculé dans le plan complexe sur le cercle de rayon ρ à partir des N zéros Z_j tels que $|Z_j| > 1$

$$G_\rho(\nu) = \prod_{j=1}^{N}(\rho e^{2i\pi\nu} - Z_j)$$

$$\tau_g = \frac{d\Phi[G_\rho](\nu)}{2\pi d\nu}$$

$$F_g = \underset{\nu}{\operatorname{argmin}}\,(\tau_g(\nu)) \tag{5.8}$$

Pour augmenter la stabilité de l'estimation de F_g, on calcule le spectre de la partie anticausale sur un cercle plus petit que le cercle unité, généralement de diamètre $\rho = 0.98$ [Bozkurt *et al.*, 2007]. On limite ainsi les ondulations de phase causées par des zéros trop proches du cercle unité. Ceci revient aussi à lisser la phase ainsi calculée pour ne laisser transparaitre que la forme globale et ainsi mesurer F_g de manière robuste vis-à-vis des instabilités du calcul. Pour contraindre l'estimation dans des valeurs réalistes, on cherche un minimum local situé entre 0.5 et 3 fois F_0.

Soit $Z(n)$ l'ensemble de N zéros anticausaux, alors on calcule τ_g du spectre $X(\omega)$ sur Ω points pour le rayon ρ selon l'équation 5.9.

$$|X(\omega)|e^{i\ \Phi(\omega)} = \prod_{j=0}^{N-1}(\rho e^{2i\pi\frac{\omega}{\Omega}} - Z(j))$$

$$\tau_g = \frac{d\Phi}{d\omega} \tag{5.9}$$

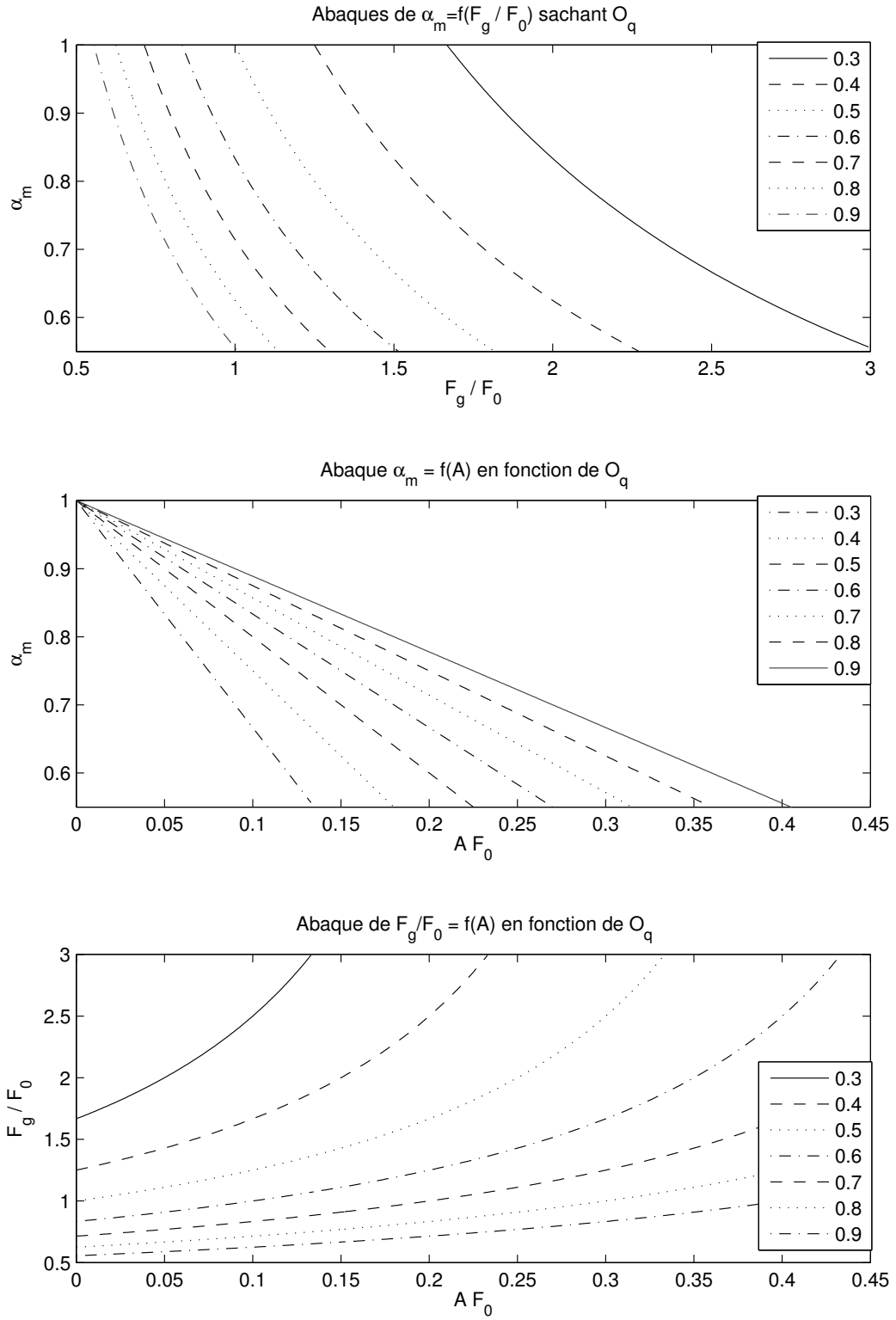


Figure 5.9 – Abaques montrant les variations, respectivement de haut en bas, de α_m en fonction du rapport $\frac{F_g}{F_0}$, de α_m en fonction de la grandeur A (normalisée par T_0) et finalement du rapport $\frac{F_g}{F_0}$ en fonction de A (normalisée par T_0) pour différentes valeurs de O_q allant de 0.3 à 0.9.

5.3.3 GCI et maximum de l'ODG

Une fois obtenue la fréquence du formant glottique, il reste à déterminer la quantité A. Cette quantité est représentée par le temps qui s'écoule entre le maximum de l'ODG et le GCI. Le GCI est censé être parfaitement connu lors d'une analyse par ZZT comme paramètre de la décomposition. En pratique l'erreur d'estimation sur le GCI par LoMA est suffisamment petite pour permettre une décomposition de bonne qualité. Par conséquent, cet instant est considéré donné. Le maximum de l'ODG est estimé directement sur la synthèse $G_{zzt}(t)$ de la partie anticausale obtenue par ZZT comme le maximum de la forme d'onde, ainsi :

$$A = T_0 - \underset{t}{\operatorname{argmax}}\,(G_{zzt}(t))$$

5.3.4 Précautions

Afin de maximiser la stabilité de l'estimation jointe $[O_q,\, \alpha_m]$, un certain nombre de règles et de précautions sont appliquées lors de l'estimation.

Forme de la fenêtre d'oubli

Contrairement à la fenêtre retenue pour le filtrage inverse par ZZT, on utilise ici la forme proposée dans [Drugman *et al.*, 2009b]. Ainsi pour les fenêtres de taille N de la famille :

$$w(t) = \frac{\alpha}{2} - \frac{1}{2}cos(\frac{2\pi n}{N-1}) - \frac{1-\alpha}{2}cos(\frac{4\pi n}{N-1})$$

La valeur optimale de α [Drugman *et al.*, 2009b] trouvée pour une décomposition à partir d'une fenêtre de taille $2T_0$ est $\alpha = 0.7$.

Plage d'estimation et critère de forme

Dans un premier temps, les plages de recherche de F_g et A sont limitées. F_g ne peut être estimé que dans une plage comprise entre $\frac{F_0}{2}$ et $3F_0$, et A normalisé par la période T_0 ne peut être trouvé que pour des valeurs allant de 0 à 0.4.

Dans un deuxième temps, il est nécessaire de régler le problème du déplacement de zéro pour une estimation convenable du débit glottique par ZZT. En effet, certains zéros sur la fréquence zéro (purement réels) ne sont pas estimés du bon coté du cercle unité. Il est proposé [Bozkurt *et al.*, 2005] de déplacer le zéro le plus proche du cercle unité du côté anticausal (à l'extérieur du cercle unité). Cette méthode semble un peu drastique et il est proposé ici de modifier le débit glottique estimé en fonction de sa forme.

Un débit glottique convenablement estimé comporte 4 points caractéristiques : le début et la fin du vecteur, ainsi que son maximum et son minimum. Un zéro manquant proche de 1 sur l'axe réel dans la transformée en Z du débit glottique revient à une intégration en trop de sa forme alors que un zéro supplémentaire sur cet axe revient à une dérivation en trop. En fonction du niveau du maximum par rapport aux premier et dernier points du débit glottique estimé par ZZT, on propose au choix, de le dériver ou l'intégrer une fois afin de corriger l'absence d'un zéro ou la présence d'un zéro supplémentaire proche de 1 par l'utilisation des équations 5.10 et 5.11 pour la dérivation ou l'intégration du vecteur g de taille N respectivement.

$$\tilde{g}(n) = g(n) - g(n-1) \tag{5.10}$$

$$\tilde{g}(n) = \tilde{g}(n-1) + g(n) \tag{5.11}$$

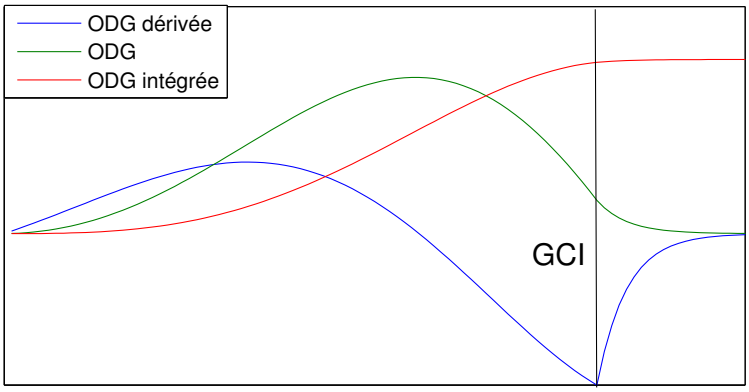


Figure 5.10 – Trois cas possibles pour l'estimation de l'ODG par ZZT, en fonction de la forme, il faut dériver ou intégrer le signal pour retrouver une ODG conforme au modèle LF.

Ainsi, d'après les observations de la figure 5.10 si $min(g) < \frac{g(1)+g(N)}{2}$, alors g est en fait la dérivée de l'ODG et il faut intégrer le signal. Si $max(g) <= g(N)$ alors g est en fait l'intégrale de l'ODG et il faut dériver le signal une fois. Ce critère de forme est nécessaire pour vérifier les équations 5.7 et 5.5.

Précision de l'estimation

L'estimation de F_g se fait sur un spectre de phase composé du nombre de descripteurs égal au nombre N de points définissant une période du signal glottique. La résolution de ce spectre est donc de $\frac{F_s}{N}$. Le spectre en question n'est censé contenir qu'une seule résonance et le calcul de TF par complément de zéros permet d'augmenter la précision de l'estimation. Une précision à 1Hz près est choisie.

L'estimation de $A = T_e - T_p$ se fait à deux échantillons près (précision sur le GCI et précision sur le maximum de l'ODG), c'est à dire $\frac{2}{F_s}$. La présence fréquente d'oscillations causées par les intéractions source/filtre ou par une mauvaise décomposition source/filtre n'a pas motivé une estimation plus précise de ce maximum.

Tant O_q que α_m sont déterminés par l'algorithme présenté ici avec une précision de l'ordre de $\frac{3}{F_s}$, soit environ 0.2ms pour une fréquence d'échantillonnage de 16kHz. Comme ces deux paramètres sont définis par rapport à la période du signal, la mesure présentée est donc plus précise pour les basses fréquences, mais aussi les valeurs plus élevées de quotient ouvert. Notons que dans ce cas la mesure de l'instant d'ouverture glottique est réalisée avec une erreur de 0.25ms [Bouzid et Ellouze, 2007], une précision similaire à notre algorithme mais présente à la fois sur l'instant d'ouverture et l'instant de fermeture ; Ceci qui génère une erreur potentielle totale sur T_e de 0.5ms.

5.3.5 Algorithme proposé

Afin d'appliquer cet algorithme il faut disposer de certaines informations *a priori* sur le signal :

- la position des instants de fermeture glottique. La méthode LoMA vue au chapitre 3 est préférée, la pratique ayant montré une meilleure compatibilité avec les besoins de la ZZT

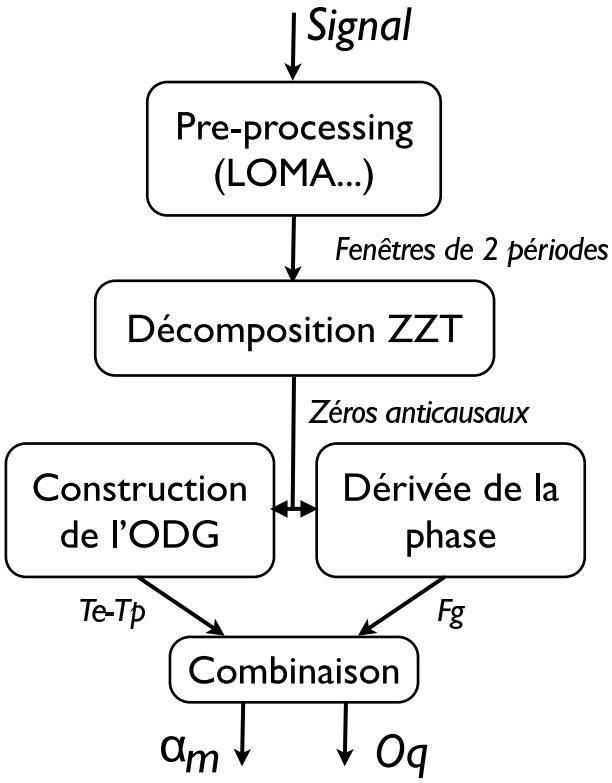


Figure 5.11 – Algorithme schématique de la méthode proposée pour l'estimation conjointe O_q-α_m.

que les autres méthodes (notamment DYPSA).
- le diamètre ρ du cercle sur lequel calculer le spectre pour en déduire le retard de groupe.

Un schéma de l'algorithme est donné en figure 5.11, il se décompose en quatre étapes :
- Tout d'abord, les étapes de traitement amont permettent de déterminer les instants de fermeture glottique, mais aussi d'appliquer une décomposition périodique-apériodique si elle se révèle nécessaire.
- L'algorithme ZZT est ensuite appliqué sur des fenêtres longues de 2 périodes pour y estimer les zéros de la partie anti-causale qui décrivent la phase ouverte du débit glottique. Les zéros anti-causaux sont ensuite utilisés pour calculer d'une part la dérivée de la phase sur le cercle complexe de diamètre ρ et reconstruire l'ODG par transformée inverse. Á cette étape, et pour des résultats similaires, il est possible d'utiliser la décomposition par cepstre complexe [Drugman *et al.*, 2008]. A ce niveau peut intervenir une étape de correction de la forme de l'ODG.
- La fréquence du formant glottique est déterminée sur la dérivée de la phase de la partie anticausale ; le maximum de l'ODG est détecté.
- Les deux valeurs précédemment déterminées sont combinées pour donner une valeur de O_q et une valeur de α_m.

5.4 Mesures préliminaires

Dans un premier temps, une petite base de données de signaux de parole naturelle est proposée pour tester l'algorithme. Les résultats de l'estimation sur cette base serviront à optimiser le comportement de l'algorithme en rapport avec les problèmes liés aux signaux naturels : ajustement de paramètres, instabilités des mesures.

5.4.1 Voyelles tenues

La base de données de voyelles tenues est composée de 18 échantillons prononcés par deux locuteurs - un homme, une femme - sur une variation de 3 voyelles et 3 qualités vocales différentes enregistrées avec un microphone statique. Un signal EGG synchrone a aussi été enregistré. Les voyelles sont tenues pendant 1 à 5 secondes selon le cas, les signaux EGG sont clairs et présentent peu de dédoublements de pics. La qualité vocale a été étiquetée à l'écoute, en tenant compte du style du locuteur.

5.4.2 Résultats de l'estimation

L'évaluation se fait sur la comparaison directe entre le quotient ouvert estimé par EGG et le quotient ouvert estimé sur le signal. Les valeurs sont comparées deux à deux. Une bonne détection est comptée lorsque l'erreur de mesure sur $\tilde{O_q}$ mesurée sur le signal par rapport à O_q mesuré sur l'EGG est inférieure au critère d'erreur fixé :

$$E = \frac{|\tilde{O_q} - O_q|}{O_q} \tag{5.12}$$

L'algorithme a été appliqué à chacun des 18 échantillons de qualité vocale. Les résultats présentés dans la table 5.2 donnent les informations sur l'échantillon (nom - voyelle - qualité vocale - fréquence fondamentale moyenne) ainsi que la valeur moyenne de O_q mesurée par EGG. On retrouve ensuite la valeur moyenne $\hat{O_q}$ estimée par la méthode à base de ZZT et le taux de détection dans les deux plages d'erreur de 17% (le seuil différentiel perceptif - JND - mesuré

sur le quotient ouvert [Henrich *et al.*, 2003]) et de 25% (précision minimum pour la séparation de trois qualités vocales cardinales). La valeur de α_m est la moyenne des valeurs couplées aux mesures de O_q détectées à l'intérieur de la plage d'erreur du JND. Enfin, le rayon du cercle du plan complexe pour le calcul de la dérivée de la phase s'il n'est pas égal à la valeur par défaut de $\rho = 0.98$.

La détection est très bonne (erreur inférieure au JND) dans plus de 60% des cas, et bonne (erreur inférieure à 25%) dans plus de 80% des cas. Même dans les cas pour lesquels les taux de détection sont bas, on remarque que la moyenne de l'estimation ($\hat{O}_q$) reste proche de la moyenne mesurée par EGG. Pour le locuteur féminin, les résultats sont moins bons.

Table 5.2 – Résultats de l'analyse sur la base de données de voyelles expressives. Deux locuteurs pour trois voyelles et trois expressions. O_q : estimé par la méthode, $\hat{O}_q$ mesuré sur l'EGG, α_m est donné comme la moyenne des estimations appariées avec un O_q dans le JND.

Sample	Voyelle	Qualité	F_0	O_q	$\hat{O}_q$	<JND	<25%	α_m	
M1	/a/	normale	127Hz	0.61	0.61	94%	95 %	0.62	
M2	/i/	normale	130Hz	0.50	0.49	66%	76%	0.60	$\rho = 0.92$
M3	/u/	normale	127Hz	0.52	0.50	85%	97%	0.60	
M4	/a/	tendue	131Hz	0.46	0.39	43%	86%	0.67	
M5	/i/	tendue	131Hz	0.41	0.39	90%	97%	0.62	
M6	/u/	tendue	128Hz	0.51	0.38	3%	13%	0.73	
M7	/a/	relâchée	123Hz	0.71	0.75	69%	82%	0.70	
M8	/i/	relâchée	130Hz	0.79	0.68	53%	74%	0.64	
M9	/u/	relâchée	128Hz	0.71	0.67	85%	88%	0.53	$\rho = 0.999$
F1	/a/	normale	235Hz	0.47	0.44	42%	70%	0.68	
F2	/i/	normale	250Hz	0.39	0.48	20%	31 %	0.90	
F3	/u/	normale	238Hz	0.47	0.49	65%	71%	0.77	$\rho = 1.05$
F4	/a/	tendue	238Hz	0.42	0.42	69%	80%	0.67	$\rho = 0.95$
F5	/i/	tendue	239Hz	0.34	0.35	50%	74%	0.87	
F6	/u/	tendue	246Hz	0.34	0.29	58%	79%	0.66	
F7	/a/	relâchée	242Hz	0.71	0.72	88%	93%	0.68	$\rho = 0.95$
F8	/i/	relâchée	250Hz	0.60	0.68	58%	74%	0.91	$\rho = 0.91$
F9	/u/	relâchée	242Hz	0.66	0.68	32%	47%	0.78	

5.4.3 Discussion

D'une manière générale, les résultats sont de bonne qualité, avec peu ou pas d'erreur. Les valeurs de α_m varient peu sur la base d'échantillons choisis. C'est un résultat attendu étant donné que l'asymétrie joue un rôle secondaire dans la dimension de tension de la voix et joue un rôle de plus grande importance pour l'aspect d'effort vocal. Les résultats obtenus [C.Sapienza *et al.*, 1998] ont montré que le quotient de vitesse - qui varie avec l'asymétrie - est fortement lié à cet aspect d'effort vocal. Discutons les résultats en fonction du locuteur :

Locuteur masculin

Les résultats sont meilleurs pour le locuteur masculin, et excellents dans le cas d'une voix modale et de la voyelle /a/ - cas classique d'analyse présenté en figure 5.5. Un seul échantillon est réellement problématique (M6) dans le sens où aucun réglage de ρ n'a permis d'obtenir des résultats exploitables lors de son analyse. L'asymétrie reste inférieure à 0.70 sauf dans deux cas :

M6 et M7. Il est probable que le locuteur ait eu besoin de produire un effort plus important pour produire ces voyelles avec ces qualités vocales. En moyenne, dans plus de 85% des cas sur la totalité des échantillons du locuteur masculin, le quotient ouvert est détecté convenablement avec une erreur inférieure à 25% par rapport à la référence EGG.

Locuteur féminin

Dans le cas du locuteur féminin, les performances sont moins bonnes, probablement en raison de la fréquence fondamentale plus élevée, diminuant de facto la précision de la mesure de O_q et α_m. Deux échantillons posent problème (F2 et dans une moindre mesure F9), pour lesquels aucun ajustement de la valeur ρ n'a permis d'améliorer les résultats assez mauvais avec moins de 30% de valeurs détectées sous le seuil différentiel perceptif. Dans le cas de F9, la valeur moyenne du quotient ouvert mesuré par le signal est tout de même proche de la mesure sur l'EGG. Les valeurs de l'asymétrie sont plus dispersées que pour le locuteur masculin allant de 0.66 à 0.91. On retrouve systématiquement des fortes valeurs de α_m pour la voyelle /i/. En moyenne, dans plus de 75% des cas sur la totalité des échantillons du locuteur féminin, le quotient ouvert est détecté convenablement avec une erreur inférieure à 25% par rapport à la référence EGG.

Asymétrie

La mesure de l'asymétrie est plus problématique. Dans le cas de l'algorithme proposé, α_m est estimé conjointement avec O_q mais l'impact de l'erreur de mesure sur A et F_g est multiplicatif dans le cas de l'asymétrie. Pour être valides, les mesures de α_m nécessitent donc une précision plus importante. Et pour être sous le JND, elles nécessitent par conséquent une estimation de O_q inférieure au JND.

On peut visualiser la relation de l'erreur d'estimation sur O_q et α_m sur la figure 5.12. Cette figure présente les résultats de l'analyse pour un signal à qualité vocale variable. Á la fois l'effort (doux - fort - doux) et la tension (lâche - tendu - lâche) varient, l'asymétrie (ligne du haut) et le quotient ouvert (estimé en noir, référence en rouge - ligne du milieu) sont estimés par la méthode présentée. On remarque que les valeurs de α_m qui correspondent le mieux aux attentes et aux travaux précédents dans le domaine (i.e. : que α_m augmente avec l'effort) sont appairées aux valeurs de $\tilde{O}_q$ présentant l'erreur la plus faible entre 0 et 1s et 5 et 7 secondes.

Entre 1 et 2 secondes sur la figure 5.12, la différence entre O_q et $\tilde{O}_q$ est d'environ 0.06-0.07, soit proche du JND, alors que l'asymétrie tombe brutalement de 0.7 à 0.6. Lorsque l'effort augmente, on s'attend à une augmentation de α_m et donc à un déplacement du formant glottique par rapport à la fréquence de modulation de pulsation ω_g. Ce déplacement par rapport à la prévision des équations 5.7 et 5.5 rend l'estimation moins précise, voire invalide.

Compte tenu des observations sur la figure 5.12, le choix du JND comme critère d'erreur minimum sur O_q n'est pas suffisant pour valider le couple de valeurs (O_q, α_m). L'introduction d'un troisième critère de précision de 5% permettrait la validation des estimations sur α_m. Ce critère n'a pas été mis en œuvre dans la mesure où il n'existe pas de manière de vérifier une précision inférieure au JND sur α_m.

5.5 Protocole d'analyse sur signaux naturels

5.5.1 Adaptation de l'algorithme aux signaux naturels

L'analyse du petit corpus expressif précédent donne des informations sur l'applicabilité de la méthode aux signaux naturels. Dans l'optique d'une analyse automatique, il faut dans un premier

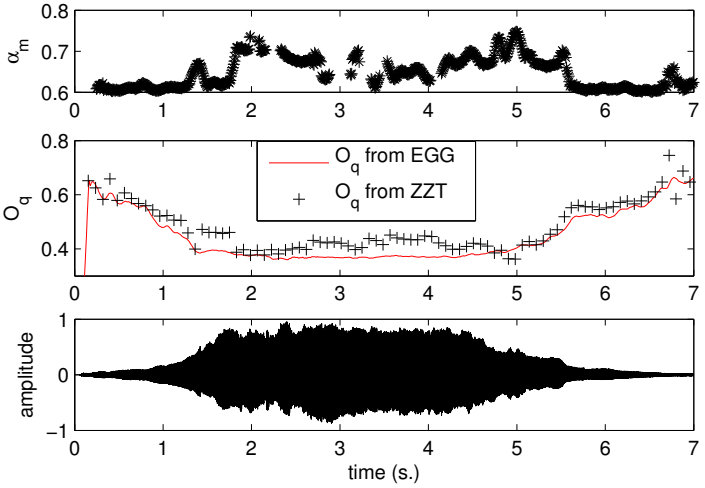


Figure 5.12 – Estimation combinée O_q-α_m sur un échantillon à qualité vocale variable. Si la mesure de O_q sur le signal suit bien les données EGG, il est important de sélectionner les valeurs de l'asymétrie couplées avec une erreur minimum de O_q sous peine de mal interpréter la mesure.

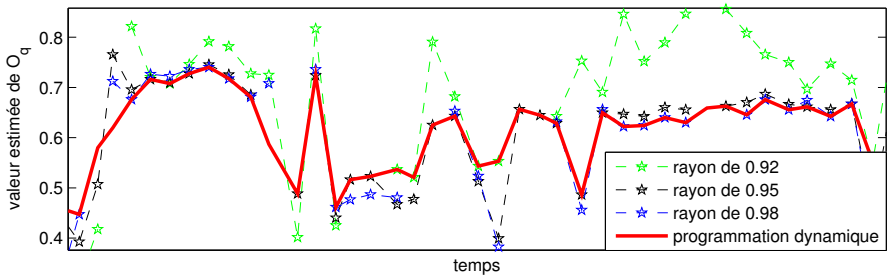


Figure 5.13 – Programmation dynamique pour sélectionner la valeur de O_q la plus cohérente en fonction des valeurs précédentes. Le chemin continu est le chemin choisi par l'algorithme. Certains chemin en pointillés sont coupés, car les valeurs estimées ne sont pas toujours réalistes (dans ce cas, elles sont mises à NaN).

temps se libérer de la contrainte du paramètre ρ et augmenter l'immunité aux instabilités de la décomposition. La finesse de l'estimation du paramètre α_m est aussi à prendre en considération. Ainsi, certaines étapes sont ajoutées à la méthode précédente :

- Dans un premier temps, on cherche à s'affranchir de la sensibilité de la mesure de F_g au rayon ρ choisi pour le calcul de la dérivée de la phase de la partie anticausale. La mesure est directement réalisée pour 3 valeurs de ρ : 0.92, 0.95 et 0.98. Le choix se fait par la suite avec une programmation dynamique (illustrée sur la figure 5.13), cherchant à minimiser le chemin parcouru dans le temps au gré des mesures de O_q. Une valeur non appropriée de ρ donnera des valeurs de formant glottique variant fortement d'une période à une autre et donc moins susceptibles d'être retenues lors de la minimisation du chemin.
- Dans un deuxième temps, il convient de tenir compte de l'instabilité de la décomposition ZZT du point de vue de la reconstruction du débit glottique. Certaines trames peuvent causer des instabilités en haute fréquence qui vont corrompre la mesure de $A = T_e - T_p$. Pour contourner ce problème, on applique un filtre moyenneur de gain 1 et de largeur 10

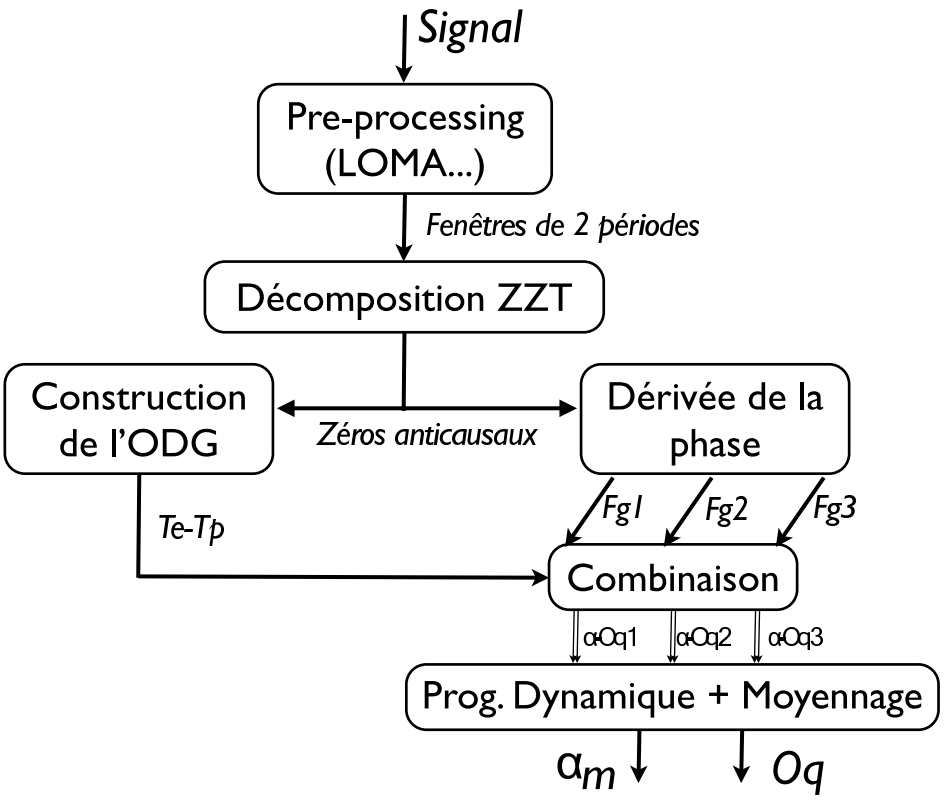


Figure 5.14 – Algorithme modifié pour la méthode proposée. La différence principale réside dans la mesure de 3 fréquences différentes de F_g (une pour chaque valeur de ρ, c.f. texte) qui donnent chacune un couple O_q-α_m. Les valeurs sont ensuite sélectionnées par programmation dynamique puis moyennées.

aux valeurs de O_q après programmation dynamique.

– Enfin, dans le cas de l'exploitation des mesures de α_m, seuls les couples $O_q - \alpha_m$ dont $\tilde{O}_q$ est mesuré dans une plage de 5% d'erreur par rapport à l'EGG sont retenus. Dans une première approche de cette analyse on donne les valeurs de α_m non vérifiées par EGG. (figure 5.14).

L'algorithme modifié est présenté en figure 5.14. On remarque que chaque valeur de ρ retenue engendre un couple O_q-α_m. Le choix se fait ensuite par programmation dynamique et moyennage.

5.5.2 Résultat de l'algorithme sur de la voix parlée

La base de données (appelée Base A pour éviter les confusions) utilisée est la même que pour le test de l'algorithme LoMA sur l'estimation des GCI. Une base de données, de français lu à partir d'article de journaux regroupant deux locuteurs différents. Les instants d'ouverture et de fermeture sont clairement identifiés sur les signaux EGG enregistrés en parallèle.

Ces résultats sont présentés sur la figure 5.15 par le biais de trois histogrammes superposés. La détection de O_q dans les valeurs associées à la voix modale est très bonne.

– L'histogramme blanc donne la dispersion des valeurs de O_q sources, mesurées sur le signal EGG.
– L'histogramme gris donne le nombre de valeurs détectées avec une erreur inférieure au JND pour chaque plage de O_q. En moyenne, le taux de détection sous cette erreur est de 80%

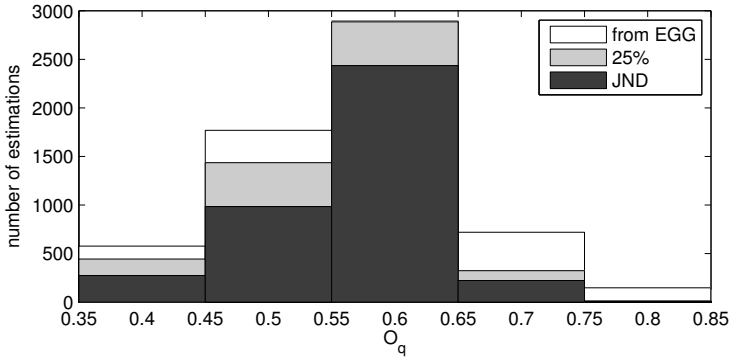


Figure 5.15 – Résultat de l'estimation du quotient ouvert sur la base de données A de voix parlée. Les résultats sont présentés sous forme d'histogrammes superposés : valeurs mesurées sur l'EGG (blanc), du nombre de détections pour chaque valeur dans chaque plage d'erreur précédemment décidée : le JND (noir) et 25% (gris).

pour une voix modale (O_q moyen), de 50% pour un quotient ouvert bas et de 20% pour un quotient ouvert élevé.

– L'histogramme noir donne le nombre de valeurs détectées avec une erreur inférieure à 25% pour chaque plage de O_q. En moyenne, le taux de détection sous cette erreur est de 97% pour une voix modale, de 80% pour un quotient ouvert bas et de 25% pour un quotient ouvert élevé.

5.5.3 Discussion

Le choix d'une base de données est critique pour évaluer la performance d'un algorithme d'estimation de quotient ouvert. La voix modale présente généralement un O_q aux environs de 0.66 [Klatt et Klatt, 1990] dans 60 à 70% des cas. Ainsi, un estimateur naïf qui donnerait $O_q = 0.6$ 100% du temps présenterait déjà un score brut très élevé sur l'ensemble de la base de données.

Des études comme [Bouzid et Ellouze, 2007, Thomas *et al.*, 2009, Degottex *et al.*, 2010], par exemple, ne tiennent pas compte de la répartition des valeurs de quotient ouvert au sein de la base de données. Les résultats présentés peuvent être biaisés par un trop grand nombre de valeurs autour du point de fonctionnement optimal de la méthode. Le parti a donc été pris de présenter des résultats détaillés en fonction de la valeur source de O_q et de critiquer les performances d'estimation en matière de détection dans une plage d'erreur autour de la valeur de O_q. Ces résultats sont présentés sur la figure 5.15. On y retrouve l'analyse des détections de quotient ouvert pour les deux seuils d'erreur (JND et 25%), en fonction de la répartition des valeurs au sein de la base de données, permettant de critiquer l'algorithme sur sa performance en fonction de la valeur même du quotient ouvert.

Les résultats de l'estimation ne sont pas constants sur toute la plage des valeurs prises par O_q. Les plus basses valeurs sont mieux détectées que les fortes valeurs. Ceci est probablement expliqué par une tendance de l'algorithme à sous-estimer les valeurs de O_q. La plage réelle de l'erreur augmente avec la valeur de O_q, le protocole utilisé devrait donc naturellement compenser cette tendance. Mais il semble que l'impact d'une analyse sur deux périodes, couplé avec la très faible fréquence du formant glottique dans le cas d'une grande valeur de O_q entrainent une distorsion dans cette plage d'estimation. En effet, dans ce cas là le formant glottique se trouve

être inférieur au premier harmonique. Sa définition spectrale (par les zéros anticausaux de la ZZT) est donc difficile et limitée par la résolution et l'instabilité de l'analyse. Une autre explication pourrait venir d'un bruit aléatoire créé dans le conduit vocal, bruit connu pour altérer de manière significative les performances de la décomposition par ZZT. Une grande valeur de quotient ouvert est généralement associée à un bruit de phonation élevé. Dans ce cas, une décomposition préalable du type périodique/apériodique pourrait se révéler adéquate. Ce point sera abordé par la suite.

Sur la plage propre à la voix modale (avec O_q entre 0.55 et 0.65 sur la figure 5.15) les performances sont très satisfaisantes : plus de 80% des quotients ouverts sont estimés avec une erreur inférieure aux 17% du seuil différentiel. Cela correspond à la configuration glottique qui répond généralement le mieux à ce type d'estimation, le formant glottique est alors placé entre le premier et le deuxième harmonique, permettant sa résolution dans le domaine spectral.

Sur la plage de voix serrée, pour laquelle O_q est compris entre 0.35 et 0.55, la quantité de détection décroît un peu mais reste importante, avec plus de 80% de détection dans une plage de 25% d'erreur.

5.5.4 Conclusion

Une méthode pour estimer conjointement les valeurs de O_q et α_m sur des signaux naturels a été présentée. Cette méthode utilise la dérivée de la phase calculée sur un cercle dans le plan complexe et la position du maximum de l'onde de débit glottique, estimée par ZZT. Sur un éventail de voyelles tenues de différente qualité vocale, l'algorithme a donné d'excellents résultats mais a aussi montré quelques faiblesses. La méthode a donc été ajustée pour tenir compte de ces faiblesses et a été appliqué sur un ensemble de signaux naturels produits par deux locuteurs (homme et femme). Les résultats ainsi obtenus sont présentés en fonction de la valeur de O_q mesurée sur EGG, et par taux de détection pour un critère d'erreur donné : erreur d'estimation inférieure à 17% (seuil différentiel) ou inférieure à 25%.

Les résultats montrent que cette méthode se révèle efficace mais présente des erreur de détections importantes pour les fortes valeurs de quotient ouvert. Une explication possible serait que ces lacunes sont causées par un bruit de phonation trop élevé.

5.6 Méthode hybride combinant ZZT et LoMA pour l'estimation du quotient ouvert

5.6.1 Principe

Deux méthodes ont été présentées pour l'estimation du quotient ouvert. La méthode présentée dans ce chapitre, basée sur une estimation par ZZT, est la méthode basée sur le décalage du premier harmonique par rapport au GCI (décalage mesuré sur la LoMA et présenté au chapitre 3). Les deux méthodes présentent des forces et des faiblesses complémentaires :

- La méthode par ZZT présente des faiblesses dans les fortes valeurs de O_q, probablement causées par la présence de bruit. Ainsi, on attend de meilleures performances pour de fortes valeurs de quotient ouvert après une décomposition PAP-A comme celle présentée au chapitre 4.
- L'analyse par ondelettes, quant à elle, tend à mieux estimer les faibles valeurs de quotient ouvert, mais aussi à détecter des valeurs de quotient ouvert sur des portions non stationnaires du signal.

Une solution consiste donc à combiner ces méthodes par une pondération. Pour ce faire, les deux méthodes seront appliquées à une base de données de signaux de parole, plus importante que

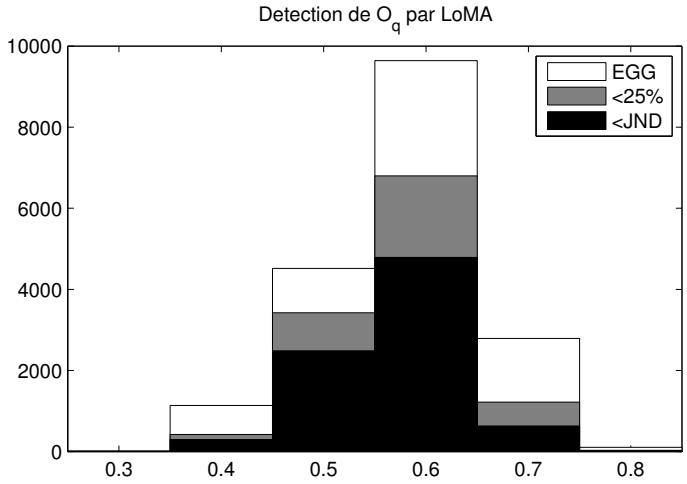


Figure 5.16 – Analyse de la base de données B par la méthode LoMA, mesure du quotient ouvert et distribution des détections pour les deux seuils.

celle utilisée dans l'étude précédente. Une pondération sera proposée et les résultats combinés seront analysés.

5.6.2 Résultats de chaque méthode

On applique chaque méthode à une base de données contenant deux locuteurs comme en 5.5.2, mais sur un nombre plus important de fichiers et notamment des fichiers auparavant exclus du fait de leur difficulté d'analyse. Cette base de données est appelée Base B pour éviter toute confusion. Au total, 18274 périodes ont permis de mesurer autant de valeurs du quotient ouvert et instants de fermeture glottique sur les signaux EGG synchrones. Tout comme précédemment, les résultats présentés sont les distributions de bonnes détections pour les deux seuils retenus : le JND et 25% d'erreur avec le même protocole pour la décision d'une bonne détection et pour la présentation des données.

La figure 5.16 présente la distribution de l'analyse de la base de données par la méthode LoMA de mesure du quotient ouvert par le décalage du premier harmonique ; la formule appliquée est celle décrite lors de la présentation de la méthode dans le chapitre 3. Les résultats ne sont pas très bons pour les fortes valeurs de quotient ouvert, mais plutôt bons sur la tranche autour de 0.5.

Sur la figure 5.17 est présentée la distribution de l'analyse de la base de données par la méthode ZZT de ce chapitre avec décomposition PAP-A préalable. Les résultats sont moins bons que pour une analyse sans décomposition PAP-A de la figure 5.15, principalement du fait de l'utilisation d'une base de donnée plus complète (Base B). Cependant, les résultats sont meilleurs pour des fortes valeurs de O_q par rapport à la méthode à base de LoMA.

5.6.3 Pondération

Á la lumière des résultats des figures 5.16 et 5.17, on remarque que les méthodes ne possèdent pas les même forces et faiblesses. On choisit de manière empirique, en fonction des résultats donnés par chaque méthode, la pondération proposée sur la figure 5.18 et d'équation 5.13 afin de pouvoir combiner les résultats des deux méthodes, notés Oq_Z et Oq_L pour respectivement l'estimation par ZZT et celle par LoMA et donner donc une valeur combinée de O_q.

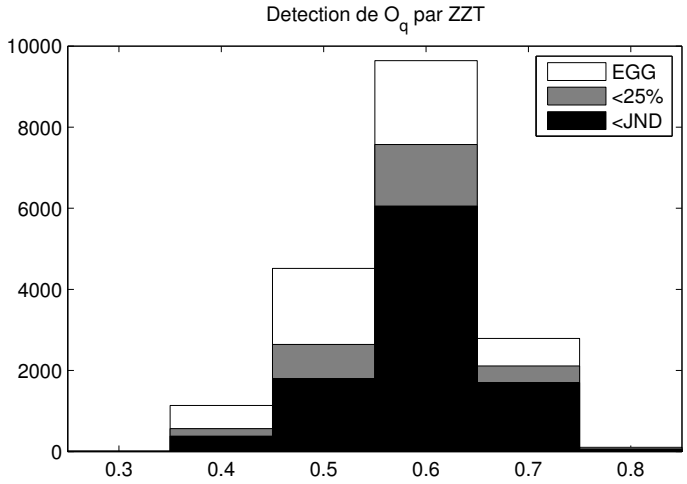


Figure 5.17 – Analyse de la base de données B par la méthode ZZT + PAP-A : mesure du quotient ouvert et distribution des détections pour les deux seuils.

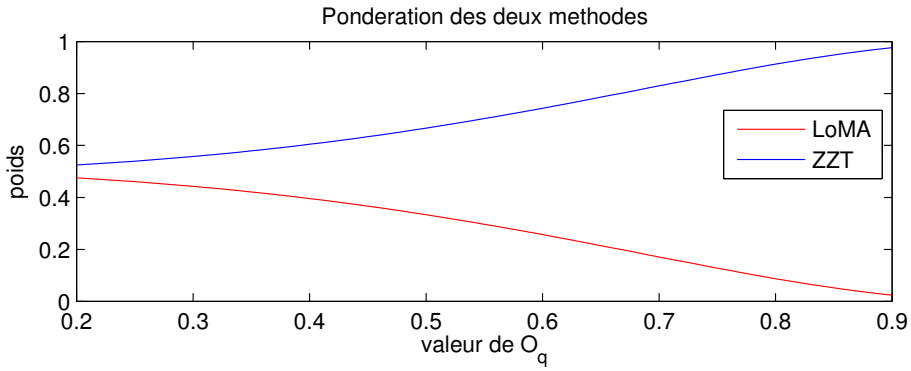


Figure 5.18 – Pondération arbitraire des deux méthodes pour favoriser les résultats obtenus par ZZT sur les O_q élevés.

Cette pondération arbitraire cherche à favoriser les valeurs hautes données par ZZT sur la partie voisée des signaux (les meilleurs résultats) par rapport au LoMA alors que pour les valeurs basses on donne un poids égal aux deux méthodes qui présentent des performances cohérentes sur cette plage.

$$O_q = \frac{Oq_Z + Oq_L cos(Oq_L \pi/2)^2}{1 + cos(Oq_L \pi/2)^2} \tag{5.13}$$

Les résultats de l'analyse de la base de données après combinaison des deux méthodes sont fournis sur la figure 5.19, on remarque alors qu'ils sont bien meilleurs que pour les méthodes seules. On arrive à 67% de détections à l'intérieur du JND et 83% de détections sous 25% d'erreur. De plus, comme deux résultats sont utilisés pour générer une seule valeur de O_q, on peut exploiter le surplus d'information pour déterminer un facteur de certitude sur la qualité finale de l'estimation.

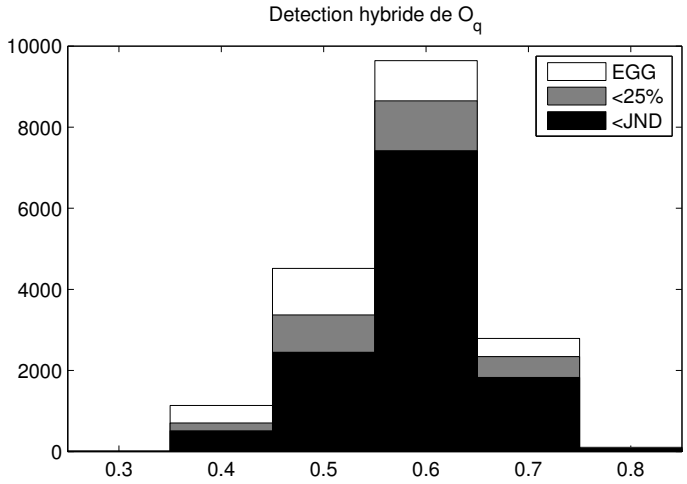


Figure 5.19 – Analyse de la base de données B par la méthode hybride LoMA + ZZT.

5.6.4 Facteur de certitude

Examinons la différence entre les deux valeurs de O_q estimées, par LoMA et par ZZT. Soit e cette valeur, alors $e = |Oq_{LoMA} - Oq_{zzt}|$.

Dans 77% des cas, on observe une différence d'estimation e inférieure à 0.12. Parmi ces estimations pour lesquelles la différence est inférieure à 0.12, 87% présentent une valeur pondérée de O_q estimée avec une erreur inférieure au JND. On peut donc en conclure le facteur de certitude suivant :

Si la différence d'estimation e est inférieure à 0.12, la valeur estimée est proche de la valeur réelle (erreur inférieure au JND) dans 87% des cas.

5.6.5 Discussion et conclusion

Dans cette section il est proposé une hybridation de deux méthodes d'estimation du quotient ouvert. Chacune des deux méthodes, estimation par ZZT après décomposition PAP-A et estimation par décalage du premier harmonique de la LoMA, présente des forces et des faiblesses. Une procédure de pondération des deux résultats permet de les combiner pour arriver à un niveau de détection supérieur à 66% pour une erreur inférieure au JND sur la Base B. Un facteur de confiance dans l'estimation a été proposé : dans plus de 87% des cas, si la différence d'estimation entre les deux méthodes est inférieure à 0.12, alors la valeur de O_q estimée par combinaison présente une erreur inférieure au JND.

Pour appuyer l'intérêt de cette pondération, des tests complémentaires ont été effectués sur deux bases de données supplémentaires issues d'enregistrements réalisés pour VOCQUAL'03. Ces bases sont indépendantes des Bases A et B en terme de langage (ici l'anglais est utilisé) et en terme de locuteur.

La première base repose sur une phrase courte (*"she has left for a great party today"*) prononcée de manière répétitive et très expressive (cirée, chuchotée, en fry, etc) par un seul locuteur : cette base totalise 75 occurrences de cette phrase (avec EGG synchrone) pour un total de 4978 valeurs de O_q estimées sur EGG et est appelée *brian*. Les résultats de l'estimation par LoMA, par ZZT (+PAP-A) et par hybridation sont donnés sur la figure 5.20. Pour LoMA on retrouve 35% et 24% de détection pour une erreur respectivement inférieur au JND ou au seuil de 25%. Pour ZZT on retrouve 36% et 25% de détection pour une erreur respectivement inférieur au JND

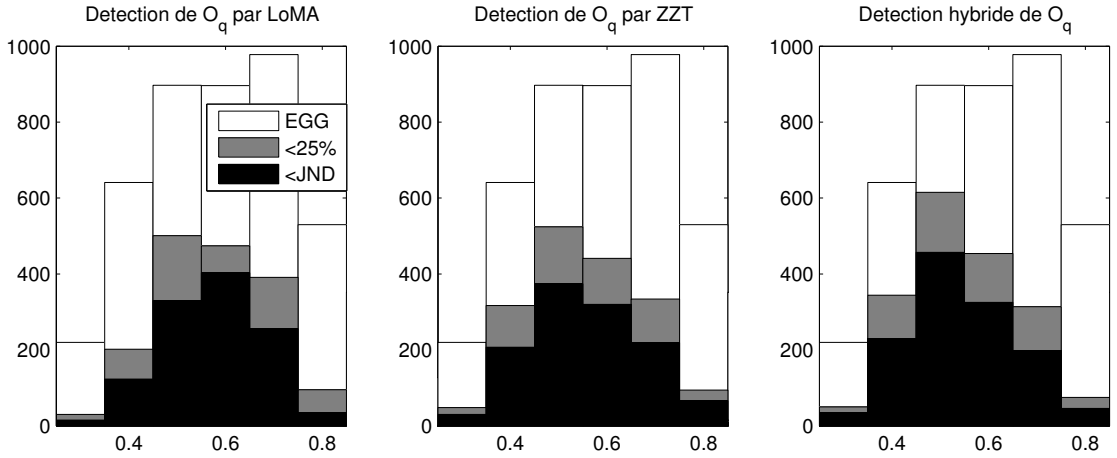


Figure 5.20 – Estimation du quotient ouvert par LoMA, par ZZT (+PAP-A) et par hybridation des deux méthodes sur la base de donnée "brian" de VOCQUAL'03.

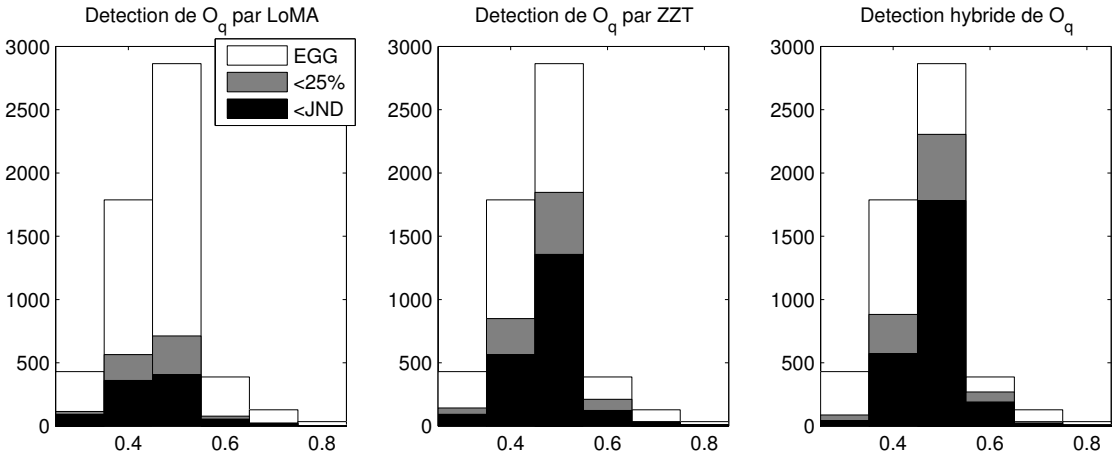


Figure 5.21 – Estimation du quotient ouvert par LoMA, par ZZT (+PAP-A) et par hybridation des deux méthodes sur la base de donnée "singing" (fichiers JF-mem-6-A et LP-mem-6-a) de VOCQUAL'03.

ou au seuil de 25%. Pour l'hybridation on retrouve 38% et 26% de détection pour une erreur respectivement inférieur au JND ou au seuil de 25%. Dans 30% des cas, une différence d'estimation inférieur à 0.1 a donné une valeur détectée par hybridation avec une erreur inférieur au JND.

Le deuxième base est un regroupement d'une séquence de 16 secondes du refrain de la chanson *Memory* de la comédie musicale *Catz* chantée par un chanteur puis une chanteuse, totalisant 5697 estimations du quotient ouvert sur EGG. Les résultats de l'estimation par LoMA, par ZZT (+PAP-A) et par hybridation sont donnés sur la figure 5.21. Pour LoMA on retrouve 25% et 18% de détection pour une erreur respectivement inférieur au JND ou au seuil de 25%. Pour ZZT on retrouve 55% et 40% de détection pour une erreur respectivement inférieur au JND ou au seuil de 25%. Pour l'hybridation on retrouve 60% et 43% de détection pour une erreur respectivement inférieur au JND ou au seuil de 25%. Dans 60% des cas, une différence d'estimation inférieur à 0.1 a donné une valeur détectée par hybridation avec une erreur inférieur au JND.

Les résultats obtenus sont moins bons que pour la base de données B mais on remarque que l'hybridation entre les méthodes fournit toujours des résultats de meilleur qualité que pour chaque méthode seule. La fonction de pondération n'est pas optimale pour toute les bases de données, mais étant basée sur les propriétés analytiques des méthodes utilisées, elle permet d'en tirer facilement les forces et faiblesses. Sur ces analyses complémentaires, le facteur de certitude proposé d'après l'analyse de la base de données B ne s'est pas révélé aussi pertinent.

5.7 Conclusion

Un aspect important de la qualité vocale est l'estimation des paramètres de la source glottique. Pour réaliser une telle estimation, une méthode hybride est proposée tenant compte des travaux des chapitres 3 et 4.

Dans un premier temps, un protocole de test du filtrage inverse par ZZT a permis de comparer deux modélisations de la production vocale. D'un côté, le modèle causal-anticausal basé sur la connaissance *a priori* des GCI et de l'autre le modèle autorégressif imposant la forme de la réponse impulsionnelle du conduit vocal. Ce test a permis de valider la viabilité du modèle sur lequel se base la décomposition ZZT, en présentant des résultats comparables aux méthodes habituelles de filtrage inverse. Des exemples sur signaux réels ont montré qu'une forme non conventionnelle de débit glottique dérivé ne donnait pas nécessairement une estimation erronée.

Le test précédent valide l'utilisation de la ZZT comme méthode systématique de filtrage inverse, ce qui permet de proposer une une méthode pour la mesure du quotient ouvert et de l'asymétrie. Cette méthode combine la valeur de la fréquence du formant glottique mesurée sur la dérivée de la phase et le maximum de l'ODG obtenu par décomposition source/filtre, ici par ZZT. Une première version de l'algorithme a été testée sur des voyelles tenues prononcées avec différentes qualités vocales et pour deux locuteurs, en évaluant la quantité de bonnes détections du quotient ouvert. Aucune méthode existante ne permettant de déterminer avec précision l'asymétrie (bien que NAQ, entre autres, donne une idée de sa valeur [Alku et Bäckström, 2002]), l'étude s'est donc limitée à l'évaluation des performances sur un seul paramètre : le quotient ouvert O_q. Une valeur de quotient ouvert est considérée valide à partir du moment où l'erreur avec la valeur mesurée sur l'EGG est inférieure au seuil différentiel perceptif, mesuré à 17% [Henrich *et al.*, 2003]. Afin de représenter les résultats plus en détails, un deuxième seuil d'erreur de 25% a été proposé. Ce seuil correspond à la limite maximum pour différencier les valeurs de 0.3, 0.5 et 0.8 considérées comme les 3 principales qualités vocales extractibles par les informations du quotient ouvert : serré, normal, lâché. Les résultats ont été donnés en pourcentage de bonnes détections pour chaque échantillon de la base de données.

Les résultats de ce premier algorithme ont montré la dépendance des résultats au paramètre ρ, spécifiant le rayon de calcul de la dérivée de la phase. Pour compenser cet effet et garder l'aspect systématique de la méthode, un nouvel algorithme a été présenté. Trois différentes valeurs du formant glottique sont alors mesurées, une programmation dynamique et un moyennage permettent de sélectionner la combinaison de valeurs la plus cohérente. Ce deuxième algorithme a été testé sur une base de données de voix parlée, montrant une bonne capacité à estimer les valeurs moyennes et basses de O_q, mais une incapacité à estimer de manière convenable les valeurs élevées de ce paramètre.

En ce qui concerne l'asymétrie, les observations montrent que sa valeur est très sensible aux variations du quotient ouvert. Il est donc proposé de ne retenir comme estimation convenable du paramètre α_m que celles dont la valeur appairée de quotient ouvert est estimée avec une précision suffisante par rapport à la référence. Dans l'état actuel, l'estimation de l'asymétrie dépend donc encore d'une référence externe comme l'EGG, à l'instar des travaux précédents [Sturmel *et al.*, 2006].

Finalement, une méthode hybride basée sur la combinaison de l'estimation par ZZT et par LoMA a été proposée. Cette méthode présente des résultats plus cohérents sur l'ensemble de la plage de variation du quotient ouvert et permet en outre de déduire un facteur de confiance sur la qualité de l'estimation, des analyse complémentaires ont montré cependant que ce facteur de confiance n'était pas toujours pertinent. Le fait même que la mesure par ZZT et le décalage du premier harmonique observé sur la LoMA soient dépendants à la fois du quotient ouvert et de l'asymétrie encourage à continuer la recherche sur ce moyen d'estimation afin d'arriver in fine à une mesure combinée O_q-α_m à l'instar de ce qui a été proposé pour la ZZT seule.

Résumé

Problématique

Pour déterminer la qualité vocale d'un enregistrement de voix, il est utile de connaitre la valeur des paramètres du flux glottique associés à la production du signal. Une technique de filtrage inverse encore peu utilisée, la ZZT, repose sur un modèle original de la production glottique : le modèle causal-anticausal. Il s'agit donc de déterminer dans quelle mesure ce modèle est viable et de l'utiliser ensuite pour déterminer les paramètres du modèle de source.

Apport scientifique

Un protocole d'expérimentation a mis en valeur les capacités du modèle causal-anticausal à rivaliser avec le modèle autorégressif de la production vocale. Une méthode simple de mesure conjointe du quotient ouvert et de l'asymétrie a été présentée. Cette méthode se base sur la partie anticausale de la décomposition ZZT, estimée sur le signal de parole. Dans un premier temps, la fréquence du formant glottique est mesurée sur la dérivée de la phase calculée sur un cercle au rayon bien précis ; puis le maximum de l'ODG est détecté sur l'onde reconstituée à partir de la partie anticausale de la ZZT. Dans le but d'appliquer cette méthode à de la parole expressive, des tests préliminaires ont été réalisés sur un corpus simple, mais relativement exhaustif, de voyelles soutenues à différentes qualités vocales.
Après affinement de l'algorithme, dont notamment l'ajout d'une étape de traitement a posteriori, une base de données plus complète a servi à l'évaluation finale de la méthode. Les résultats sur cette base de données sont encourageants, mais montrent aussi les limites de la méthode pour certaines configurations glottiques bien précises.
Dans un second temps, une méthode hybride a été présentée, combinant à la fois LoMA et ZZT, donnant des résultats plus cohérents sur l'ensemble de la plage de variation du quotient ouvert. Cette méthode hybride a été testée sur trois bases de données différentes.

En conclusion

Les résultats de l'évaluation de cette méthode montrent que la méthode hybride d'estimation du quotient ouvert donne de bons résultats sur le corpus. Cette méthode est à la fois efficace pour estimer les faibles et fortes valeurs de O_q. Cependant, l'estimation du quotient d'asymétrie n'est pas encore assez précise pour être exploitée de manière systématique.

Troisième partie

Application à de la parole expressive

Chapitre 6

Analyse d'un grand corpus

Sommaire

6.1 Constitution de la base . . . 167
6.2 Analyse et Protocole . . . 168
6.3 Résultats . . . 169
6.3.1 Tendances générales par style . . . 171
6.3.2 Tendances par phonème voisé . . . 171
6.3.3 Analyse par phonème voisé . . . 174
En ce qui concerne le barycentre de la LoMA . . . 174
En ce qui concerne les valeurs de quotient ouvert . . . 175
En ce qui concerne les valeurs de jitter . . . 175
En ce qui concerne les valeurs du rapport harmonique sur bruit . 175
6.4 Confirmation des tendances par analyse statistique . . . 175
6.4.1 Indépendance des paramètres . . . 176
6.4.2 Classement hiérarchique paramètre par paramètre . . . 176
6.4.3 Classement hiérarchique par phonème . . . 177
6.5 Interactions source-filtre . . . 177
6.6 Corrélation entre les estimations . . . 177
6.6.1 Corrélation par style . . . 177
6.6.2 Corrélation par phonème . . . 181
6.7 Caractérisation des styles . . . 181
6.7.1 Commentaires style par style . . . 181
Contraste Bas . . . 181
Dialogue . . . 181
Dictée . . . 181
Didactique . . . 182
Enjoué . . . 182
Fort . . . 182
Grave . . . 182
Insiste . . . 182

Narratif calme 183
Narratif tendu 183
Point 183
Suspensif 183
Triste 183
Vieux 183
Virgule 183
Voix basse 183
6.7.2 Visualisation 184
6.7.3 Méta-classes de qualité vocale 184
6.8 Conclusion 184

Les trois chapitres précédents ont proposé de nouvelles approches pour l'analyse des signaux de parole naturelle. Il a été vu comment les segmenter, séparer les contributions périodiques et apériodiques et finalement y estimer les paramètres de source glottique. Chacun de ces axes peut se lier à une dimension de la qualité vocale (raucité, voisement, tension, force). Il s'agit donc d'explorer dans quelle mesure les quantités estimées par les méthodes proposées peuvent être liées au style d'élocution (i.e. : à de la voix expressive).

Pour ce faire, un corpus fournit par Orange Labs regroupant des phrases ou segments de phrases destinées à la synthèse vocale a été analysé. Ce corpus est intéressant car il regroupe un nombre important de styles (16) sur le même contenu linguistique, mais ne comporte qu'un seul locuteur. Ainsi, chaque style sera décrit en fonction des paramètres de qualité vocale estimés à partir des méthodes précédemment développées.

Après avoir décrit la base de données et le protocole d'analyse, ce chapitre présentera les résultats obtenus sur cette base. Des représentations statistiques seront utilisées (dendrogrammes, boîtes à moustaches) pour faciliter le dépouillement des résultats. Finalement, des méta-classes seront proposées, visant à regrouper les styles proches en terme de qualité vocale en tenant compte de la valeur moyenne des paramètres estimés.

6.1 Constitution de la base

La base analysée contient de nombreuses occurrences prononcées selon 16 styles différents par un seul locuteur. En marge des analyses prosodiques (contour de F_0 notamment), des analyses sur la qualité vocale ont été nécessaires. Les méthodes développées précédemment sont donc appliquées aux fichiers de la base. Les 16 styles sont présentés sur la table 6.1.

Table 6.1 – Énoncé et description des 16 consignes (appelés plus tard *styles*) de la base de données.

Consigne	Description
Contraste Bas	Voix moyennement grave, effort faible
Dialogue	Phrases prononcées rapidement, modale
Dictée	Phrases articulées lentement avec emphase, modale
Didactique	Phrases articulées avec soin, modale
Enjoué	Voix gaie, modale
Fort	Effort important
Grave	Voix très basse en F_0
Insiste	Emphase sur la prononciation, mais plus rapide que dictée
Narratif Calme	Voix neutre, prosodie neutre
Narratif Tendu	Phrases prononcées sous le coup du stress
Point	Voix neutre, produite comme suvuie d'un point
Suspensif	Voix neutre, produit sans fin de phrase particulière
Triste	Voix triste, grave
Vieux	Voix tendue, chevrotante
Virgule	Voix neutre, produite comme suivie d'une virgule
Voix Basse	Voix basse presque chuchotée mais avec voisement

Cette base contient 39947 fichiers. Comme les algorithmes présentés sont encore assez gourmands en puissance, bien que l'analyse soit effectuée sur un serveur dédié à cette tâche (8 coeurs doublés par HyperTreading^{TM} et 32Go de mémoire vive), seuls 590 fichiers par style ont été

Table 6.2 – Table donnant la correspondance entre l'appellation des voyelles dans ce chapitre et leur représentation phonétique.

Appellation	API	exemple	Appellation	API	exemple	Appellation	API	exemple
OU	/u/	mou	AN	/ɑ̃/	rang	AU	/ɔ/	port
EU	/ø/	deux	O	/o/	rot	UN	/œ̃/	brun
AI	/ɛ/	belle	IN	/ɛ̃/	pain	OE	/œ/	leur
A	/a/	rat	ON	/ɔ̃/	on	UI	/ɥi/	suite
EI	/e/	ré	I	/i/	lit	U	/y/	rue

analysés, les phrases numérotées de 10 à 600. Par ailleurs, certains styles ne présentaient pas de données au-delà de la phrase 600, c'est donc le nombre maximum disponible pour comparer les 16 styles sur les mêmes données. L'analyse porte donc sur 9600 phrases ou morceaux de phrases répartis en 16 styles. Pour chaque style, un total de 10549 phonèmes ont été analysés.

La base analysée comporte 15 voyelles différentes[1]. Pour des facilités d'affichage et de dépouillement des données elles seront désignées par la manière dont s'écrit couramment en français le son qu'elles produisent : ces étiquettes étaient fournies avec les enregistrements de la base de données. Le tableau 6.2 représente la correspondance entre l'écriture d'un phonème et sa représentation phonétique. La majorité du dépouillement se fera sur ces 15 voyelles seulement.

6.2 Analyse et Protocole

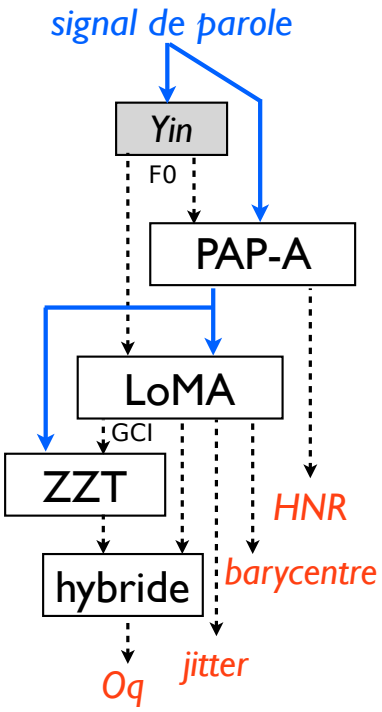


Figure 6.1 – Enchainement des méthodes d'analyse.

Les méthodes présentées précédemment ont permis d'extraire un certain nombre de paramètres décrivant la qualité vocale. Les quatre paramètres les plus fiables sont retenus :

- Le quotient ouvert, par hybridation entre ZZT et LoMA
- Le Jitter instantané à partir des GCI obtenus par LoMA
- Le barycentre spectral obtenu par LoMA
- Le rapport harmonique/bruit R obtenu pour tout instant t du signal selon l'équation 6.1 utilisant les parties harmoniques v et non harmoniques b estimées sur le fichier d'après la méthode présentée au chapitre 4. Le calcul se fait pour une pondération par une fenêtre de Hann w de taille $N = 320$ échantillons à 16kHz (20ms).

L'agencement des méthodes d'analyse pour l'estimation des 4 paramètres est donné en figure 6.1. Ces mesures seront exclusivement appliquées aux phonèmes voisés décrits dans la table 6.2 : en effet, la notion de fermeture glottique ou de forme de l'ODG n'est valable que pour des sons voisés.

1. Le glide /ɥ/ est inclu dans la voyelle /i/ et l'ensemble est considéré comme une voyelle à part entière.

$$R(n) = 10log_{10}\left(\frac{\sum_{k=n-N/2}^{n+N/2}(v(k)w(k-n))^2}{\sum_{k=n-N/2}^{n+N/2}(b(k)w(k-n))^2}\right) \tag{6.1}$$

Un exemple d'analyse est donné sur la figure 6.2. On y voit trois estimations précédemment décrites pour trois styles différents et pour la même phrase (n°20) "vous êtes professeur de sciences politiques ?". On constate déjà des différences entre les styles pour les 3 estimations données : le quotient ouvert, le barycentre et le rapport harmonique sur bruit.

Le dépouillement des résultats présente une grande importance. Il est nécessaire de se poser la question de la pertinence des représentations utilisées. Car la qualité vocale ne dépend pas que du style, mais aussi du phonème prononcé. Il est bien plus facile de crier /a/ que /i/ (et c'est bien pour ça que nous nous exclamons naturellement en /a/) et un locuteur peut très bien tenter de faire passer un style avec plus d'application sur certains phonèmes que sur d'autres. Les résultats doivent donc en tenir compte.

6.3 Résultats

De l'analyse des phonèmes voisés des 600 fichiers sont extraits 4 paramètres dont les résultats sont explicités dans le présent chapitre :

- La valeur moyenne du quotient ouvert selon la méthode hybride du chapitre 5
- Le jitter moyen (dans les cas où il est inférieur à 30%)
- Le barycentre spectral moyen exprimé en Hz selon les relations exposées au chapitre 3.
- La rapport harmonique sur bruit moyen, en décibels pour les parties voisées détectées comme telles par Yin.

La fréquence fondamentale ne sera pas étudiée dans ce chapitre car elle correspond d'avantage à un trait prosodique qu'à un paramètre de qualité vocale. La majorité des valeurs analysées est d'ailleurs normalisée par la fréquence fondamentale (jitter, shimmer, quotient ouvert).

L'importante quantité de résultats obtenus nécessite de porter une attention particulière à leur présentation. Deux représentations des distributions seront principalement utilisées :

- Les boites à moustaches, très adaptées à ce genre de résultats. Elles permettent de visualiser la position de la médiane de la distribution ainsi que les quarts supérieurs et inférieurs à cette médiane. Dans le cas de distributions ayant un écart type important mais pour lesquelles la majorité des résultats est concentrée autour d'une valeur, cette représentation est plus efficace que la représentation à base de moyenne et écart type.
- Les histogrammes, pour déterminer de quelle manière les valeurs estimées se répartissent selon la plage de variation du paramètre considéré. En effet, pour deux médianes très proches, on peut cependant disposer de distributions légèrement différentes.

Les résultats seront donc représentés selon ces deux formes et suivant quatre étapes :

- Dans un premier temps, les résultats seront représentés de manière globale par paramètre et par style, afin de dégager des tendances générales dans leur valeur.
- Dans un deuxième temps, les résultats seront séparés par voyelles. Ils seront alors représentés par style et par paramètre en fonction des phonèmes, ou alors par phonème et par paramètre en fonction des styles. Cette analyse permettra de dégager des tendances plus fines.
- Dans un troisième temps, une analyse statistique sera menée afin de déterminer les distances respectives entre styles et paramètres.
- Dans un quatrième temps, le lien entre les paramètres sera étudié.

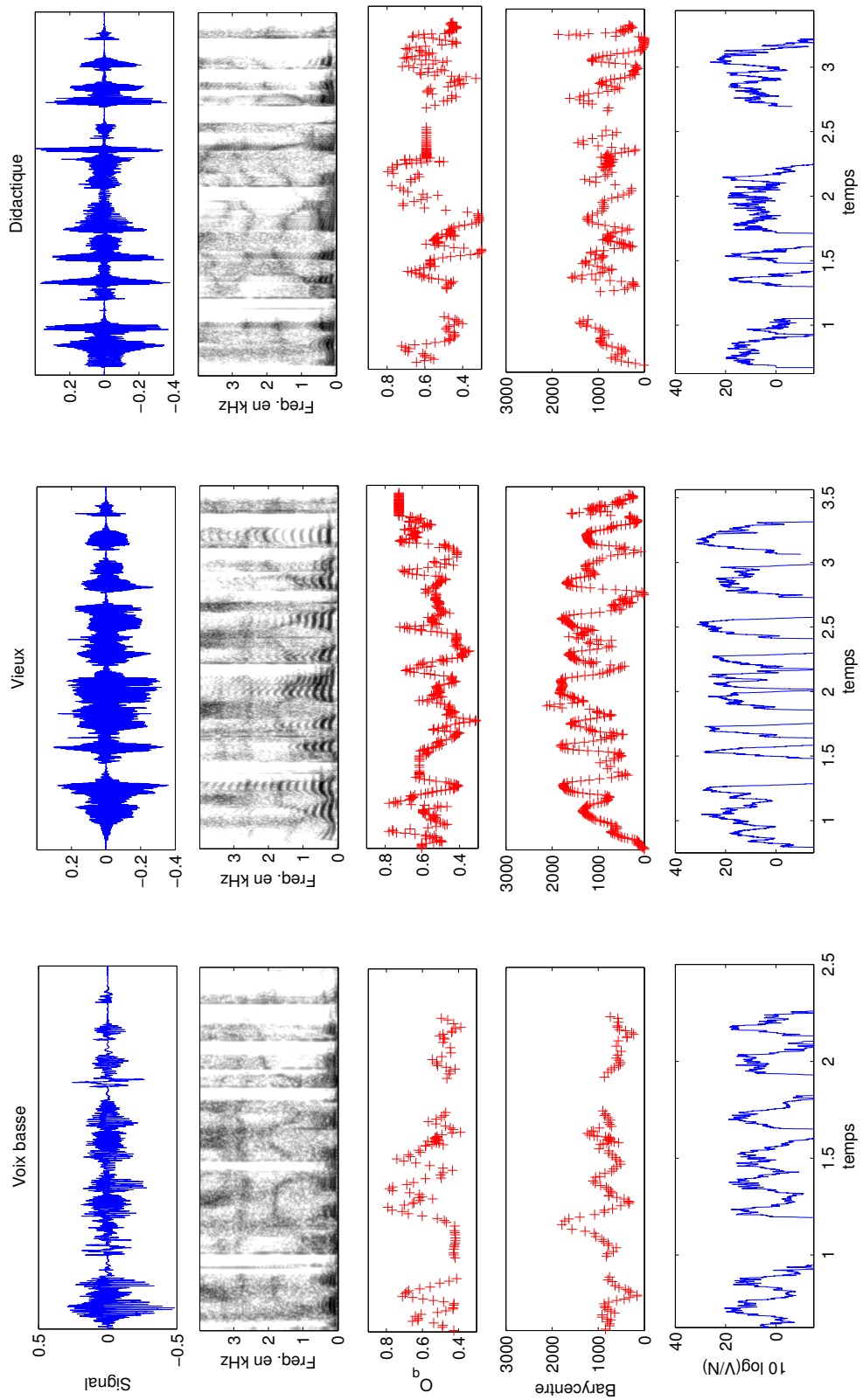


Figure 6.2 – Analyse de la phrase numéro 20 : "vous êtes professeur de sciences politiques ?", pour trois styles (voix basse, vieux, didactique). De bas en haut : le signal, son spectrogramme en bande étroite, les valeurs mesurées pour O_q, le barycentre (en Hz), et le rapport Harmonique sur Bruit en dB.

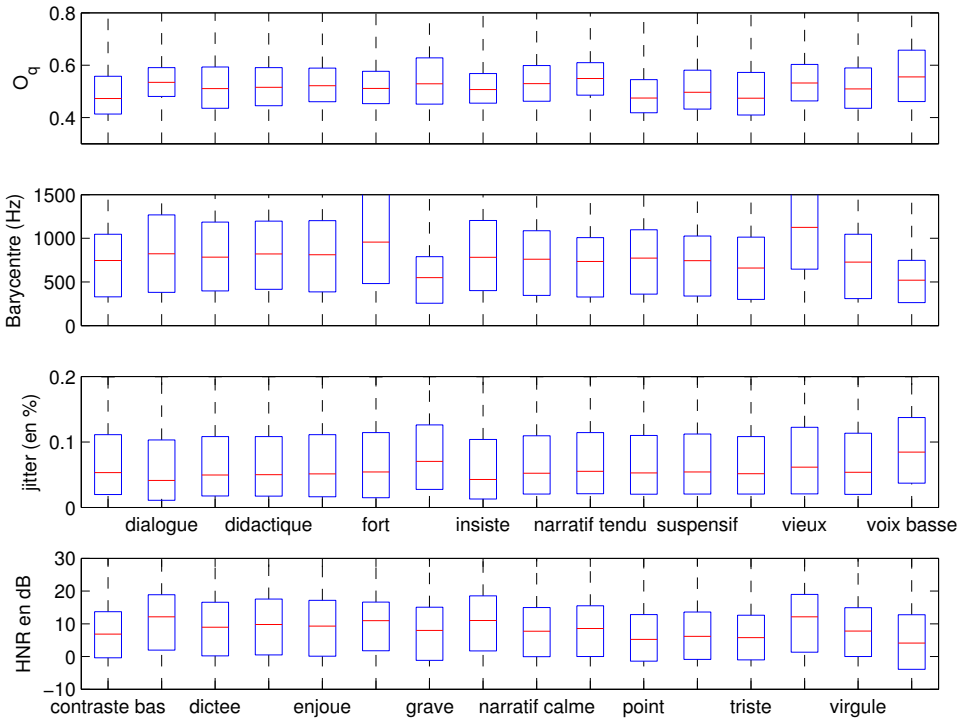


Figure 6.3 – Tendance générale par style sur les 4 paramètres estimés.

De ces observations seront bien entendu déduites des tendances générales permettant de discriminer un style d'un autre si cela est possible.

6.3.1 Tendances générales par style

Sur la figure 6.3 sont présentées les boites à moustaches pour les styles analysés et les 4 paramètres retenus. De prime abord, les styles *voix basse*, *fort*, *vieux* et *grave* se démarquent clairement en ce qui concerne les valeurs médianes pour le barycentre de la LoMA. Les styles *vieux* et *voix basse* se démarquent par leur rapport harmonique sur bruit alors que *voix basse* et *grave* présentent un quotient ouvert élevé.

6.3.2 Tendances par phonème voisé

Sur les figures 6.4, 6.5, 6.6 et 6.7 sont analysées en détails les distributions statistiques de détection de valeurs des paramètres estimés pour chaque style. Le nombre d'items pour chaque distribution dépend bien entendu du nombre de cycles glottiques et donc de la fréquence fondamentale, c'est pourquoi les figures sont présentées sous une forme normalisée. La quantité de paramètres estimés est importante par rapport aux nombre de styles présents, on s'attend donc à des variations importantes d'un style à l'autre et d'un paramètre à l'autre.

Sur la figure 6.4 sont présentées les distributions de quotient ouvert par style. Dans tous les cas, les distributions sont assez larges, recouvrant généralement l'ensemble des valeurs possibles mais dans des proportions différentes. On voit ainsi, par exemple, que les styles *contraste bas* et *suspensif* sont généralement plus tendus (valeur de O_q plus basse) que les styles *virgule* et *didactique*, eux même plus tendus que *grave* ou encore *voix basse*.

Sur la figure 6.5 sont présentées les distributions de jitter en % de la fréquence fondamentale.

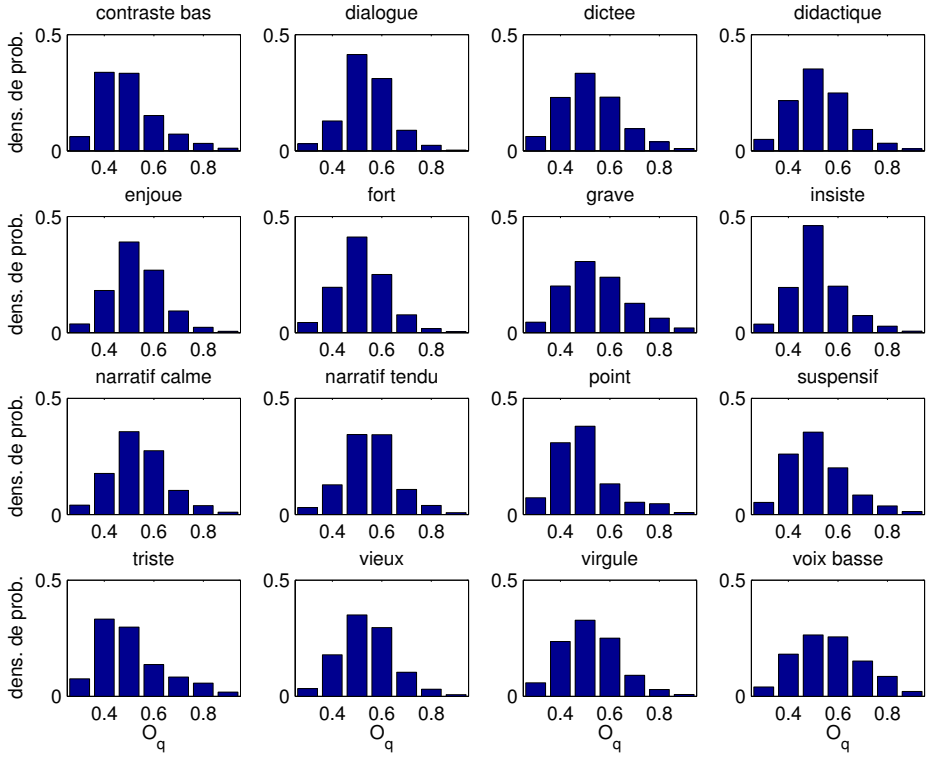


Figure 6.4 – Histogramme des distributions statistiques par style du quotient ouvert.

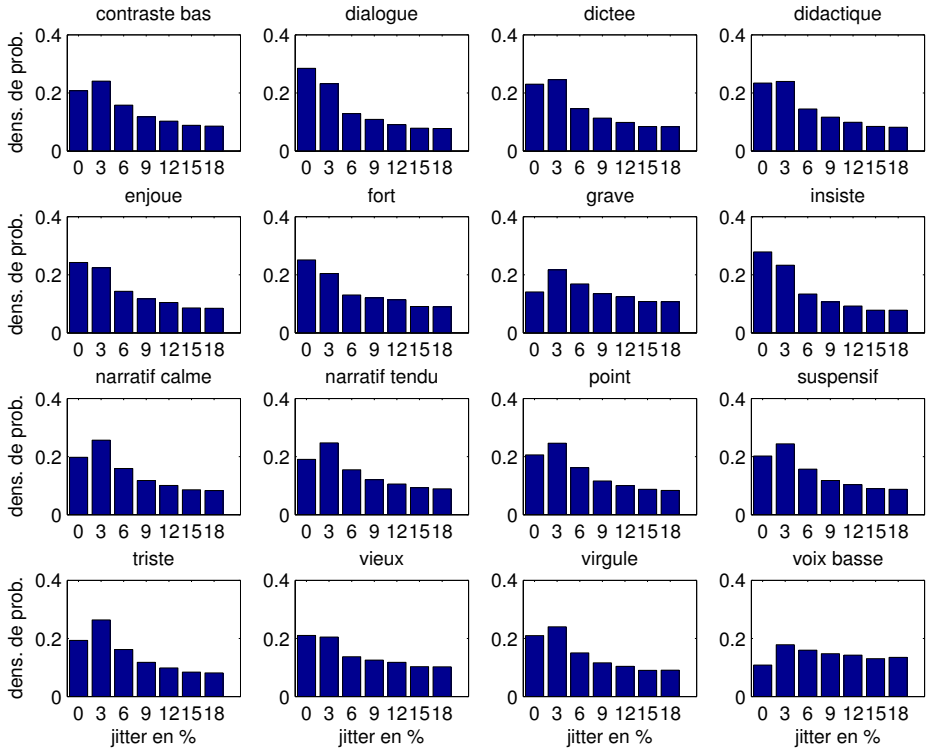


Figure 6.5 – Histogramme des distributions statistiques par style du jitter exprimé en %.

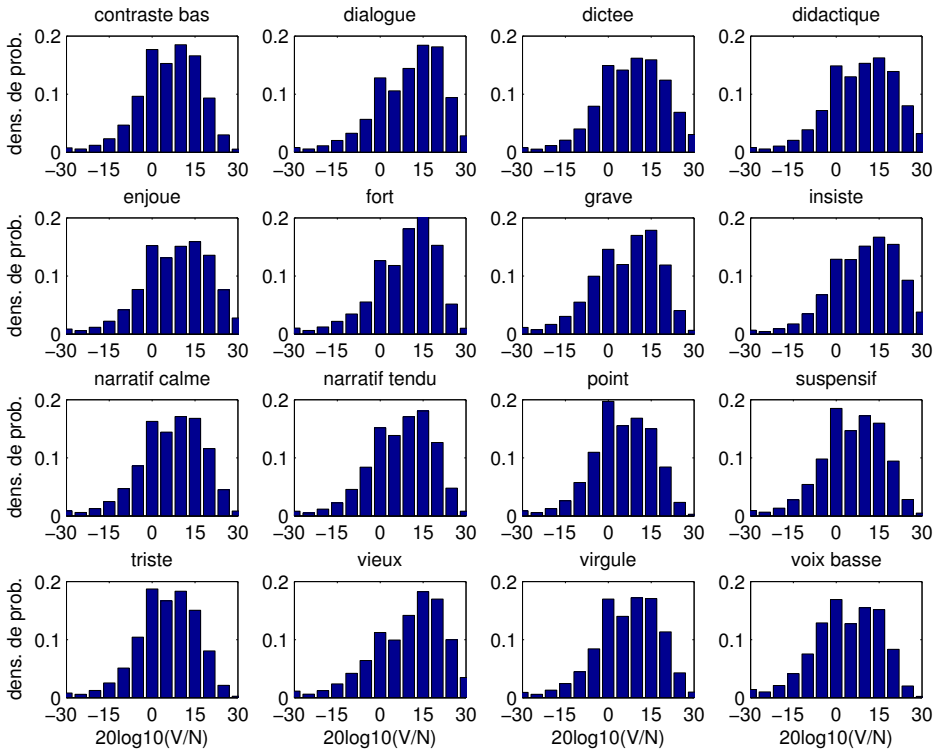


Figure 6.6 – Histogramme des distributions statistiques par style du rapport harmonique sur bruit exprimé en dB.

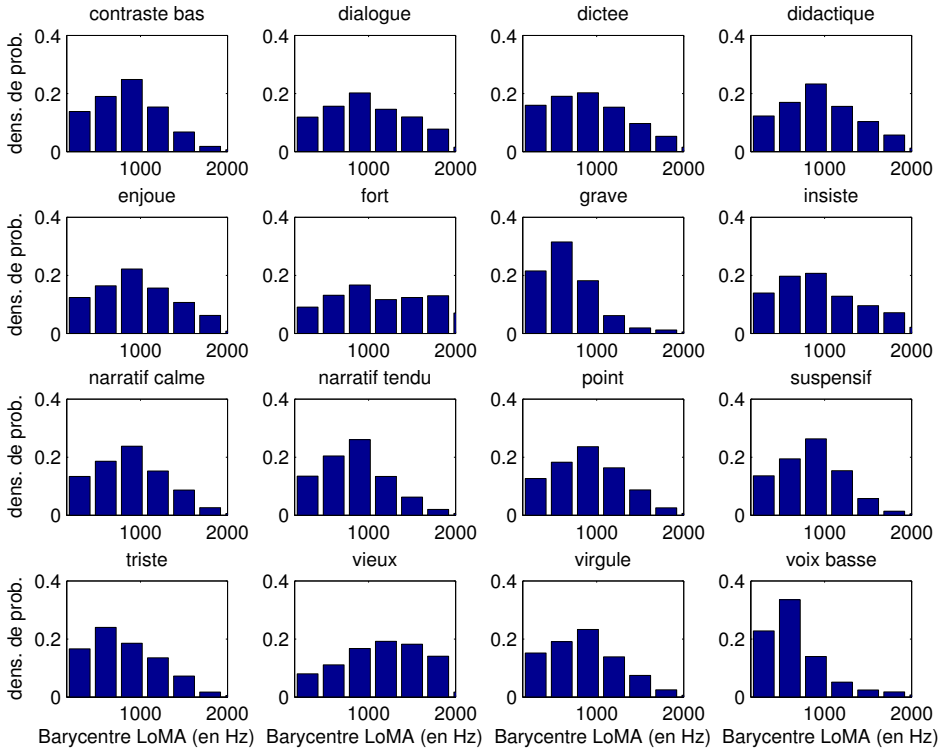


Figure 6.7 – Histogramme des distributions statistiques par style du barycentre de la LoMA.

Le jitter mesure les aspects de raucité et d'irrégularités dans la phonation. Dans tous les cas, les valeurs de jitter mesurée sont elevées, et il n'est pas rare de trouver statistiquement 25% des valeurs mesurée au dessus de 10% de jitter : ceci est probablement causé par des erreurs de détections au niveau de l'algorithme LoMA qui faussent légèrement les résultats. N'oublions cependant pas que ce sont des signaux réels, où la variation de fréquence fondamentale suffit elle seule à générer du jitter. On remarque trois types distincts de distributions.

- La première est une distribution décroissante quand le jitter augmente, avec un maximum à 0. Cette distribution montre un jitter très faible. Cette forme est visible notamment sur le style *vieux*.
- La deuxième présente toujours une distribution décroissante, mais avec un maximum autour de 3%. Cette distribution traduit un jitter faible à modéré, mais présent tout au long du discours. Cette forme est visible notamment sur le style *suspensif*
- La troisième distribution présente une forme plate, traduisant un jitter important et très présent tout au long du discours. Cette forme est visible notamment sur le style *grave*.

Ainsi, comme le précisent les études sur ce paramètre [Brockmann *et al.*, 2008] on remarque qu'il y a plus de jitter sur les styles *voix basse* et *grave* que sur le style *narratif calme*, par exemple.

Sur la figure 6.6 sont présentées les distributions de rapport harmonique sur bruit (HNR) en dB pour chaque style. Les distributions sont sensiblement identiques avec une présence légèrement plus importante dans les valeurs basses (bruit important) pour le style *voix basse* et un bruit clairement moins présent pour les styles *vieux* et *fort*.

Sur la figure 6.7 sont présentées les distributions de barycentre de LoMA pour chaque style. On retrouve encore ici des formes très caractéristiques. Un barycentre de LoMA plus haut en fréquence signifiant une excitation glottique plus riche (onde plus asymétrique, phase de retour plus rapide), on s'attend à une distribution placée plutôt en basse fréquence pour *voix basse* et *grave* et du contraire pour *fort* et *vieux*. On retrouve ces comportements sur la figure.

6.3.3 Analyse par phonème voisé

Il y a deux manières de représenter les analyses par phonème voisé :

- Détecter les variations par phonème pour des styles différents. Cette représentation est donnée sur les figures A.1 (barycentre de la LoMA), A.2 (quotient ouvert) , A.3 (jitter) et A.4 (rapport harmonique sur bruit).
- Détecter les variations par style pour des phonèmes différents. Cette représentation est donnée sur les figures A.5 (barycentre de la LoMA), A.6 (quotient ouvert), A.7 (jitter) et A.8 (rapport harmonique sur bruit).

Pour faciliter la lecture, ces figures sont regroupées en annexes.

En ce qui concerne le barycentre de la LoMA on constate que l'influence du phonème est très faible au regard des résultats présentés sur la figure A.1. En effet, les variations par qualité vocale sont similaires d'un phonème à l'autre. Cependant, la figure A.5 nous informe que certaines voyelles, notamment les voyelles fermées /eu/, /u/, /i/, /y/ et /ui/ présentent un barycentre de la LoMA généralement plus faible. Le fait que ces voyelles soient fermées produit naturellement une barrière à l'augmentation de la richesse spectrale. En effet, le troisième formant est placé très bas et ne permet pas de restituer beaucoup d'énergie au-delà de 2kHz. Sur la même figure, on remarque que les distributions par phonème sont très serrées mais que la variation phonème par phonème est importante. La dispersion observée sur la figure 6.3 pour le barycentre est donc principalement le fait d'une variation au sein du jeu des phonèmes analysés. Cette variation se

retrouve de manière identique pour tous les styles, renforçant d'autant plus l'utilisation de la médiane pour la caractérisation de l'influence de ce paramètre.

En ce qui concerne les valeurs de quotient ouvert on constate encore une fois que l'influence du phonème est faible au regard des résultats sur la figure A.2. Cependant, les voyelles fermées ont tendance à produire des distributions légèrement plus larges que les autres. Le phonème /œ̃/ présente une distribution atypique, probablement en raison du faible nombre d'occurrences rencontrées sur les 600 fichiers. En ce qui concerne les relations aux styles présentées en figure A.6, certaines - notamment *narratif tendu*, *narratif calme*, *point*, *dialogue*, *insiste* et *enjoué* - présentent des distributions beaucoup plus compactes.

En ce qui concerne les valeurs de jitter on constate que la quantité de jitter dans les phonèmes voisés est beaucoup plus variable, tant du point de vue du phonème - figure A.3 - que du style - figure A.7 -. On retrouve tout de même un jitter très présent sur les styles *fort* et *voix basse*, un jitter moyennement présent sur les styles *grave* et *contraste bas* et généralement faible à l'exception de deux phonèmes (/y/ et /œ̃/) pour le style *triste*. Au niveau de l'analyse par style, on remarque que malgré le niveau moyen de jitter assez élevé pour les styles *fort* et *voix basse*, la symétrie de la distribution n'est pas la même. La médiane est beaucoup plus basse pour le style *fort* confirmant quelque peu que le jitter sur les voix fortes est un phénomène qui apparait par "à coups" : de fortes valeurs, mais pas souvent.

En ce qui concerne les valeurs du rapport harmonique sur bruit présentées sur les figures A.4 et A.8, on constate que l'influence du phonème est très faible. Les variations selon les styles sont similaires d'un phonème à l'autre. Au niveau du style, les dispersions sont identiquement étendues à l'exception de *voix basse*, beaucoup plus étendue. Les voyelles /i/ et /u/ présentent généralement une distribution plus compacte et une médiane plus élevée du rapport harmonique sur bruit. Dans une moindre mesure on retrouve aussi ce comportement sur la voyelle /u/ au contraire de la voyelle /o/ et on peut donc généraliser ce comportement aux voyelles fermées. Une explication pourrait se trouver dans la volonté du locuteur de se faire comprendre : comme les voyelles fermées ont un rendement plus faible, la nécessité de voiser peut être plus présente.

6.4 Confirmation des tendances par analyse statistique

La majorité des analyses statistiques demandent un nombre d'observations identiques pour chaque expérience comparée. Or, le nombre d'estimations de O_q, du barycentre de la LoMA ou même du rapport harmonique sur bruit varie en fonction du style, le locuteur parlant plus ou moins vite, plus ou moins grave.

Le dénominateur commun à chaque expérience demeure le nombre de phonèmes prononcés. Les paramètres seront donc regroupés par phonème (par une simple moyenne algébrique) et analysés ainsi afin de déterminer des relations statistiques entre style et valeur de paramètre.

Dans un premier temps, les résultats de l'analyse par composante principale des quatre paramètres seront présentés.

Dans un deuxième temps, c'est le classement hiérarchique qui a été choisi pour présenter les résultats. Une distance est calculée entre les groupes (par style). Ces styles sont ensuite ordonnés dans un arbre (un dendrogramme) par distance relative. On visualise alors sur les figures 6.9 et

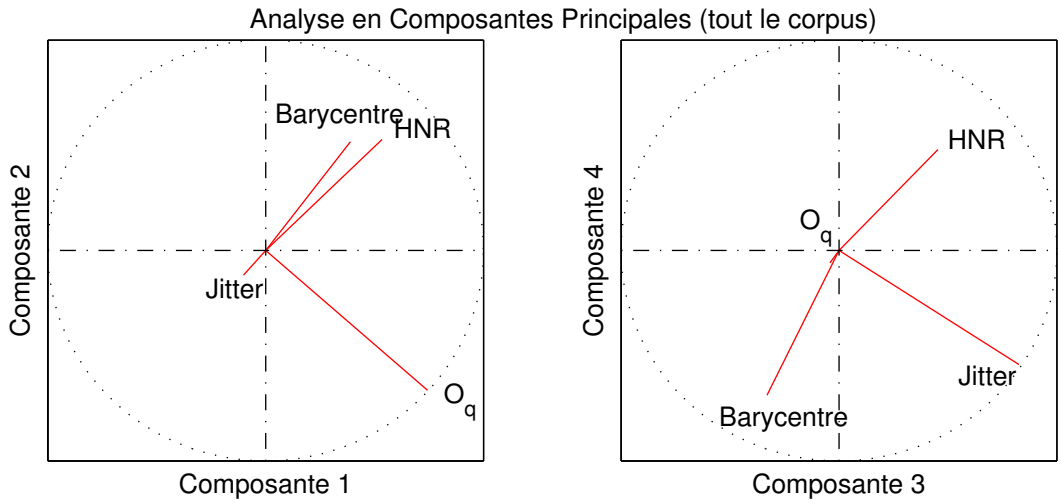


Figure 6.8 – Résultats de l'analyse par composantes principales des quatre paramètres sur tout le corpus

6.10 le classement hiérarchique par paramètre. Plus les styles sont proches les un des autres, et plus leurs distributions sont proches en valeurs pour le paramètre considéré.

6.4.1 Indépendance des paramètres

Les projections de chaque paramètre sur chacune des quatre composantes principales obtenues par analyse statistique sont données sur la figure 6.8. Chaque vecteur a pour norme 1. Sur les composantes les plus fortes, on remarque l'émergence de 3 dimensions prédominantes : une dimensions donnée par le quotient ouvert, une dimension donnée par le jitter (orthogonale au plan des composantes 1 et 2) et une dimension regroupant le HNR et le barycentre. Sur les deux composantes suivantes (3 et 4), HNR et Barycentre sont aussi colinéaires.

Cette analyse montre la pertinence du choix des quatre paramètres pour l'analyse et la possibilité de décrire la qualité vocale selon trois dimensions indépendantes grâce aux paramètres choisis.

6.4.2 Classement hiérarchique paramètre par paramètre

- Pour le quotient ouvert (haut de la figure 6.9) : deux groupes se distinguent au niveau de leur moyenne générale : d'un coté (en rouge) le groupe présentant des quotients ouverts faibles, et de l'autre (en noir) un groupe présentant des quotients ouverts élevés. Cependant, de ces deux sous-groupes, aucune structure hiérarchique claire ne se démarque.
- Pour le barycentre de la LoMA (bas de la figure 6.9) : trois groupes se distinguent. En rouge, un groupe est relié aux faibles valeurs du barycentre (*voix basse*, *grave*), en vert un groupe présentant des valeurs élevées et en bleu un groupe intermédiaire. La structure hiérarchique est claire, et les groupes rouge et vert se détachent nettement du reste.
- Pour le jitter (en haut de la figure 6.10) tout comme pour le quotient ouvert, deux groupes se distinguent, ici de manière assez nette. Un groupe (en noir) présente des valeurs élevées de jitter. On remarquera que *fort* et *vieux* font partie de la même sous classe, confirmant l'aspect très similaire de leur distribution. En rouge, on retrouve un groupe aux éléments difficilement séparés correspondant aux valeurs faibles de jitter.

– Pour le rapport harmonique sur bruit, trois groupes se distinguent : un groupe présentant un HNR élevé (en rouge), un groupe présentant un HNR moyen (en bleu) et un groupe présentant un HNR faible (en noir). La dissociation à l'intérieur de ces groupes est assez importante. On retrouve quelques sous groupes indépendants et notamment le sous groupe bleu.

6.4.3 Classement hiérarchique par phonème

Tout comme l'analyse de la distribution des valeurs estimées, se pencher sur le classement hiérarchique par phonème apporte beaucoup d'informations pour déterminer les principales caractéristiques de chaque style. Les classements en dendrogrammes par paramètre et par style des phonèmes voisés sont donnés sur les figures A.9 à A.24. Les classements hiérarchiques sont très similaires d'un style à l'autre, mais certains styles présentent un classement différent, soulignant une distribution à l'allure atypique. Faire référence à chaque graphique des figures A.9 à A.24 serait trop long et répétitif, les données sont fournies en annexe pour mémoire.

6.5 Interactions source-filtre

L'analyse des figures A.1 à A.4 permet aussi de visualiser les interactions source-filtre. Pour un style donné, les résultats ne sont pas similaires selon les voyelles. L'effet est particulièrement marqué sur le jitter et le barycentre de la LoMA. Cependant, comme le barycentre de la LoMA est mesuré sur le signal de parole (et non sur la source), on peut considérer qu'il dépend tout autant du filtre que de la source, et donc que sa variation en fonction de la voyelle ne dépend pas nécessairement d'interactions source filtre.

Dans une moindre mesure, des valeurs de paramètres de source dépendant de la configuration du conduit vocal peuvent aussi être observées sur la figure A.1 où est représentée la distribution du quotient ouvert. On y remarque des différences tant au niveau de l'amplitude de la distribution que de la médiane. On remarque systématiquement la même forme de variation pour tous les styles, preuve qu'il est plus facile dans certaines configurations du filtre vocalique de produire certaines valeurs de quotient ouvert.

Enfin, pour le rapport harmonique sur bruit, les différences sont encore moins marquées que pour le quotient ouvert, mais il est toujours possible de voir un certain schéma de variation à travers les voyelles qui se reproduit pour tous les styles.

6.6 Corrélation entre les estimations

6.6.1 Corrélation par style

Les corrélations sont présentées sur la table 6.3 par combinaisons successives de deux paramètres et par style. On remarque que les valeurs sont généralement basses, proches de 0 pour beaucoup de combinaisons. La seule combinaison présentant une faible corrélation entre les valeurs des paramètres est la combinaison Jitter-HNR sauf pour les styles *fort*, *grave*, et *voix basse*. Ces 3 styles sont placés dans les extrêmes de variations en terme de qualité vocale, il est compréhensible qu'ils présentent des évolutions prédominantes d'un de ces deux paramètres sur l'autre. Pour cette combinaison Jitter-HNR, on remarque aussi que du fait des valeurs négatives, une hausse du HNR entraine une baisse du jitter. Dans une moindre mesure, ces valeurs sont aussi corrélées au barycentre : une hausse du barycentre de la LoMA entraine une baisse du jitter et une hausse du rapport harmonique sur bruit.

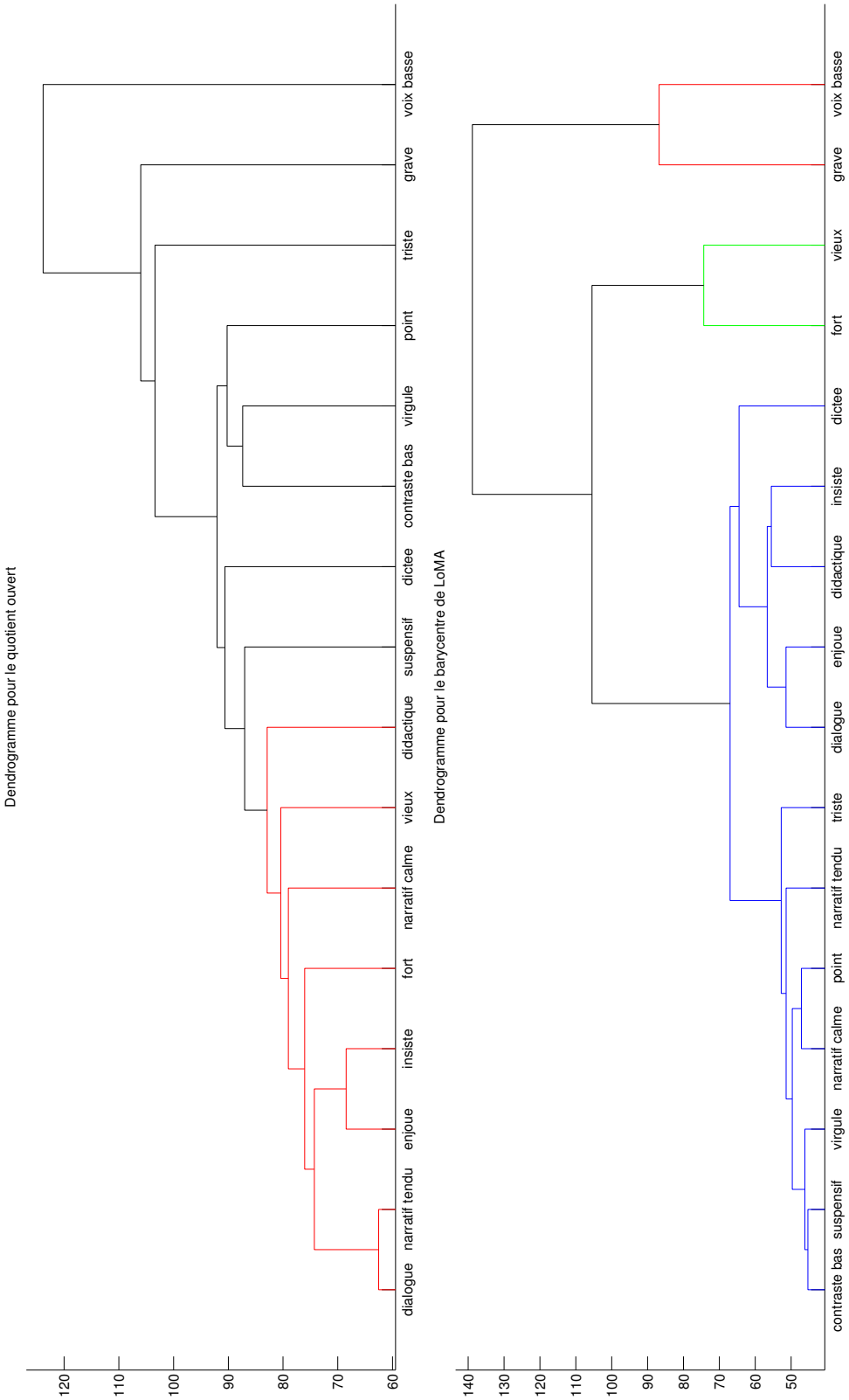


Figure 6.9 – Dendogrammes sur toutes les analyses, par styles, pour le quotient ouvert et le barycentre de la LoMA. Distance euclidienne standardisée moyennée et non pondérée. Les groupements de couleurs indiquent des groupes aux éléments très proches les uns des autres.

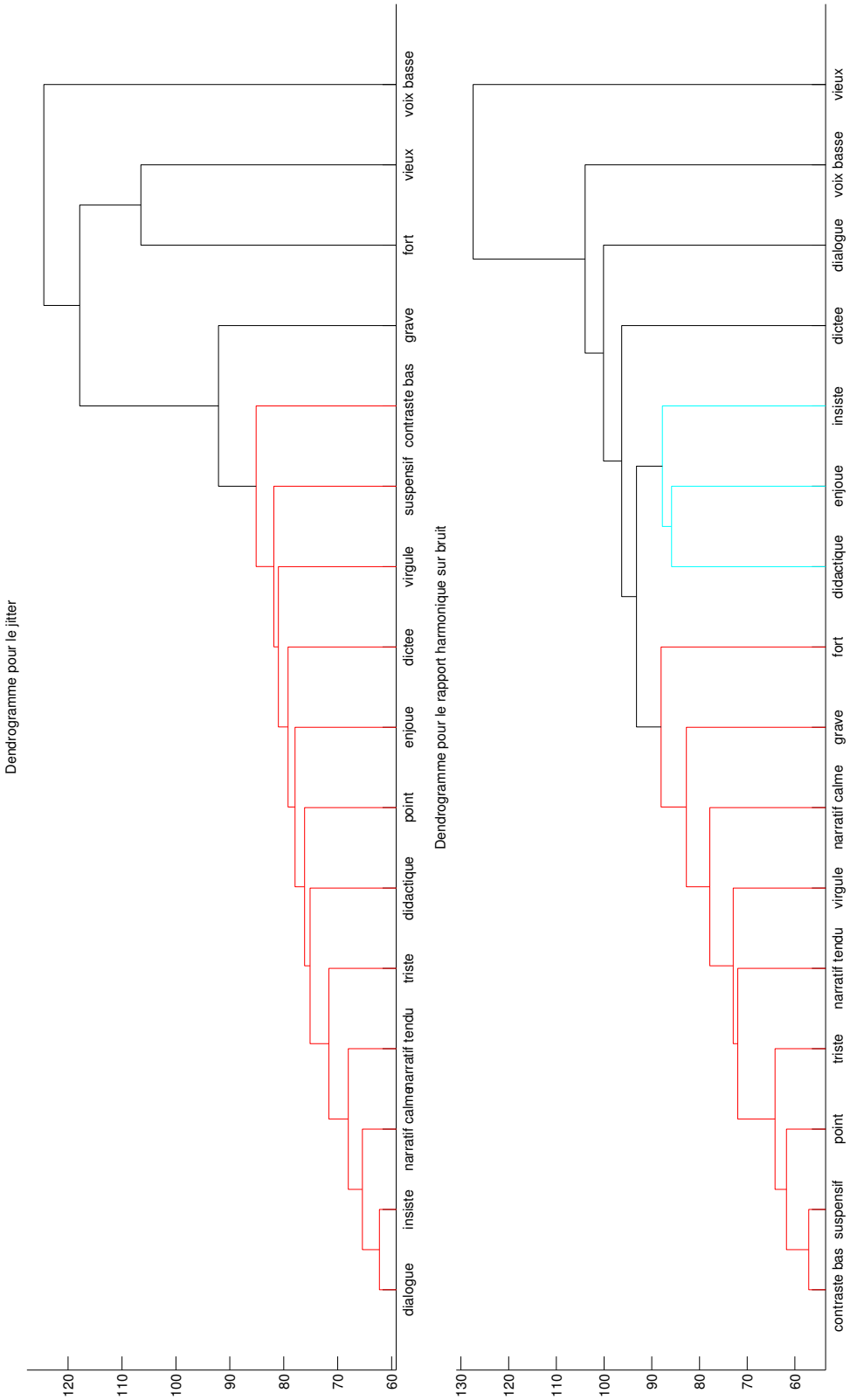


Figure 6.10 – Dendogrammes sur toutes les analyses, par styles, pour le jitter et le rapport harmonique sur bruit. Distance euclidienne standardisée moyennée et non pondérée. Les groupements de couleurs indiquent des groupes aux éléments très proches les uns des autres.

Table 6.3 – Corrélation des résultats par style.

	O_q-Bary.	O_q-jitter	O_q-HNR	Bary.-jitter	Bary.-HNR	jitter-HNR
contraste bas	-0.2342	0.28971	-0.26036	-0.51597	0.43697	-0.57825
dialogue	-0.065436	0.004531	0.049226	-0.64216	0.60944	-0.69662
dictée	-0.077304	0.16	-0.022046	-0.5038	0.30948	-0.53267
didactique	-0.10559	0.13764	0.042751	-0.54113	0.37773	-0.56046
enjoué	-0.097344	0.095932	0.06061	-0.56632	0.49492	-0.51948
fort	-0.15183	-0.049053	0.15397	-0.2917	0.31874	-0.44026
grave	-0.19077	0.1491	-0.12936	-0.21001	0.213	-0.37707
insiste	-0.17285	0.14356	-0.019985	-0.60953	0.44392	-0.6352
narratif calme	-0.11509	0.12516	0.012278	-0.54763	0.44931	-0.58479
narratif tendu	-0.06022	0.031634	0.050485	-0.51445	0.45994	-0.61119
point	-0.17334	0.19309	-0.11464	-0.53846	0.44781	-0.59392
suspensif	-0.18851	0.23614	-0.20006	-0.50748	0.42234	-0.58168
triste	-0.27206	0.29172	-0.27396	-0.50174	0.38276	-0.53565
vieux	-0.2303	0.010173	0.046179	-0.44889	0.31156	-0.49415
virgule	-0.17607	0.23065	-0.080838	-0.52853	0.41989	-0.58157
voix basse	0.0062204	0.077674	0.054844	0.13672	0.046512	-0.18509

Table 6.4 – Corrélation des résultats par phonème.

	O_q-Bary.	O_q-jitter	O_q-HNR	Bary.-jitter	Bary.-HNR	jitter-HNR
EU	-0.20231	0.16426	-0.025231	-0.21024	0.38229	-0.43846
AI	-0.24655	0.1676	0.06413	-0.19841	0.21044	-0.33001
A	-0.26506	0.19984	0.049121	-0.27783	0.21182	-0.38925
OU	-0.10558	0.13299	0.10811	-0.1059	0.21716	-0.2809
ON	-0.23565	0.17545	0.040334	-0.29042	0.28948	-0.38563
I	-0.015433	0.23946	0.044495	-0.13998	0.30266	-0.37173
AU	-0.23863	0.16952	0.019108	-0.22089	0.35162	-0.35295
AN	-0.27646	0.22296	0.076161	-0.32116	0.25772	-0.37645
O	-0.15212	0.10084	0.1439	-0.15763	0.25602	-0.28958
IN	-0.2875	0.21901	0.087107	-0.28905	0.14438	-0.33752
EI	-0.23704	0.18734	-0.011097	-0.16188	0.29442	-0.40217
OE	-0.20956	0.16135	0.2124	-0.14469	0.109	-0.23541
U	-0.010738	0.24022	0.056837	-0.082717	0.24901	-0.30034
UI	0.0058611	0.24429	0.0097881	-0.068246	0.19334	-0.48443
UN	-0.38583	0.30497	0.059288	-0.31598	0.15092	-0.35935

D'une manière générale, également, on remarque que le quotient ouvert est décorrélé des trois autres paramètres mesurés. Ceci peut être choquant, dans la mesure où il est admis qu'une forte valeur de quotient ouvert engendre un HNR moins élevé, ce n'est toutefois pas le cas pour ce locuteur.

6.6.2 Corrélation par phonème

Les corrélations sont présentées sur la table 6.4 par combinaisons successives de deux paramètres et par phonème voisé. Ce calcul de corrélation à travers tous les styles donne des résultats similaires à deux de la table 6.3. En effet, on retrouve la corrélation entre barycentre, jitter et rapport harmonique sur bruit, mais dans une moindre mesure. Il est tout à fait légitime de trouver une corrélation, même faible, entre quotient ouvert et barycentre spectral pour une voyelle donnée. En effet, le barycentre spectral de LoMA représente l'énergie de chaque cycle glottique, qui est fortement conditionnée par la forme du filtre vocalique alors que le quotient ouvert, par interaction source-filtre, peut aussi être conditionné par la voyelle prononcée.

On retrouve tout de même une forte décorrélation du quotient ouvert avec le rapport harmonique sur bruit.

6.7 Caractérisation des styles

L'analyse des résultats obtenus permet de tirer des propriétés pour chaque paramètre et chaque style. Il peut donc être intéressant d'essayer de positionner ces styles dans un domaine en 4 dimensions. On rappelle que les styles sont décrits dans le tableau 6.1 en début de chapitre.

6.7.1 Commentaires style par style

Contraste Bas

Ce style présente des valeurs de O_q assez basses, un niveau de jitter tout à fait dans la normale, un barycentre de la LoMA positionné bas dans le spectre, généralement sous les 1000Hz et un rapport harmonique sur bruit assez bas, majoritairement inférieur à 10dB. On retrouve ces caractéristiques dans les dendrogrammes présentés aux figures 6.9 et 6.10. Cependant, ce style ne se démarque pas nettement des autres.

Dialogue

Ce style présente des valeurs concentrées du quotient ouvert autour de 0.5 et barycentre de LoMA assez dispersé, un jitter très faible et un rapport harmonique sur bruit élevé, concentré autour de 15dB. Sur la figure 6.9 on trouve que ce style est groupé avec *narratif tendu* pour le quotient ouvert et *vieux, fort* pour le barycentre. Ce dernier groupe se démarque particulièrement des autres. Sur la figure 6.10 on retrouve le style à l'extrême du classement hiérarchique pour le jitter, confirmant que les faibles valeurs de jitter observées sont un caractère déterminant du style. Au niveau du rapport harmonique sur bruit, ce style se situe dans un sous-groupe assez large de 5 styles. On retiendra donc la faible valeur de jitter comme caractère fondamental de ce style, avec une propension à des rapports harmonique sur bruit élevé.

Dictée

Tout comme *Contraste bas*, ce style ne possède pas vraiment de paramètre caractéristique. On remarque tout de même des différences sensibles en décomposant voyelle par voyelle. En

effet, le style *Dictée* présente des valeurs de barycentre spectral légèrement plus hautes pour les voyelles comme /a/ ou /o/. Ce style présente un placement central sur les quatres classifications hiérarchiques des figures 6.9 et 6.10.

Didactique

Tout aussi proche de *Dictée* et *Contraste Bas*, ce style se distingue tout de même par des valeurs de quotient ouvert légèrement plus hautes. Tendance qui est confirmée pour la majorité des phonèmes analysés sur la figure A.2. Le classement hiérarchique nous informe également que ce style se situe dans le sous-groupe présentant un rapport harmonique sur bruit élevé.

Enjoué

Ce style présente des distributions proches de *Didactique* et *Dictée* mais une distribution plus étroite du quotient ouvert. On remarque sur la figure 6.4 que les valeurs de O_q sont principalement distribuées autour de 0.5. On retrouve l'étroitesse de cette distribution sur toutes les voyelles de la figure A.2. Sa forme caractéristique est probablement la raison du classement hiérarchique de la figure 6.9 (en haut) où le style est dans le dernier sous-groupe pour le quotient ouvert. La principale caractéristique de ce style est donc la stabilité des valeurs de O_q et la forte probabilité de les trouver sur la valeur médiane de la distribution.

Fort

Ce style est réellement à part. La distinction se situe principalement au niveau du barycentre de la LoMA et du rapport harmonique sur bruit. Le barycentre est situé haut en fréquence alors que le rapport harmonique sur bruit est exceptionnellement élevé. L'inspection du jitter sur les phonèmes voisés de la figure A.3 montre qu'il présente une distribution large avec une médiane basse. Ces distributions atypiques placent systématiquement ce style dans des sous-groupes restreints de classements hiérarchiques. Ce sont donc des valeurs hautes du rapport harmonique sur bruit et du barycentre de la LoMA, ainsi qu'une distribution reconnaissable du jitter qui caractérisent ce style.

Grave

Ce style se distingue des autres. On remarque principalement des valeurs basses du barycentre de la LoMA alors que le jitter et le rapport harmonique sur bruit présentent des distributions normales. Au niveau du quotient ouvert, la dispersion de la distribution est plus importante mais l'analyse par voyelle sur la figure A.2 n'indique pas de différence marquante. C'est donc principalement les valeurs du barycentre de la LoMA qui caractérisent ce style, on remarque d'ailleurs que c'est dans le classement hiérarchique du barycentre de la LoMA qu'il est le plus isolé, groupé avec *voix basse*.

Insiste

Ce style présente des valeurs de paramètres normales, mais le rapport harmonique sur bruit est élevé. Ce style se distingue donc uniquement sur ce paramètre. Groupée avec *enjoué* sur les classements hiérarchiques du quotient ouvert et du barycentre de la LoMA on retrouve l'exception sur le rapport harmonique sur bruit par son placement dans le sous groupe bleu de la figure 6.10.

Narratif calme

Ce style présente une distribution de valeurs proches de *Contraste Bas*. La seule différence se situant au niveau de la distribution des valeurs de quotient ouvert.

Narratif tendu

Ce style présente des distributions similaires à *Narratif Calme* avec comme principale différence les valeurs de quotient ouvert, comme le montre la figure A.2. Étonnamment, *Narratif Calme* présente des valeurs de quotient ouvert plus faibles que *Narratif tendu*. Sur les figures 6.9 et 6.10 ces deux styles sont classés proches l'un de l'autre et de manière très médiane, sauf pour le quotient ouvert.

Point

Les distributions de ce style sont proches de *Constraste Bas* à l'exception de O_q dont la distribution est, quant à elle, proche des styles *enjoué* et *fort*. On constate cependant un jitter légèrement moins marqué et un rapport harmonique sur bruit un peu moins marqué que sur le premier style (*contraste bas*) lors de l'analyse des figures A.3 et A.4. Cette tendance est confirmée par le classement hiérarchique : toujours à part, et jamais dans des sous groupes très caractéristiques.

Suspensif

Ce style ne présente pas de réelles différences avec *dictée*, il faut chercher les différences par voyelle sur le jitter et le quotient ouvert pour arriver à différencier les deux. On remarque d'ailleurs que ce style n'est jamais classé de manière singulière sur les dendrogrammes.

Triste

Ce style ne montre pas de caractéristique précise, mais l'analyse au niveau des phonèmes montre une tendance à un quotient ouvert plus bas et un jitter moins élevé.

Vieux

Ce style se distingue principalement par les fortes valeurs du rapport harmonique sur bruit - majoritairement au dessus de 15dB - et un barycentre de LoMA très haut en fréquence. On retrouve ces caractéristiques sur le classement hiérarchique pour lequel ce style est placé dans des groupes spécifiques sur ces deux mêmes paramètres.

Virgule

Ce style ne se distingue pas vraiment au niveau de la qualité vocale, tout comme *Point* ou encore *Narratif calme* et *narratif tendu*. Il est d'ailleurs généralement classé de manière inintéressante sur les dendrogrammes.

Voix basse

Ce style, une vraie qualité vocale à lui seul, se distingue sur les 4 paramètres estimés dans cette étude. Le quotient ouvert est globalement plus élevé, le barycentre de la LoMA beaucoup plus bas que la normale, le jitter élevé, avec une médiane haute sur toutes les voyelles et le

rapport harmonique sur bruit exceptionnellement bas. On retrouve bien ce placement à part sur le classement hiérarchique des figures 6.9 et 6.10.

6.7.2 Visualisation

Sur la figure 6.11 sont présentées les visualisations en 2 dimensions des valeurs moyennes mesurées sur chaque style. Cette représentation permet de visualiser plus facilement les remarques faites à la section précédente, du point de vue de la moyenne de la distribution.

On remarque notamment que certains styles se démarquent plus facilement que d'autres. Ainsi, un groupe constitué des styles *narratif calme, narratif tendu, point, suspensif, didactique, dictée*, et *contraste bas* reste distribué de manière très compacte sur les 6 plans proposés en figure 6.11. Á l'intérieur de ce groupe, les variations sont légères et il est probable que leurs différences se situent davantage sur un plan prosodique.

D'autres styles, dont *voix basse, grave, fort, vieux*, se démarquent clairement et on retrouve les propriétés de chacun d'entre eux, énoncées à la section précédente.

6.7.3 Méta-classes de qualité vocale

A la lumière des résultats présentés et de leurs analyses, on est en mesure de grouper les 16 styles d'élocutions en 4 méta-classes de qualité vocale :

1. Une classe **modale**, regroupant les styles ne faisant pas état de valeurs remarquables dans les 4 paramètres analysés : *narratif calme, narratif tendu, suspensif, virgule, contraste bas, enjoué, didactique, insiste, grave*
2. Une classe **voix basse** dédiée au style *voix basse* qui tient une place particulière pour les la valeurs des paramètres mesurés.
3. Une classe **d'effort** dans laquelle on retrouve les deux styles *vieux* et *fort*.
4. Une classe **claire** dans laquelle on retrouve le style *dialogue* qui produit tous les caractères d'une voix claire : fort HNR, jitter bas, barycentre spectral moyen.
5. Une classe **tendue** dans laquelle on retrouve les styles *triste, point et dictée*, généralement groupés sur la figure 6.11 mais qui présentent une dispersion importante en matière de jitter.

6.8 Conclusion

Dans ce chapitre, les méthodes présentées précédemment ont été utilisées pour analyser un grand corpus de données, composé de fichiers de parole produits par un seul locuteur, séparés en 16 styles répartis sur des dimensions à la fois de qualité vocale, mais aussi de prosodie. L'analyse de ce corpus est axée uniquement sur les caractéristiques de qualité vocale dans l'idée de séparer les styles par les 4 paramètres estimés à partir de méthodes développées dans ce travail de thèse.

Les résultats de l'analyse de 600 fichiers de cette base de données sur ces quatre paramètres ont été analysés de manière statistique. Entre autres visualisations, l'utilisation d'une représentation hiérarchique a permis de dégager des groupes cohérents au niveau des résultats pour chacun des 4 paramètres étudiés. Dans un souci d'analyse approfondie du corpus, une analyse au niveau des phonèmes voisés a aussi été proposée, à la fois pour la distribution et le classement hiérarchique. Une étude de la corrélation entre les résultats a montré une faible dépendance des paramètres entre eux, à l'exception du quotient ouvert qui semble décorrélé des autres paramètres.

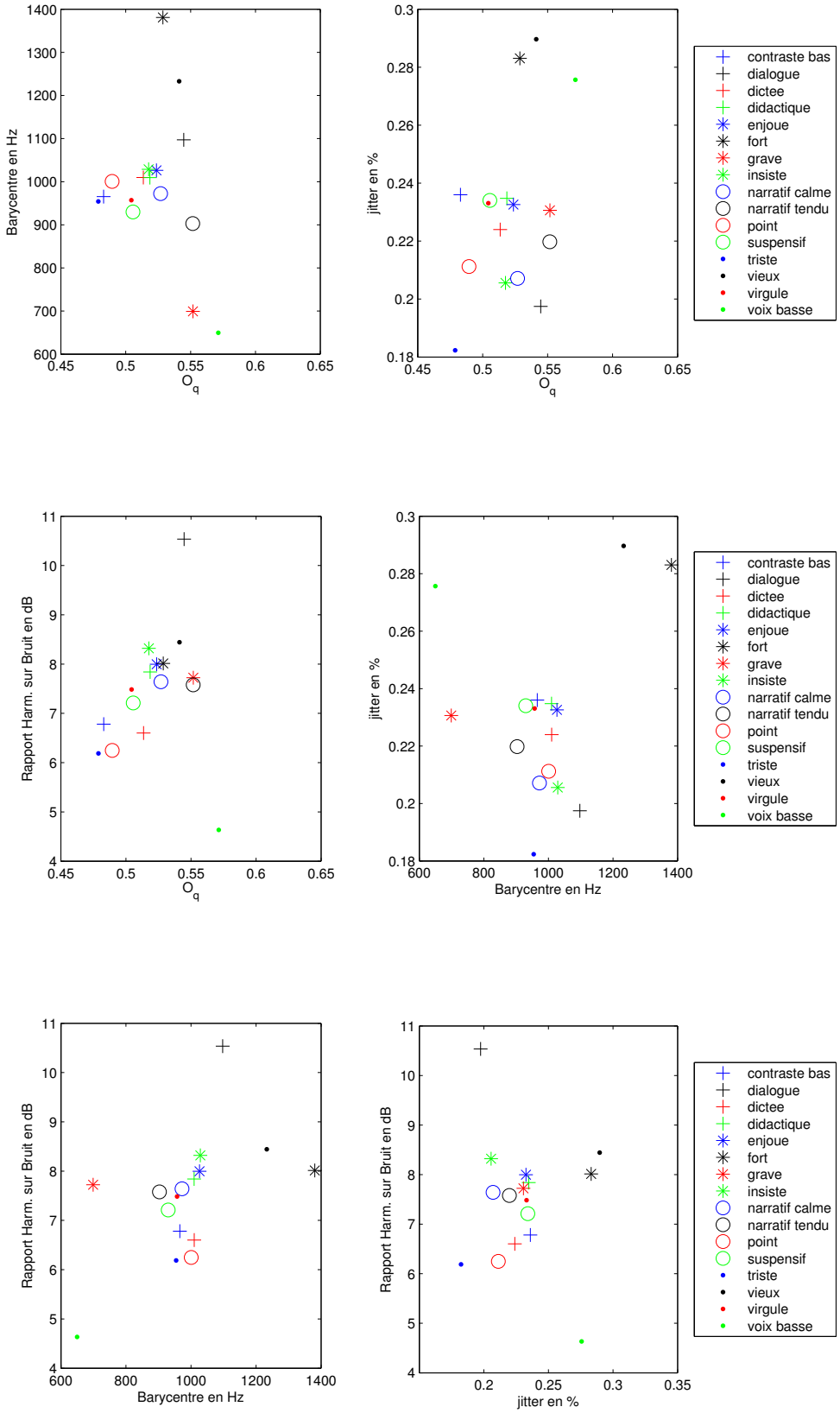


Figure 6.11 – Représentation des moyennes des analyses pour chaque style en deux dimensions. Certains styles se démarquent plus clairement que d'autres.

Cette analyse à grande échelle montre la viabilité des méthodes d'analyse proposées, et montre la pertinence du choix des 4 paramètres analysés par la faible corrélation présente entre eux. Du point de vue de la production vocale, l'analyse du corpus a ainsi montré que certains styles se démarquent clairement au niveau de la qualité vocale *fort, vieux, grave, triste* alors que les autres présentent des qualités vocales qui peuvent être considérées comme modales.

Résumé

Analyse d'un corpus

Un corpus séparé en 16 styles a été analysé, pour un total de 9600 phrases ou bouts de phrases. Les résultats ont été analysés en terme de distribution des valeurs estimées et de hiérarchie, à la fois par style mais aussi par phonèmes voisés. Une étude de la corrélation entre paramètres et phonèmes a aussi été effectuée. Cette analyse a permis de mettre en valeur 4 styles qui ont recours à des qualités vocales spécifiques, alors que les 12 autres styles proposent des variations plus fines des paramètres. Cette tendance est confirmée par une visualisation de la répartition des moyennes sur les quatre dimensions analysées.

En conclusion

Ce test a permis de vérifier l'applicabilité à grande échelle des méthodes développées dans cette thèse. Il a aussi été montré que les 4 paramètres retenus sont indépendants au niveau de leur variation.
Enfin, cette analyse a permis de valider la notion de qualité vocale comme un sous espace de la prosodie : chaque style présentant des différences plus ou moins nettes sur un ou plusieurs des paramètres analysés.
Des 16 styles proposés pour l'analyse, cinq méta-classes ont été proposées, reflétant chacune une qualité vocale particulière.

Chapitre 7

Conclusion

Sommaire

Segmentation des signaux vocaux 190
Décomposition harmonique + bruit 191
Déconvolution source-filtre 191
Analyse d'une grande base de données 192
Travail à venir 193

Les études réalisées au cours de cette thèse se sont concentrées sur l'analyse de la qualité vocale de la voix parlée, en tenant compte de la grande variabilité des signaux dans le cas de la parole expressive. Après avoir défini la production vocale chez l'homme et ses propriétés du point de vue perceptif, différentes dimensions de ce qui définit la qualité vocale ont été introduites : relâché-tendu, rauque-clair, doux-fort et la dimension de voisement. Un modèle de la partie voisée a été présenté : le modèle LF [Fant *et al.*, 1985] généralisé par [Doval *et al.*, 2006] sur un jeu de paramètres sans dimensions. Ce modèle comporte des paramètres dont le rôle s'est montré déterminant dans la quantification et la détermination de la qualité vocale, et notamment le quotient ouvert O_q pour la dimension lâché-tendu. D'autres paramètres, comme l'asymétrie ou la phase de retour ont été moins étudiés, et bien que l'intuition prédise qu'ils soient reliés à la dimension doux-fort, les études exhaustives font encore défaut sur ce point. Ces études se concentrent sur les parties voisées (harmoniques) du signal vocal, mais celui-ci comporte aussi une partie non voisée causée par des flux turbulents ou des constrictions lors de la phonation. La détermination du rapport harmonique sur bruit est un bon indicateur de la qualité vocale sur la dimension de voisement [de Krom, 1993].

Suite à ce constat, des méthodes ont été proposées, afin de permettre l'analyse des signaux vocaux et, in fine, de résoudre le problème de l'estimation des paramètres responsables de chacune des quatre dimensions de la qualité vocale : raucité, effort, serrage et voisement[d'Alessandro C., 2006]. Le problème a été séparé en 3 parties, chacune abordant l'analyse des signaux de parole dans un niveau d'analyse plus profond. Ainsi, la première approche a été de segmenter les signaux vocaux : nous avons remarqué alors que beaucoup d'informations propres à la qualité vocale pouvaient être extraites à cette étape. Dans un deuxième temps, la dimension de voisement a été abordée par la présentation et l'amélioration d'une technique de séparation périodique/apériodique. Dans un troisième temps, c'est la configuration de l'appareil de production vocal, et en particulier la forme de l'onde de débit glottique qui a été abordée via une méthode d'estimation conjointe. Enfin, dans un quatrième temps et pour valider les méthodes proposées, elles ont été appliquées à un grand corpus de voix expressive.

Segmentation des signaux vocaux

La segmentation des signaux vocaux par analyse multi-échelles a donné naissance à la méthode LoMA [Tuan et d'Alessandro, 2000] - Lines of Maximum Amplitude. Cette méthode a été reprise au cours de cette thèse afin d'en améliorer l'algorithme et de permettre l'extraction de paramètres supplémentaires. Des tests sur des signaux synthétiques et réels ont montré que cette méthode permettait la mesure de :

1. L'instant de fermeture glottique, donnant ainsi la localisation temporelle du moment où la glotte possède une surface minimale. Cet instant correspond généralement au moment où l'énergie induite dans le conduit vocal est maximale. La méthode LoMA estime l'instant de fermeture glottique avec une précision suffisante pour permettre d'estimer le jitter.
2. L'énergie apportée lors de l'instant de fermeture glottique par cumul des maxima rencontrés sur la LoMA. L'analyse des variations d'énergie d'une période à l'autre donne une idée du shimmer alors que le calcul du centroïde de la distribution d'énergie permet de déterminer le barycentre de la LoMA, qui donne alors une image de la richesse spectrale du signal de source, et donc de la dimension fort-doux de la qualité vocale, comme le montrent des exemples de signaux naturels.
3. Le quotient ouvert pour chaque cycle glottique, par analyse du décalage temporel du premier harmonique par rapport à l'instant de fermeture glottique. Une simulation numérique

de la phase du modèle de source a permis de relier de manière linéaire ce décalage aux valeurs de O_q.

Cette approche innovante montre qu'il est possible d'extraire un grand nombre paramètres des signaux vocaux sans utiliser de décomposition ou déconvolution particulière (bien que le fait même de tracer une LoMA puisse s'apparenter à une déconvolution partielle). Un travail important d'optimisation de l'algorithme permettrait sa diffusion au sein de la communauté scientifique. Des tests approfondis devraient être menés pour tester notamment la pertinence de l'utilisation du barycentre spectrale de la LoMA par rapport au barycentre spectral classique. A la lumière des dernières recherches en matière d'amplitude de voisement [Murty *et al.*, 2009], il devrait être possible de déterminer dans quelle mesure l'amplitude de la LoMA reflète précisément cet aspect des signaux vocaux.

Décomposition harmonique + bruit

Afin de séparer la composante harmonique modélisée par le modèle LF de la partie apériodique, la méthode de décomposition périodique/apériodique [Yegnanarayana *et al.*, 1998] a été reprise. Cette méthode ne tient pas compte de la fréquence fondamentale du signal et pose des problèmes de décomposition sur les phases non stationnaires, pour lesquelles la fréquence fondamentale varie beaucoup.

En s'inspirant des travaux précédents [Jackson et Shadle, 2001], la méthode a été modifiée pour adapter la durée d'observation à la fréquence fondamentale courante. Une étude théorique a permis de déterminer le nombre minimum de périodes à observer pour permettre une décomposition de qualité, l'adaptation de la durée d'observation tenant compte de ce nombre. Pour permettre d'accroitre encore plus cette durée d'observation, une fonction de coût a été proposée pour tenir compte de la dispersion des harmoniques en fréquence.

Cette nouvelle méthode, appelée PAP-A signifiant périodique/apériodique adaptative, a été testée sur une grande base de signaux synthétiques. Ces signaux comportent une variation à la fois du niveau de bruit mais aussi du niveau d'apériodicités structurelles (jitter, shimmer). Les résultats montrent une décomposition de meilleure qualité avec une sensibilité moins importante à la précision de l'estimation préalable de la fréquence fondamentale du signal.

Des analyses de signaux réels ont permis de valider cette méthode, à la fois sur des passages voisés, sur des passages mixtes (consonnes voisées), mais aussi sur des phrases complètes.

Bien que dans le cadre de cette thèse cet algorithme se limite à l'estimation du rapport harmonique sur bruit, la qualité des décomposition obtenues devrait permettre une caractérisation plus fine de la partie apériodique. En matière de qualité vocale, caractériser la partie apériodique en terme de composition spectrale, voire de localisation temporelle plutôt que par sa seule énergie devrait donner des informations capitales notamment pour l'identification des locuteurs.

Pour améliorer encore la qualité de la décomposition, il faudra probablement tenir compte de la redondance d'informations apportée par la transformée de Fourier à court terme (superposition des fenêtres) afin de reconstruire un signal cohérent non seulement sur la durée d'observation, mais aussi par rapport aux données amont et aval. Dans cette optique, les travaux sur la reconstruction itératives de signaux à base de spectrogramme d'amplitude [Griffin et Lim, 1984] font autorité, et la connaissance partiele du spectre non voisé par PAP devrait autoriser la transposition du paradigme de Griffin & Lim à la décomposition voisé/non voisé.

Déconvolution source-filtre

Une fois les signaux décomposés, la composante voisée peut être analysée afin de retrouver la contribution de la source dans le modèle linéaire. La méthode choisie pour opérer la déconvolution

est la méthode ZZT [Bozkurt *et al.*, 2005], basée sur l'évolution de phase différente entre la source (composante anti-causale, possédant des zéros hors du cercle unité) et le filtre (composante causale, possédant ses zéros à l'intérieur du cercle unité). Cette méthode permet d'opérer une déconvolution par séparation des zéros obtenus par calcul de la transformée en Z d'une trame de signal.

Étant donné l'originalité du modèle sur lequel repose la décomposition, il a fallu dans un premier temps confirmer sa capacité à fournir des résultats cohérents sur de grandes variations de qualités vocales. Pour ce faire, et dans l'idée de pouvoir qualifier objectivement les résultats obtenus, des signaux synthétiques ont été générés et analysés par ZZT et par méthode de prédiction linéaire. Les résultats ont montré que cette nouvelle méthode présente une sensibilité moins importante à la variation de paramètres (notamment l'asymétrie α_m) plus importante pour la prédiction linéaire. Il a aussi été montré que les résultats de la décomposition par analyse ZZT sont statistiquement comparables à ceux obtenus par prédiction linéaire.

Dans un deuxième temps, l'utilisation de l'onde de débit glottique estimée par ZZT a permis l'estimation des paramètres du modèle LF sur des signaux réels. Une méthode linéaire basée sur les propriétés spectrales (dérivée de la phase) et temporelles de la source estimée a été présentée, permettant l'estimation conjointe du quotient ouvert O_q et de l'asymétrie α_m. Cette méthode a été testée en deux temps :

- Des voyelles tenues, présentant des qualités vocales différentes ont été analysées dans un effort de réglage de l'algorithme. Une sensibilité au paramètre ρ servant à calculer la dérivée de la phase a été identifiée et l'algorithme a été adapté en y incluant une phase de programmation dynamique.
- Une grande base de données de parole naturelle, composée de textes lus par deux locuteurs (homme et femme) a été analysée. Les résultats ont montré des résultats bons dans le cas où l'estimation est limitée à la détection du quotient ouvert dans une plage d'erreur inférieure à 25%.

Cette méthode a présenté des résultats aux propriétés différentes de l'estimation du quotient ouvert par LoMA. L'idée a donc été proposée de combiner les deux méthodes pour une estimation plus complète et précise du quotient ouvert. Les résultats présentés ont montré une nette amélioration des quantités de détection. Un indicateur de fiabilité de la mesure a été proposé. Ainsi, dans plus de 85% des cas, la détection du quotient ouvert se fait sous 25% d'erreur.

Les tests complémentaires pour valider la méthode ont montré que la pondération choisie n'était pas optimale. Si elle permet toujours d'améliorer les résultats par rapport à chaque méthode isolée, il y a clairement des signaux où l'estimation par LoMA est plus fidèle (et inversement). En outre, l'asymétrie du débit glottique joue probablement aussi un rôle important dans le décalage du premier harmonique utilisé pour estimer le quotient ouvert par LoMA : le développement d'une co-estimation $O_q - \alpha_m$ devrait être possible pour à la fois améliorer la précision sur O_q mais aussi permettre de déterminer avec précision le coefficient α_m, toujours problématique à estimer.

Ce chapitre montre en outre les qualités indéniables de la décomposition par ZZT, basée sur une approche radicalement différente pour l'estimation du filtre inverse permettant d'obtenir le débit glottique. Actuellement déjà utilisée conjointement à la prédiction linéaire dans le cadre d'estimation par contraintes non linéaires, les avancées en matière de cepstre complexe (accélérant considérablement le calcul) devraient permettre le développement de techniques de codages basées sur ce paradigme.

Analyse d'une grande base de données

Les méthodes développées au cours de ce travail ont été utilisées pour analyser une grande base de données de voix parlée expressive. Cette base est composée de plusieurs centaines de fichiers séparés en 16 styles différents et produits par un seul locuteur. Les fichiers ont été analysés et quatre paramètres ont été retenus pour l'analyse : le quotient ouvert (O_q), le barycentre de la LoMA, le jitter et le rapport harmonique sur bruit.

Cette analyse a permis de mettre en lumière plusieurs résultats :

- Dans un premier temps, la cohérence des résultats obtenus pour les différents styles est gage de l'efficacité de la stabilité des méthodes proposées.
- La très faible corrélation entre les résultats des paramètres indique une forte indépendance des quatre paramètres choisis.
- Les styles ont pu être séparés en deux groupes. Un groupe de style possédant une forte composante de qualité vocale et un groupe présentant une qualité vocale normale, pour lesquelles la différence doit principalement jouer sur la prosodie.
- Les interactions source-filtre ont été mises en évidence par l'observation des résultats décomposés par phonèmes voisés.

Cette analyse à grande échelle a permis de confirmer les résultats pressentis au sein de la communauté scientifique sur les rôles que jouent les différents paramètres de production vocale sur la qualité vocale perçue. Ainsi, a-t-on montré que la voix basse, typiquement associée à un rapport harmonique sur bruit faible, a recours à de fortes valeurs de quotient ouvert mais aussi un jitter important. Ces résultats ont été prédits dans des travaux précédents [Michaelis *et al.*, 1998]. De même, les voix fortes, considérées spectralement riches, sont aussi accompagnées d'une distribution caractéristique du jitter et de faibles valeurs du quotient ouvert.

Le travail de ce chapitre se limite cependant à l'analyse de la qualité vocale. Pour aller plus loin, il faudrait lier les résultats obtenus aux analyses prosodiques et aux différents types de représentations disponibles à la fois sur la qualité vocale et sur la prosodie. Certains style ne se démarquent clairement que au niveau de la prosodie, et une analyse conjointe en considérant la qualité vocale comme un trait prosodique à part entière serait bien plus intéressante. De plus, une analyse perceptive de la base de données, donnant des informations comme le niveau de tension, raucité, effort ou voisement permettrait de lier les paramètres estimés sur le signal à leur dimension en matière de qualité vocale.

Travail à venir

En complément des remarques précédentes, cette thèse devrait déboucher scientifiquement sur 3 axes principaux :

- Dans un premier temps, sur l'analyse des signaux vocaux. Des pistes intéressantes ont été explorées lors de cette thèse, et méritent probablement une attention plus importante. Un travail important reste aussi à réaliser pour estimer de manière robuste les paramètres de l'ODG, notamment l'asymétrie. La phase de retour, aujourd'hui uniquement atteinte par son lien avec l'effort, reste un problème majeur et difficile à résoudre du fait de l'action ce paramètre située principalement en hautes fréquences du spectre du débit glottique (noyé dans le bruit de voisement). Arriver à estimer fidèlement ce paramètre de la phase de retour permettrait des transformations de signaux vocaux de bien meilleure qualité dans la dimension d'effort, en particulier dans le sens doux→fort.
- De nombreuses améliorations sont possibles en ce qui concerne la décomposition périodique/apériodique, en tenant compte des avancées récentes en matière de séparation de source et de reconstruction de signaux. Les travaux proposés au cours de cette thèse res-

tent superficiels et n'utilisent qu'une faible partie des possibilités d'une telle décomposition, dont les applications en matière de modification et synthèse vocales devraient être importantes notamment sur la dimension de voisement.

– Finalement, que seraient toutes ces avancées en matière d'analyse et de modification des signaux vocaux sans une analyse du lien entre qualité vocale et expressivité de la voix ? Une analyse approfondie de ce lien s'impose, en ayant recours notamment à des techniques d'analyse-synthèse et de tests perceptifs pour déterminer quelle variation de paramètre est responsable de tel ou tel aspect de la qualité vocale. Sans aucun doute, les variations temporelles de la qualité vocale jouent aussi un rôle prépondérant dans la prosodie : mais ce lien reste encore à établir (changement de mécanisme en fin de phonation dans certaines langues, par exemple).

Références

[Agiomyrgiannakis et Rosec, 2009] Agiomyrgiannakis, Y. et Rosec, O. (2009). Arx-lf-based source-filter methods for voice modification and transformation. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 0:3589–3592.

[Airas *et al.*, 2005] Airas, M., Pulakka, H., Bäckström, T. et Alku, P. (2005). A toolkit for voice inverse filtering and parametrisation. *In : 9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, pages 2145–2148.

[Alku, 1992] Alku, P. (1992). Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11:109–118.

[Alku, 2003] Alku, P. (2003). Parametrisation methods for glottal flow estimation by inverse filtering. *VOQUAL'03*.

[Alku *et al.*, 2005] Alku, P., Airas, M., Bäckström, T. et Pulakka, H. (2005). Using group delay function to assess glottal flows estimated by inverse filtering. *Electronics Letters*, 41 (9):562–563.

[Alku *et al.*, 2006] Alku, P., Airas, M., Björkner, E. et Sundberg, J. (2006). An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity. *J. Acoust. Soc. Am.*, 120(2):1052–1062.

[Alku et Bäckström, 2002] Alku, P. et Bäckström, T. (2002). Normalized amplitude quotient for parametrization of the glottal flow. *J. Acoust. Soc. Am.*, 112 (2):701–710.

[Alku *et al.*, 2009] Alku, P., Magi, C., Yrttiaho, S., Bäckström, T. et Story, B. (2009). Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering. *J. Acoust. Soc. Am.*, 125 (5):3289–3305.

[Allen et Rabiner, 1977] Allen, J. B. et Rabiner, L. R. (1977). A unified approach to short-time fourier analysis and synthesis. *Proc. IEEE*, 65:1558–1564.

[Alteris et Paliwal, 2003] Alteris, L. D. et Paliwal, K. (2003). Usefulness of phase spectrum in human speech perception. *In Eurospeech 2003*, pages 2117–2120.

[Ananthapadmanabha S. et Yegnanarayana, 1979] Ananthapadmanabha S., T. V. et Yegnanarayana, B. (1979). Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE trans. on ASSP*, 27(4):309–318.

[Audibert *et al.*, 2006] Audibert, N., Vincent, D., Auberg'e, V. et Rosec, O. (2006). Expressive speech synthesis : evaluation of a voice quality centered coder on the different acoustic dimensions. *In proceedings of Speech Prosody 2006, Dresden, Germany.*

[Bachu *et al.*, 2010] Bachu, R., Kopparthi, S., Adapa, B. et Barkana, B. (2010). Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy. *In* Elleithy, K., éditeur : *Advanced Techniques in Computing Sciences and Software Engineering*, pages 279–282. Springer Netherlands.

[Badin *et al.*, 1990] BADIN, P., HERTEGARD, S. et KARLOSSON, I. (1990). Notes on the rothenberg mask. *STL-QPSR*, 31:1–7.

[Blackman et Tukey, 1959] BLACKMAN, R. B. et TUKEY, J. W. (1959). *The Measurement of Power Spectra, From the Point of View of Communications Engineering.* New York : Dover.

[Bouzid et Ellouze, 2007] BOUZID, A. et ELLOUZE, N. (2007). Open quotient measurements based on multiscale product of speech signal wavelet transform. *Research Letters in Signal Processing*, 2007.

[Bozkurt, 2005] BOZKURT, B. (2005). *Zeros of the z-transform (ZZT) representation and chirp group delay processing for the analysis of source and filter characteristics of speech signals.* Thèse de doctorat, Faculté Polytechnique de Mons.

[Bozkurt *et al.*, 2007] BOZKURT, B., COUVREUR, L. et DUTOIT, T. (2007). Chirp group delay analysis of speech signals. *Speech Commun.*, 49:159–176.

[Bozkurt *et al.*, 2004a] BOZKURT, B., DOVAL, B., D'ALESSANDRO, C. et DUTOIT, T. (2004a). Appropriate windowing for group delay analysis and roots of z-transform of speech signals. *12th European Signal Processing Conference (EUSIPCO04), Vienna, Austria.*

[Bozkurt *et al.*, 2004b] BOZKURT, B., DOVAL, B., D'ALESSANDRO, C. et DUTOIT, T. (2004b). A method for glottal formant frequency estimation. *In International Conference on Spoken Language Processing (ICSLP).*

[Bozkurt *et al.*, 2005] BOZKURT, B., DOVAL, B., D'ALESSANDRO, C. et DUTOIT, T. (2005). Zeros of z-transform representation with application to source-filter separation in speech. *IEEE signal proc. letter*, 12:344–347.

[Brockmann *et al.*, 2008] BROCKMANN, M., STORCK, C., CARDING, P. et DRINNAN, M. (2008). Voice loudness and gender effects on jitter and shimmer in healthy adults. *J Speech Lang Hear Res.*, 51 (5):1152–1160.

[Brookes, 2000] BROOKES, M. (2000). Voicebox : Speech processing toolbox for matlab. world wide web. *http ://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/.*

[Calliope, 1989] CALLIOPE (1989). *La Parole et son traitement automatique.* Dunod.

[Cheng et O'Shaughnessy, 1989] CHENG, Y. M. et O'SHAUGHNESSY, D. (1989). Automatic and reliable estimation of glottal closure instant and period. *IEEE Trans. on ASSP*, 37(12):1805–1814.

[Childers et Lee, 1991] CHILDERS, D. et LEE, C. (1991). Vocal quality factors : Analysis, synthesis, and perception. *J. Acoust. Soc. Am.*, 90 (5):2394–2410.

[Childers *et al.*, 1977] CHILDERS, D. G., SKINNER, D. P. et KEMERAIT, R. C. (1977). The cepstrum : A guide to processing. *Proceedings of the IEEE*, pages 1428–1443.

[C.Sapienza *et al.*, 1998] C.SAPIENZA, E.STATHOPOULOS et C.DROMEY (1998). Approximations of open quotient and speed quotient from glottal airflow and egg waveforms : Effects of measurement criteria and sound pressure level. *Journal of Voice*, 12 (1):31–43.

[d'Alessandro *et al.*, 1998] D'ALESSANDRO, C., DARSINOS, V. et YEGNANARAYANA, B. (1998). Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources. *IEEE Trans. on Speech and Audio Processing*, 6(1):12–23.

[d'Alessandro et Sturmel,] D'ALESSANDRO, C. et STURMEL, N. Glottal closure instants and voice source analysis using time-scale lines of maximum amplitude. *submitted to sadhana.*

[D'Alessandro *et al.*, 2007] D'ALESSANDRO, N., WOODRUFF, P., FABRE, Y., DUTOIT, T., LE BEUX, S., DOVAL, B. et D'ALESANDRO, C. (2007). Realtime and accurate musical control of expression in singing synthesis. *Journal on Multimodal User Interfaces*, 1(1):31–39.

[d'Alessandro C., 2006] d'Alessandro C. (2006). Voice source parameters and prosodic analysis. *In* S. SUDHOFF, D. et AL., éditeurs : *Method in Empirical Prosody Research*, pages 63–87. Walter de Gruyter, Berlin, New York.

[Dalsgaard *et al.*, 2008] DALSGAARD, P., PEDERSEN, C. F., ANDERSEN, O. et YEGNANARAYANA, B. (2008). Using zeros of the z-transform in the analysis of speech signals. *In ISCA ITRW Speech Analysis and Processing for Knowledge Discovery, june 4-6, 2008, Aalborg, Denmark, 4 pages.*

[de Cheveigné et Kawahara, 2002] de CHEVEIGNÉ, A. et KAWAHARA, H. (2002). Yin, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.*, 111(4):1917–1930.

[de Krom, 1993] de KROM, G. (1993). A cepstrum based technique for determining a harmonics-to-noise ratio in speech signals. *J. Speech Hear. Res.*, 36(2):254–266.

[Degottex *et al.*, 2010] DEGOTTEX, G., ROEBEL, A. et RODET, X. (2010). Phase minimization for glottal model estimation. *IEEE trans. on Acoust. Sp. and Lang. Proc.*, page in press.

[Deller, 1981] DELLER, J. (1981). Some notes on closed phase glottal inverse filtering. *IEEE Trans. on acoustics, speech and language processing*, 29:917–919.

[Delprat, 1992] DELPRAT, N. (1992). *Analyse temps-fréquence des sons musicaux : exploration d'une nouvelle méthode d'extraction de données pertinentes pour un modèle de synthèse.* Thèse de doctorat, Université Aix-Marseilles II.

[Doval *et al.*, 2003] DOVAL, B., D'ALESSANDRO, C. et HENRICH, N. (2003). The voice source as a causal/anticausal linear filter. *In VOQUAL'03 ISCA Workshop, Geneva*, pages 6–10.

[Doval *et al.*, 2006] DOVAL, B., D'ALESSANDRO, C. et HENRICH, N. (2006). The spectrum of glottal flow models. *Acta Acustica united with Acustica*, 92:1026–1046.

[Drugman *et al.*, 2009a] DRUGMAN, T., BOZKURT, B. et DUTOIT, T. (2009a). Chirp decomposition of speech signals for glottal source estimation. *In ISCA workshop on Non-LInear Speech Processing, 4 pages.*

[Drugman *et al.*, 2009b] DRUGMAN, T., BOZKURT, B. et DUTOIT, T. (2009b). Complex cepstrum-based decomposition of speech for glottal source estimation. *In Interspeech09, Brighton, U.K, 4 pages.*

[Drugman *et al.*, 2008] DRUGMAN, T., DUBUISSON, T., D'ALESSANDRO, N., MOINET, A. et DUTOIT, T. (2008). Voice source parameters estimation by fitting the glottal formant and the inverse filtering open phase. *In EUSIPCO'08, Lausanne, Switzerland, 4 pages.*

[El-Jaroudi et Makhoul, 1991] EL-JAROUDI, A. et MAKHOUL, J. (1991). Discrete all-pole modeling. *IEEE Trans. on signal processing*, 39:411–423.

[Fabre, 1957] FABRE, P. (1957). Un procédé électrique percutané d'inscription de l'accolement glottique au cours de la phonation. *Bull. Nat. Med.*

[Fant, 1960] FANT, G. (1960). *Acoustic Theory of Speech Production.* Mouton, The Hague, Netherlands.

[Fant, 1995] FANT, G. (1995). The lf-model revisited. *STL-QPSR*, 2-3:119–156.

[Fant, 1997] FANT, G. (1997). The voice source in connected speech. *Speech Communication*, 22:125–139.

[Fant *et al.*, 1985] FANT, G., LILJENCRANTS., J. et LIN, Q. (1985). A four-parameter model of glottal flow. *STL-QPSR*, 4:1–13.

[Fletcher et Munson, 1933] FLETCHER, H. et MUNSON, W. (1933). Loudness, its definition, measurement and calculation. *J. Acoust. Soc. Am.*, 5:82–108.

[Fujisaki et Ljungqvist, 1987] FUJISAKI, H. et LJUNGQVIST, M. (1987). Estimation of voice source and vocal tract parameters based on arma analysis and a model for the glottal source waveform. *In ICASSP'87*, volume 12, pages 637–640.

[Gauffin et Sundberg, 1989] GAUFFIN, J. et SUNDBERG, J. (1989). Spectral correlates of glottal voice source waveform characteristics. *J. Speech Hear. Res.*, 32:556–565.

[Griffin et Lim, 1984] GRIFFIN, D. et LIM, J. (1984). Signal reconstruction from short-time fourier transform magnitude. *IEEE Trans. Acoust., Speech, and Signal Proc.*, 32(2):236–243.

[Guruprasad et Yegnanarayana, 2009] GURUPRASAD, S. et YEGNANARAYANA, B. (2009). Perceived loudness of speech based on the characteristics of glottal excitation source. *J. Acoust. Soc. Am.*, 126 (1):2061–2071.

[Hanson, 1994] HANSON, H. (1994). Acoustic correlates of glottal characteristics of female speakers. *J. Acoust. Soc. Am.*, 95:2922–2923.

[Hanson, 1995] HANSON, H. M. (1995). *Glottal Characteristics of Female Speakers.* Thèse de doctorat, Harvard University.

[Harris, 1978] HARRIS, J. F. (1978). On the use of windows for harmonic analysis with the discrete fourier transform. *proceedings of the IEEE*, 66.

[Henrich, 2001] HENRICH, N. (2001). *Etude de la source glottique en voix parlée et chantée.* Thèse de doctorat, Université Paris VI.

[Henrich *et al.*, 2001] HENRICH, N., D'ALESSANDRO, C. et DOVAL, B. (2001). Spectral correlates of voice open quotient and glottal flow asymmetry : theory, limits and experimental data. *In Eurospeech 2001, Aalborg, Denmark.*

[Henrich *et al.*, 2004] HENRICH, N., D'ALESSANDRO, C., DOVAL, B. et CASTELLENGO, M. (2004). On the use of derivative electroglottographic signals for characterization of nonpatological phonation. *J. Acoust. Soc. Am.*, 115 (3):1321–1332.

[Henrich *et al.*, 2003] HENRICH, N., SUNDIN, G., AMBROISE, D., D'ALESSANDRO, C., DOVAL, B. et CASTELLENGO, M. (2003). Just noticeable differences of open quotient and asymmetry coefficient in singing voice. *Journal of Voice*, 17:481–494.

[Jackson et Shadle, 2001] JACKSON, P. J. B. et SHADLE, C. H. (2001). Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech. *IEEE trans. Speech Audio Process.*, 9 (7):713–726.

[Jensen *et al.*, 1999] JENSEN, J., JENSEN, S. et HANSEN, E. (1999). Exponential sinusoidal modeling of transitional speech segments. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 1:473–476.

[Kadambe et Boudreaux-Bartels, 1992] KADAMBE, S. et BOUDREAUX-BARTELS, G. (1992). Application of the wavelet transform for pitch detection of speech signals. *IEEE trans. on IT*, 38(2):917–924.

[Karakozoglou *et al.*, 2010] KARAKOZOGLOU, S.-Z., HENRICH, N., D'ALESSANDRO, C. et STYLIANOU, Y. (2010). Automatic glottal segmentation using local-based active contours. *In 9th conference on Advances in Quantitative Lanryngology (AQL).*

[Kelly et Lochbaum, 1962] KELLY, J. L. et LOCHBAUM, C. C. (1962). Speech synthesis. *In Fourth Int. Congr. Acoust., Paper G42*, page pp. 1 ?4.

[Kendall, 2009] KENDALL, K. A. (2009). High-speed laryngeal imaging compared with videostroboscopy in healthy subjects. *Arch Otolaryngol Head Neck Surg.*, 135(3):274–281.

[Kim et Hahn, 2007] Kim, S.-J. et Hahn, M. (2007). Two-band excitation for hmm-based speech synthesis. *IEICE - Trans. Inf. Syst.*, E90-D:378–381.

[Klatt et Klatt, 1990] Klatt, D. et Klatt, L. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am.*, 87:820–857.

[Kounoudes, 2001] Kounoudes, A. (2001). *Epoch Estimation for Closed-Phase Analysis of Speech.* Thèse de doctorat, Imperial College.

[Kounoudes *et al.*, 2002] Kounoudes, A., Naylor, P. A. et Brookes, M. (2002). The dypsa algorithm for estimation of glottal closure instants in voiced speech. *In proceedings of ICASSP'2002*, pages 349–352.

[Kreiman et Gerratt, 2010] Kreiman, J. et Gerratt, B. R. (2010). Perceptual Assessment of Voice Quality : Past, Present, and Future. *Perspectives on Voice and Voice Disorders*, 20(2):62–67.

[Kreiman *et al.*, 2004] Kreiman, J., Vanlancker-Sidtis, D. et Gerratt, B. (2004). Defining and measuring voice quality. *From Sound to Science*, June 11 - June 13:C–163 to C–168.

[Krishnamurthy et Childers, 1986] Krishnamurthy, A. K. et Childers, D. G. (1986). Two-channel speech analysis. *IEEE Trans. on acoustics, speech and signal processing*, 34:730–743.

[Laprie et Mathieu, 1998] Laprie, Y. et Mathieu, B. (1998). A variational approach for estimating vocal tract. shapes from the speech signal. *In proceedings of ICASSP'98*, pages 929–932.

[Lehto *et al.*, 2007] Lehto, L., Airas, M., Bjorkner, E., Sundberg, J. et Alku, P. (2007). Comparison of two inverse filtering methods in parameterization of the glottal closing phase characteristics in different phonation types. *Journal of Voice*, 21:138–150.

[Makhoul, 1975] Makhoul, J. (1975). Linear prediction : A tutorial review. *Proc. IEEE*, 63(5): 561–580.

[Mallat et Hwang, 1992] Mallat, S. et Hwang, W. L. (1992). Singularity detection and processing with wavelets. *IEEE trans. on IT*, 38(2):617–643.

[Markel et Gray, 1976] Markel, J. D. et Gray, A. H. J. (1976). *Linear Prediction of Speech.* Springer Verlag, Berlin.

[Mcadams, 1999] Mcadams, S. (1999). Perspectives on the contribution of timbre to musical structure. *Comput. Music J.*, 23:85–102.

[McAulay et Quatieri, 1986] McAulay, R. et Quatieri, T. (1986). Speech analysis synthesis based on a sinusoidal representation. *IEEE Trans.*, 34(4):744–754.

[Meike *et al.*, 2010] Meike, B., Michael J., D., Claudio, S. et Paul, N. C. (2010). Reliable jitter and shimmer measurements in voice clinics : The relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task. *Journal of Voice*, In Press, Corrected Proof:–.

[Michaelis *et al.*, 1998] Michaelis, D., Fröhlich, M., Strube, H., Kruse, E., Story, B. et Titze, I. (1998). Some simulations concerning jitter and shimmer measurement. *In Advances in Quantitative Laryngoscopy, Voice and Speech Research (Proc. 3rd International Workshop, June 19-20 1998)*, pages 71–80.

[Michaelis *et al.*, 1997] Michaelis, D., Gramss, T. et Strube, H. (1997). Glottal-to-noise excitation ratio ? a new measure for describing pathological voices. *acta acustica*, 83:700–706.

[Moores et Torres, 2008] Moores, E. et Torres, J. (2008). A performance assessment of objective measures for evaluating the quality of glottal waveform estimation. *Speech Communication*, 50:56–66.

[Moulines et R., 1990] MOULINES, E. et R., D. F. (1990). Detection of the glottal closure by jumps in the statistical properties of the speech signal. *Speech Communication*, 9(5/6):401–418.

[Murthy *et al.*, 1989] MURTHY, H., MADHU MURTHY, K. et YAGNANARAYANA, B. (1989). Formant extraction from phase using weighted group delay function. *Electronics Letters*, 25:1609–1611.

[Murthy et Yegnanarayana, 1991] MURTHY, H. A. et YEGNANARAYANA, B. (1991). Formant extraction from group delay function. *Speech Commun.*, 10(3):209–221.

[Murty *et al.*, 2009] MURTY, K., YEGNANARAYANA, B. et JOSEPH, M. (2009). Characterization of glottal activity from speech signals. *Signal Processing Letters, IEEE*, 16(6):469–472.

[Naylor *et al.*, 2007] NAYLOR, P., KOUNOUDES, A., GUDNASON, J. et BROOKES, M. (2007). Estimation of the glottal closure instant using the dypsa algorithm. *IEEE Trans. on acoustics, speech and language processing*, 15:34–46.

[Nordstrom et Driessen, 2006] NORDSTROM, K. I. et DRIESSEN, P. F. (2006). variable pre-emphasis lpc for modeling vocal effort in the singing voice. *In Proc. of the 9th Int. Conference on Digital Audio Effects (DAFx-06), Montreal, Canada, September 18-20.*

[Paliwal, 1984] PALIWAL, K. K. (1984). Effect of preemphasis on vowel recognition performance. *Speech Communication*, 3(1):101–106.

[Pfitzinger, 2006] PFITZINGER, H. R. (paper KN2, 2006). Five dimensions of prosody : intensity, intonation, timing, voice quality, and degree of reduction. *In Speech Prosody 2006.*

[Plante *et al.*, 1994] PLANTE, F., MEYER, G. et AINSWORTH, W. (1994). A pitch extraction reference database. *In 4th european conference on speech communication and technology, madrid, spain.*

[Quatieri, 2001] QUATIERI, T. F. (2001). *Discrete-Time Speech Signal Processing : Principles and Practice.* Prentice Hall.

[Rabiner et Schafer, 1978] RABINER, L. R. et SCHAFER, R. W. (1978). *Digital processing of speech signals / Lawrence R. Rabiner, Ronald W. Schafer.* Prentice-Hall, Englewood Cliffs, N.J. :.

[Riegelsberger et Krishnamurthy, 1993] RIEGELSBERGER, E. et KRISHNAMURTHY, A. K. (1993). Glottal source estimation : Methods of applying the lf-model to inverse filtering. *In Proc. Int. Conf. on Acoustic Speech Signal Process*, pages 542–545.

[Rosenberg, 1971] ROSENBERG, A. E. (1971). Effect of glottal pulse shape on the quality of natural vowels. *J. Acous. Soc. Am.*, 49:583–590.

[Rothenberg, 1973] ROTHENBERG, M. (1973). A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. *J. Acoust. Soc. Am.*, 53 (6):1632–1645.

[Rothenberg, 1977] ROTHENBERG, M. (1977). Measurement of airflow in speech. *J. Speech Hear. Res.*, 20 (6):155–176.

[Rothenberg, 1992] ROTHENBERG, M. (1992). A multichannel electroglottograph. *Journal of Voice*, 6(1):36–43.

[Roy et Kailath, 1989] ROY, R. et KAILATH, T. (1989). Esprit - estimation of paramaters via rotational invariance technique. *IEEE trans. on Acoust. Speech and Sig. Proc.*, 37:984–995.

[Scherer, 2003] SCHERER, K, R. (2003). Vocal communication of emotion : A review of research paradigms. *Speech Communication*, 40:227–256.

[Sekey et Hanson, 1984] SEKEY, A. et HANSON, B. A. (1984). Improved 1-bark bandwidth auditory filter. *J. Acoust. Soc. Am.*, 75:1902–1904.

[Serra et Julius Smith, 1990] Serra, X. et Julius Smith, I. (1990). Spectral modeling synthesis : A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14:12–24.

[Smith et Abel, 1999] Smith, J. O. et Abel, J. S. (1999). Bark and erb bilinear transforms. *IEEE Transactions on Speech and Audio Processing*, 7:697–708.

[Smits et Yegnanarayana, 1995] Smits, R. et Yegnanarayana, B. (1995). Determination of instants of significant excitation on speech using group delay function. *IEEE trans. on SAP*, 3(5):325–333.

[Steiglitz et Dickinson, 1977] Steiglitz, K. et Dickinson, B. (1977). Computation of the complex cepstrum by factorization of the z-transform. *Proc. ICASSP77*, 2:723–726.

[Sturmel et d'Alessandro, 2010] Sturmel, N. et d'Alessandro, C. (2010). Glottal parameters estimation on speech using the zeros of the z transform. *In Interspeech 2010, 4 pages.*

[Sturmel *et al.*, 2006] Sturmel, N., d'Alessandro, C. et Doval, B. (2006). A spectral method for estimation of the voice speed quotient and evaluation using electroglottography. *In 7th Conference on Advances in Quantitative Laryngology, Groningen, 4 pages.*

[Sturmel *et al.*, 2007] Sturmel, N., d'Alessandro, C. et Doval, B. (2007). A comparative evaluation of the zeros of z transform representation for voice source estimation. *Interspeech07.*

[Sturmel *et al.*, 2009] Sturmel, N., d'Alessandro, C. et Rigaud, F. (2009). Glottal closure instant detection using lines of maximum amplitudes (loma) of the wavelet transform. *In ICASSP'09, 4 pages.*

[Stylianou, 1996] Stylianou, Y. (1996). *Harmonic plus Noise Models for Speech combined with Statistical Methods for Speech and Speaker Modification.* Thèse de doctorat, Telecom Paris.

[Sundberg, 1994] Sundberg, J. (1994). Vocal fold vibration patterns and phonatory modes. *STL-QPSR*, 35:69–80.

[Sundberg et Nordström, 1976] Sundberg, J. et Nordström, P.-E. (1976). Raised and lowered larynx - the effect on vowel formant frequencies. *STL-QPSR*, 17:35–39.

[Thomas *et al.*, 2009] Thomas, M. R. P., Gudnasson, J. et Naylor, P. A. (2009). Detection of glottal closing and opening instants using an improved dypsa framework. *In EUSIPCO 2009, Glasgow, Scotland*, pages 2191–2195.

[Traunmüller et Eriksson, 2000] Traunmüller, H. et Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *J. Acoust. Soc. Am.*, 107(6):3438–3451.

[Tuan et d'Alessandro, 1999] Tuan, V. N. et d'Alessandro, C. (1999). Robust glottal closure detection using the wavelet transform. *In European Conference on Speech Technology, Eurospeech*, pages 2805–2808, Budapest.

[Tuan et d'Alessandro, 2000] Tuan, V. N. et d'Alessandro, C. (2000). Glottal closure detection using egg and the wavelet transform. *in Proceedings 4th International Workshop on Advances in Quantitative Laryngoscopy, Voice and Speech Research, Jena*, pages 147–154.

[van Dinther *et al.*, 2005] van Dinther, R., Veldhuis, R. et Kohlraush, A. (2005). Perceptual aspects of glottal-pulse paramater variations. *Speech Communication*, 46 (1):95–112.

[Vasilakis et Stylianou, 2009] Vasilakis, M. et Stylianou, Y. (2009). Spectral jitter modeling and estimation. *Biomedical Signal Processing and Control*, 4(3):183 – 193.

[Veldhuis, 1998] Veldhuis, R. (1998). A computationally efficient alternative for the liljencrants-fant model and its perceptual evaluation. *J. Acous. Soc. Am.*, 103:566–571.

[Vincent *et al.*, 2005] VINCENT, D., ROSEC, O. et CHONAVEL, T. (2005). Estimation of lf glottal source parameters based on an arx model. *In proceedings of Interspeech'2005, Lisboa, Portugal*, pages 333–336.

[Vincent *et al.*, 2006] VINCENT, D., ROSEC, O. et CHONAVEL, T. (2006). Glottal closure instant estimation using an appropriateness measure of the source and continuity constraints. *In proceedings of ICASSP'2006*, volume 1, pages 381–384.

[Vincent *et al.*, 2007] VINCENT, D., ROSEC, O. et CHOVANEL, T. (2007). A new method for speech synthesis and transformation based on an arx-lf source-filter decomposition and hnm modeling. *In proceedings of ICASSP'07, Honolulu)*, pages 525–528.

[Werner, 1983] WERNER, W. (1983). A generalized companion matrix of a polynomial and some applications. *Linear algebra appl.*, 55:19–36.

[Winckel, 1954] WINCKEL (1954). Scientific appraisal of singing voice. *Nature*, 173, 574.

[Wong *et al.*, 1979] WONG, D., MARKEL, J. et GRAY, A. H. J. (1979). Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Trans. on Speech and Audio Processing*, 35:350–355.

[Yegnanarayana *et al.*, 1998] YEGNANARAYANA, B., D'ALESSANDRO, C. et DARSINOS, V. (1998). An interactive algorithm for decomposition of speech signals into periodic and aperiodic components. *IEEE Trans. on Speech and Audio Processing*, 6(1):1–11.

[Yule, 1927] YULE, G. U. (1927). On a method of investigating periodicities in disturbed series, with special reference to wolfer ?s sunspot numbers. *Phil. Trans. Roy. Soc.*, pages 267–298.

[Zubrycki et Petrovsky, 2010] ZUBRYCKI, P. et PETROVSKY, A. (2010). Quasi-periodic signal analysis using harmonic transform with application to voiced speech processing. *In Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pages 2374 –2377.

Quatrième partie

Annexes

Annexe A

Analyses complémentaires du grand corpus de parole naturelle et expressive

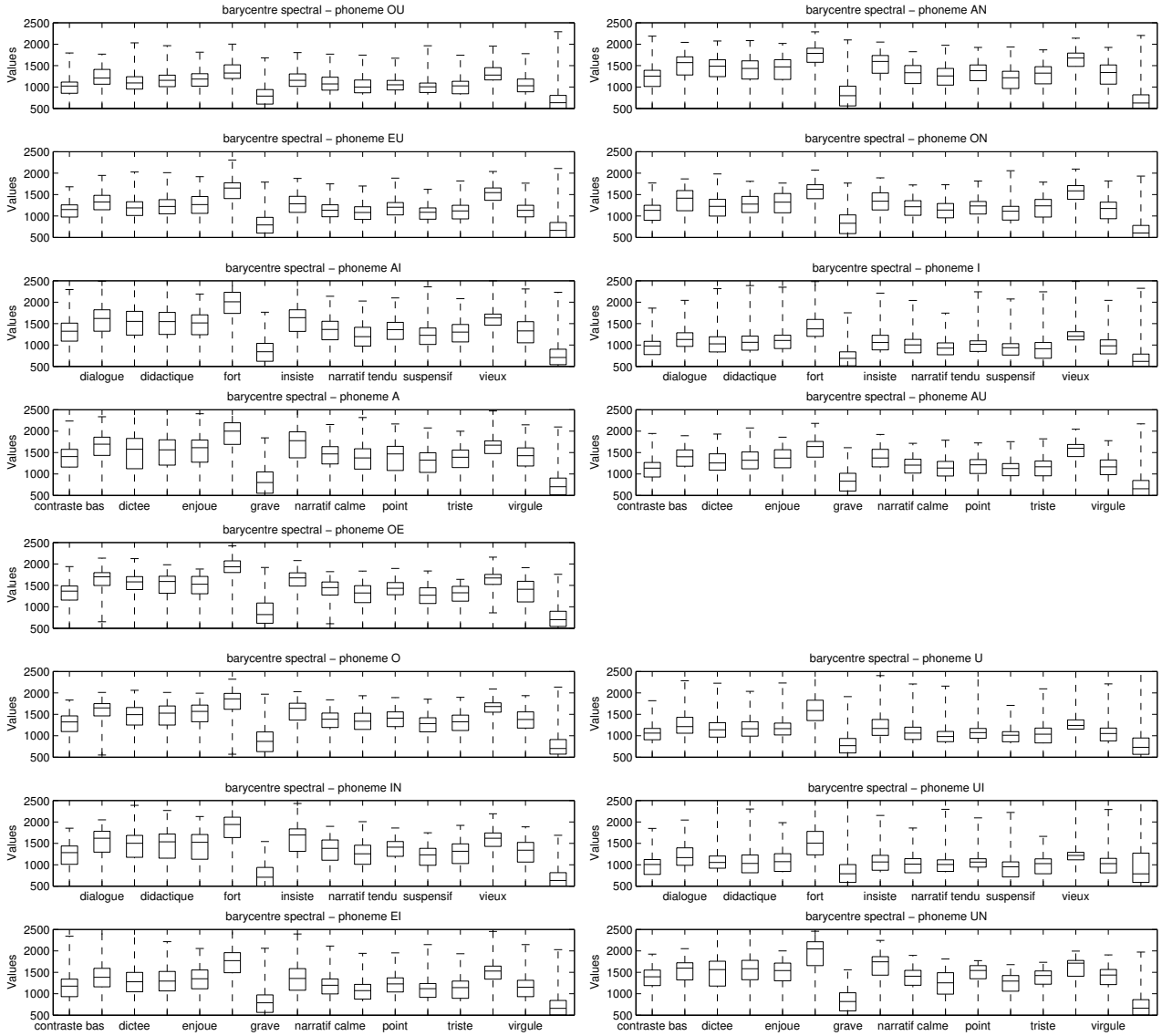


Figure A.1 – Tendance par phonèmes, pour les valeurs mesurées du barycentre de la LoMA.

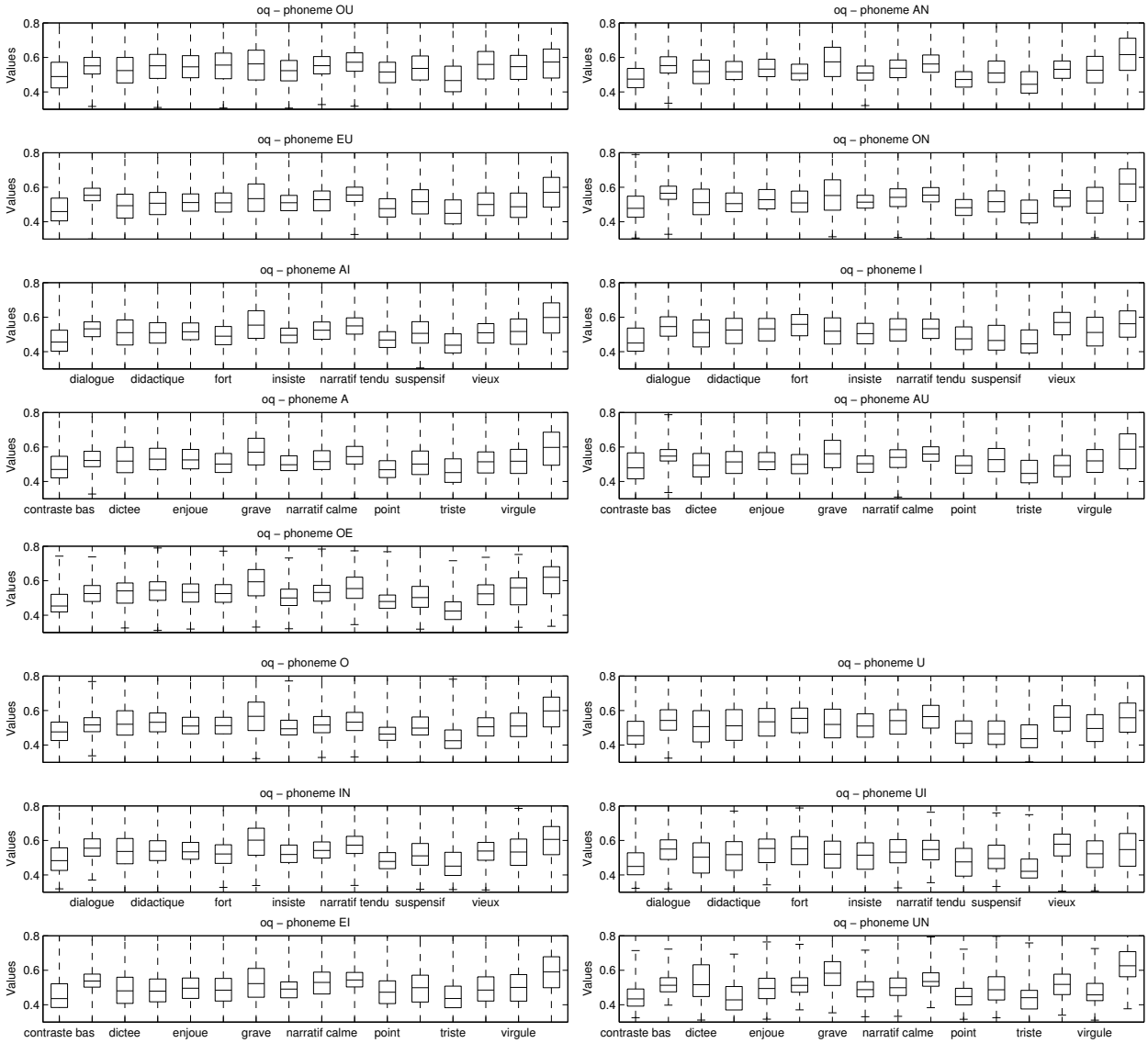


Figure A.2 – Tendance par phonèmes, pour les valeurs mesurées du quotient ouvert.

Figure A.3 – Tendance par phonèmes, pour les valeurs mesurées du jitter.

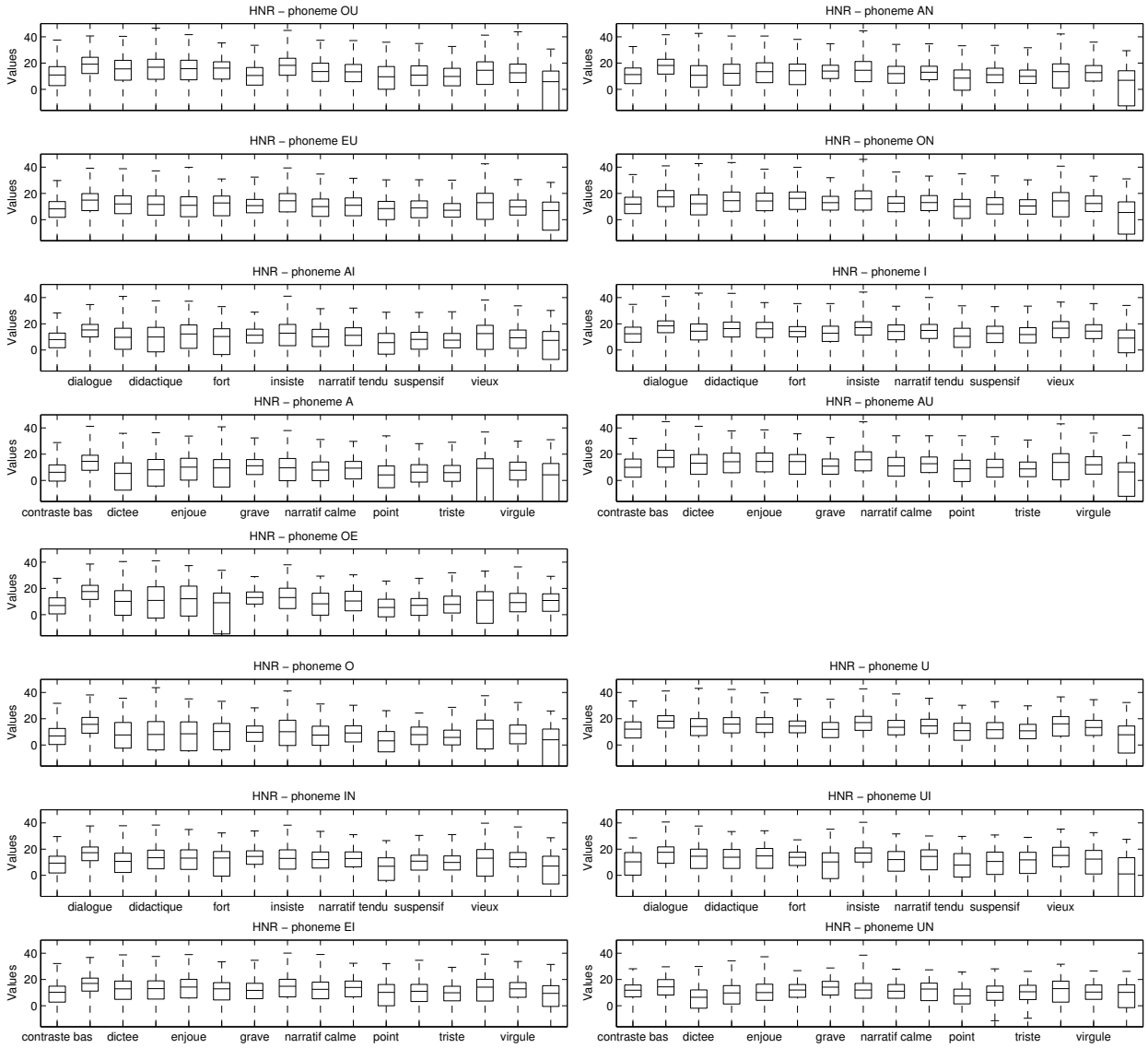


Figure A.4 – Tendance par phonèmes, pour les valeurs mesurées du rapport harmonique sur bruit.

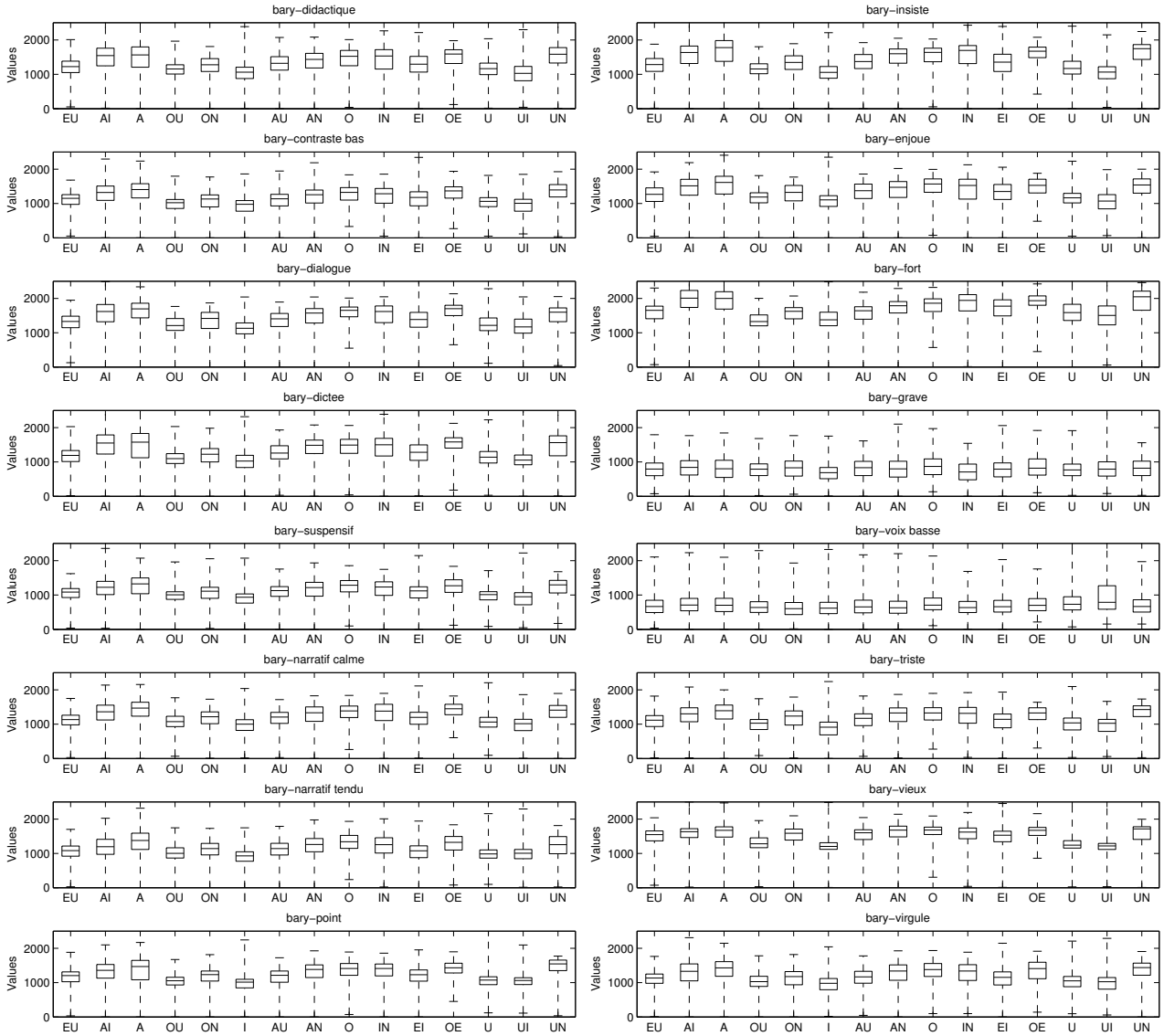


Figure A.5 – Tendance par styles, pour les valeurs mesurées du barycentre de la LoMA.

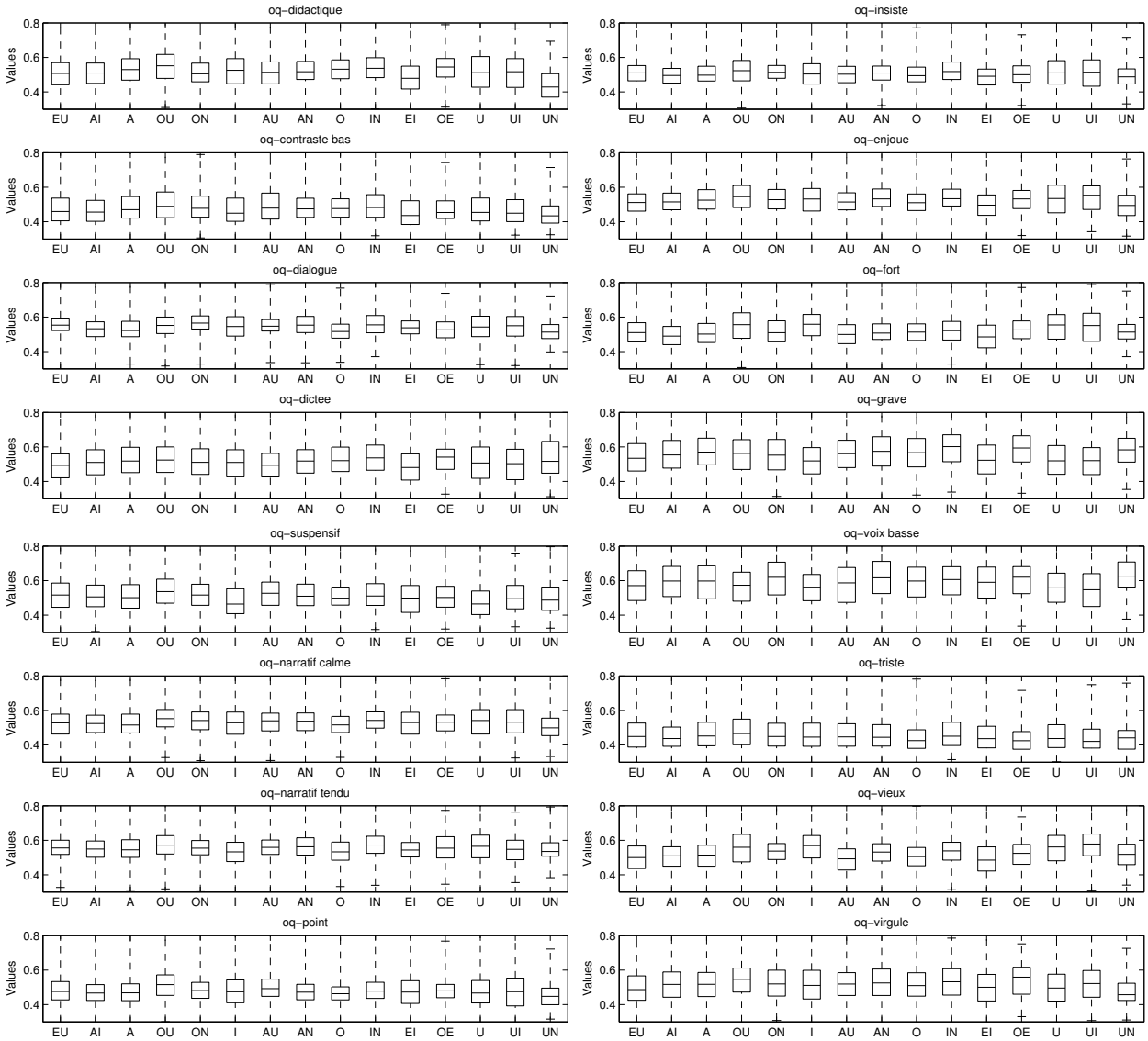


Figure A.6 – Tendance par styles, pour les valeurs mesurées du quotient ouvert.

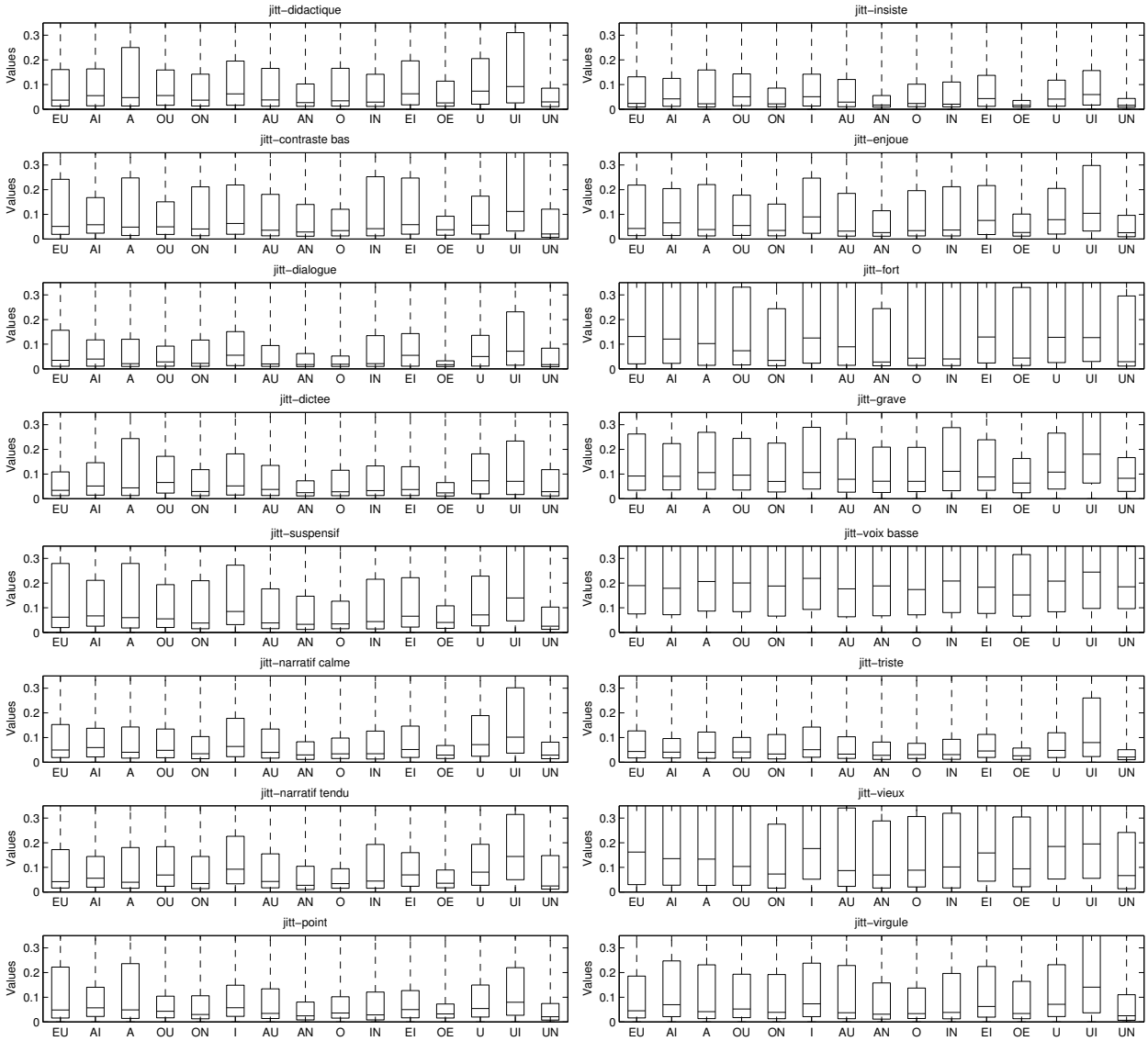


Figure A.7 – Tendance par styles, pour les valeurs mesurées du jitter.

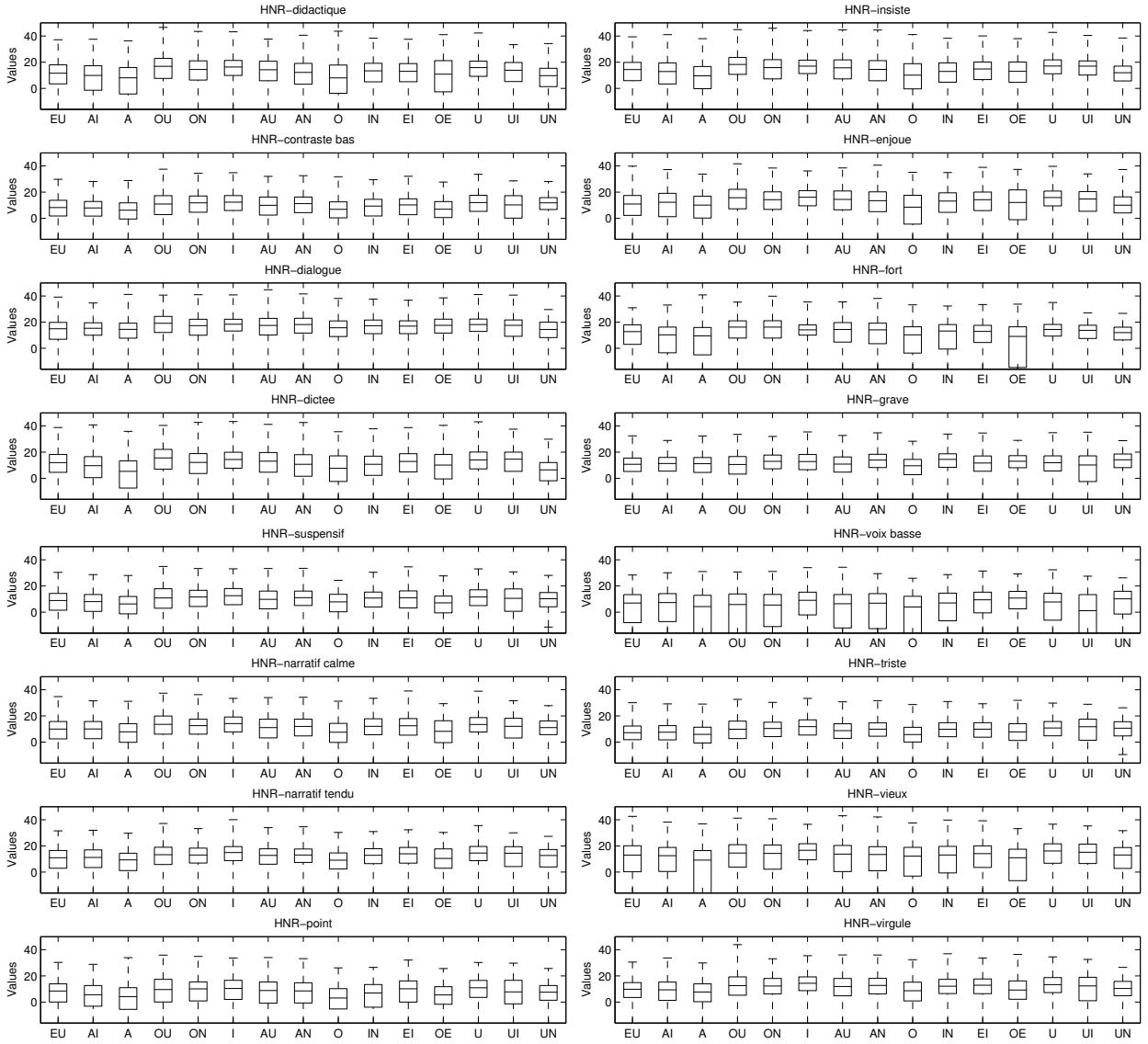


Figure A.8 – Tendance par styles, pour les valeurs mesurées du rapport harmonique sur bruit.

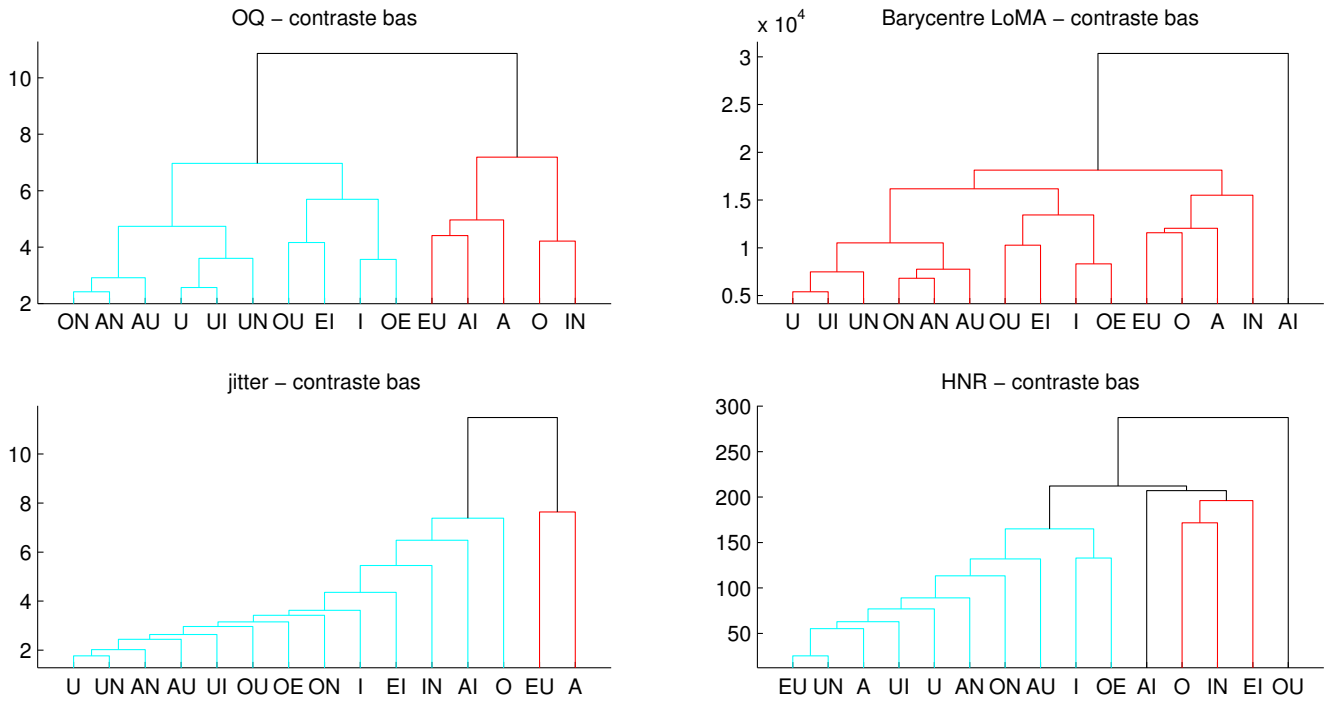


Figure A.9 – Hiérarchie des voyelles pour le style *contraste bas*

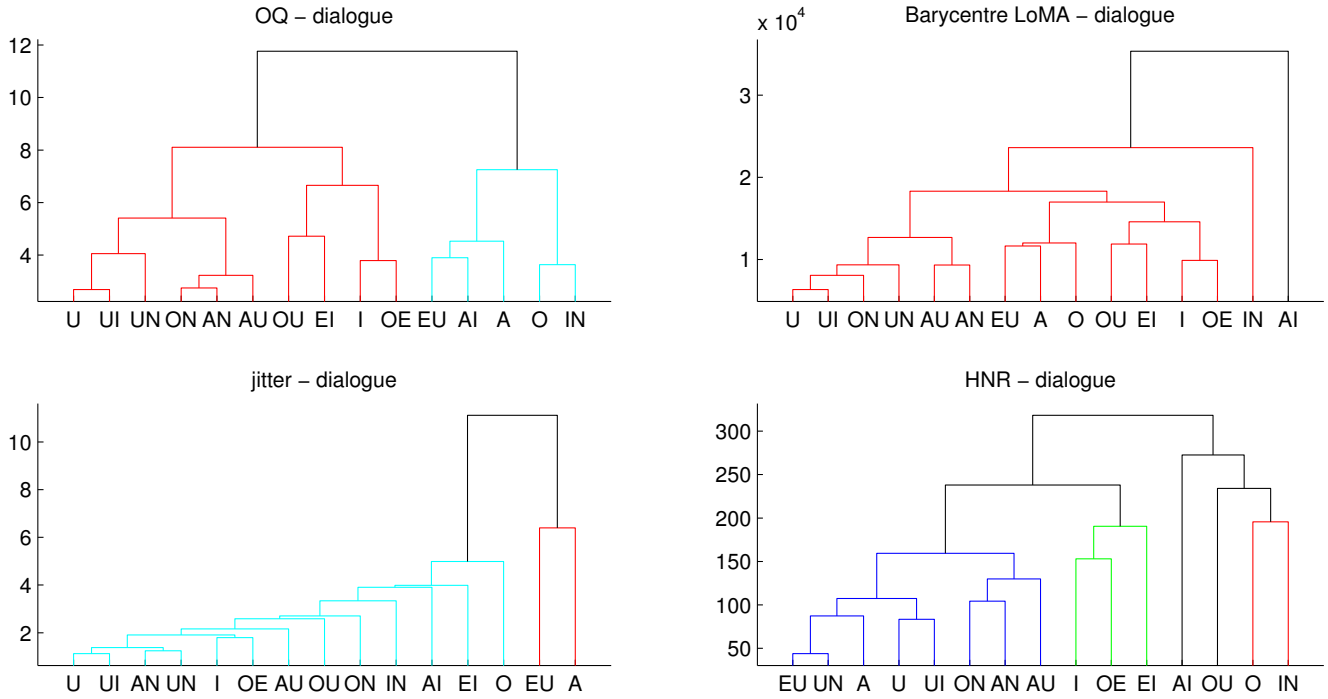


Figure A.10 – Hiérarchie des voyelles pour le style *dialogue*

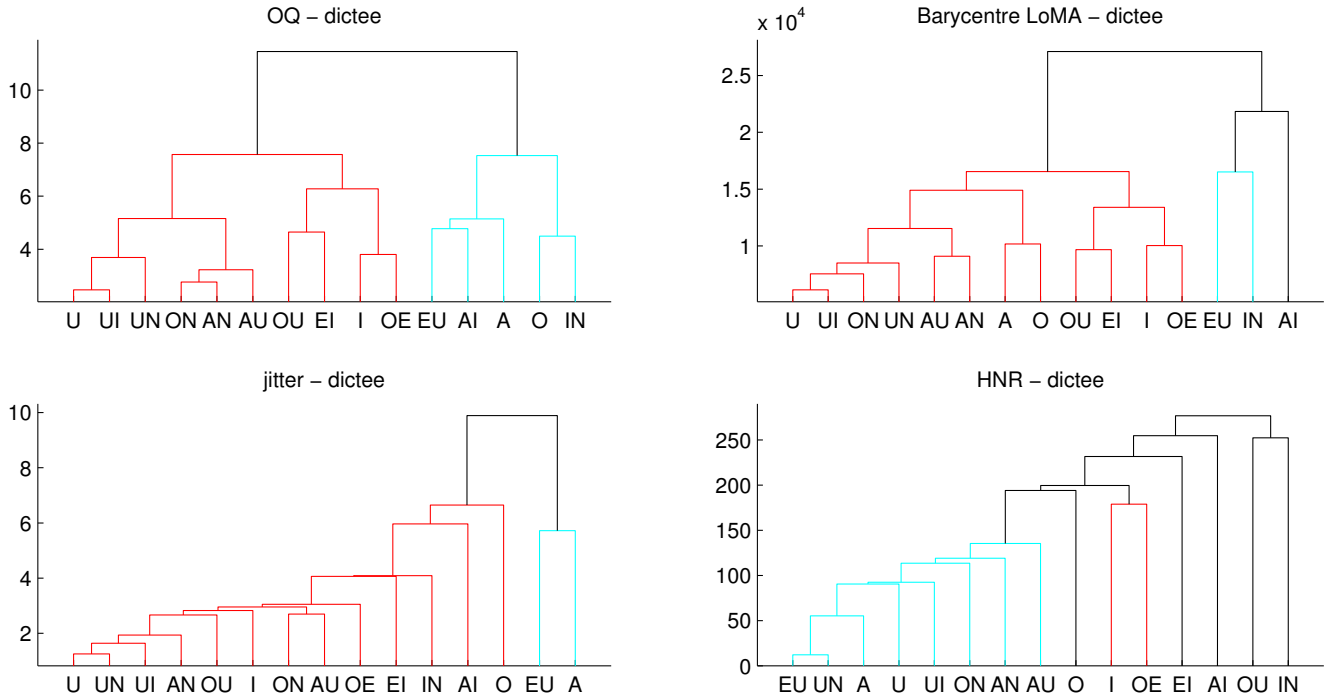


Figure A.11 – Hiérarchie des voyelles pour le style *dictée*

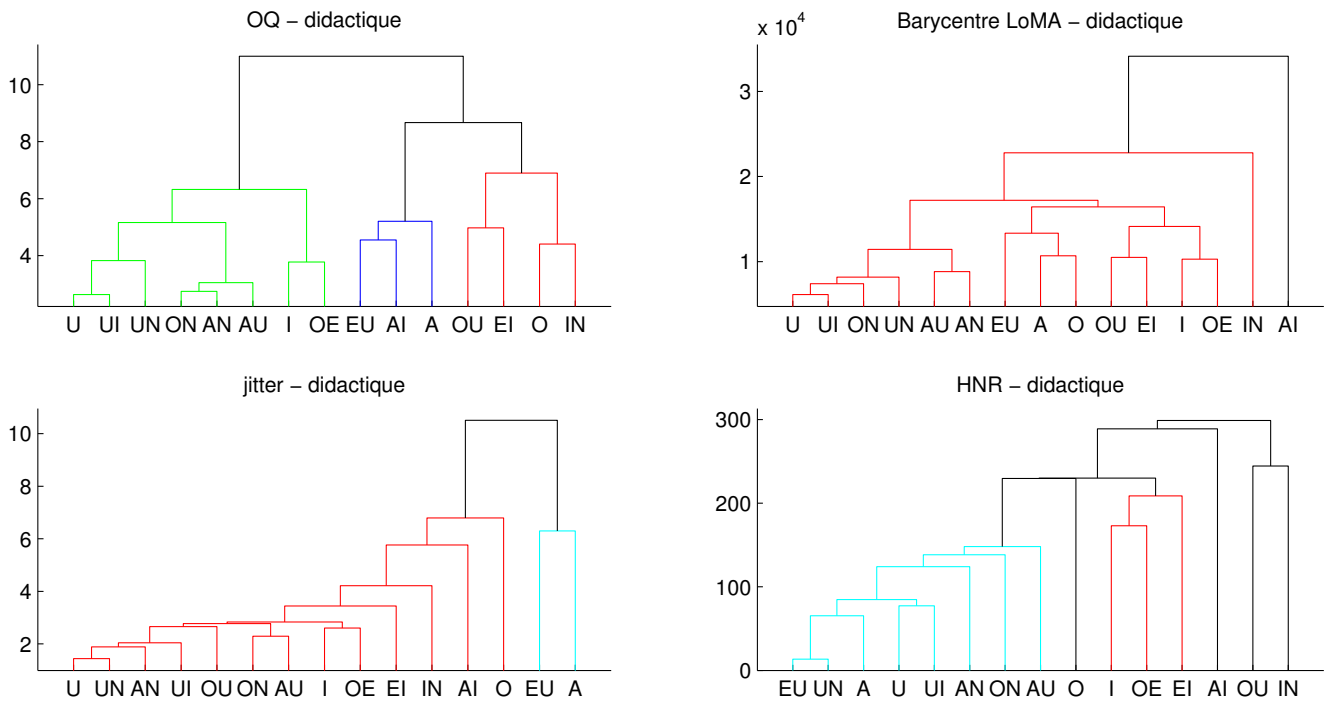


Figure A.12 – Hiérarchie des voyelles pour le style *didactique*

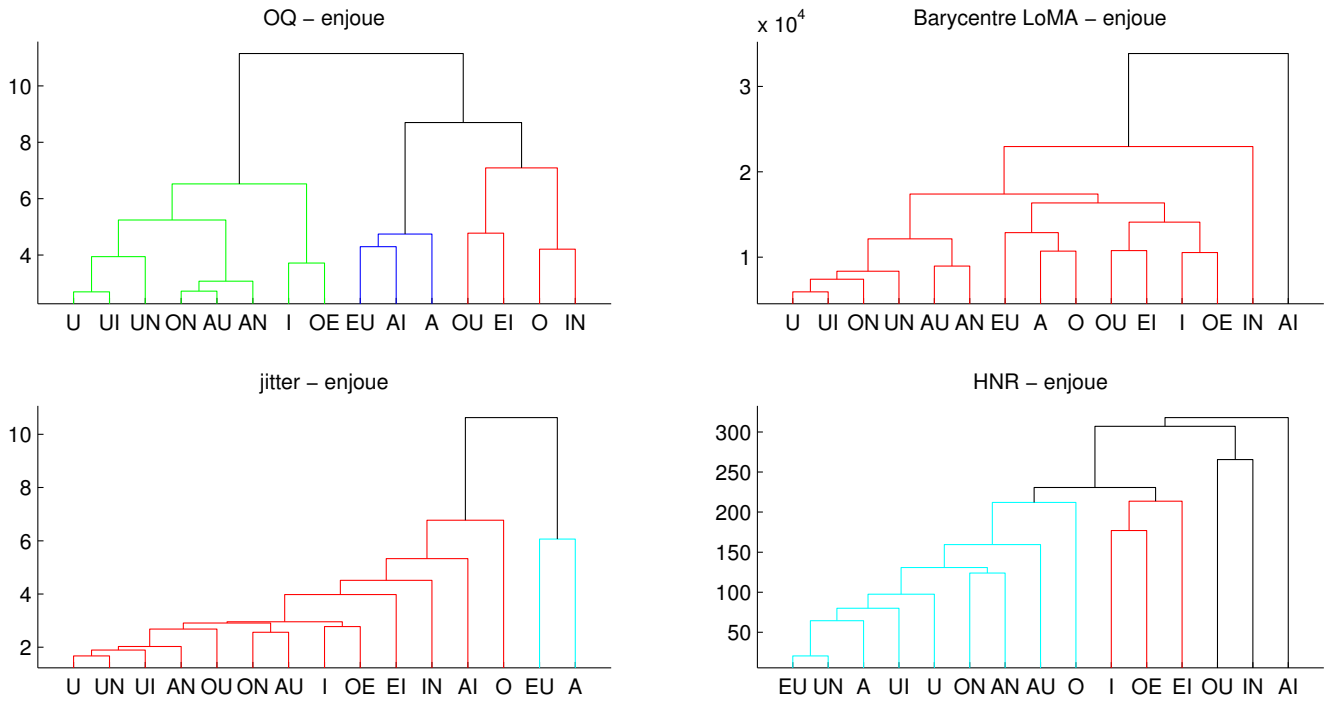


Figure A.13 – Hiérarchie des voyelles pour le style *enjoué*

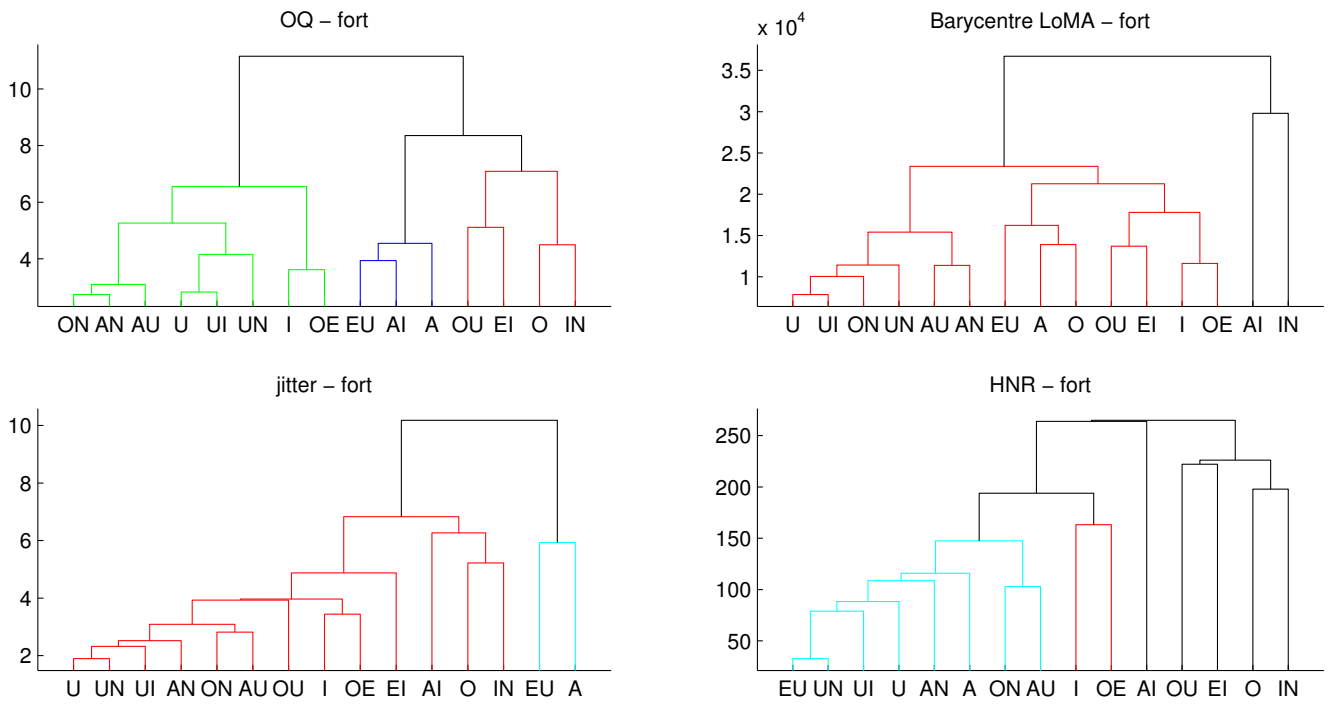


Figure A.14 – Hiérarchie des voyelles pour le style *fort*

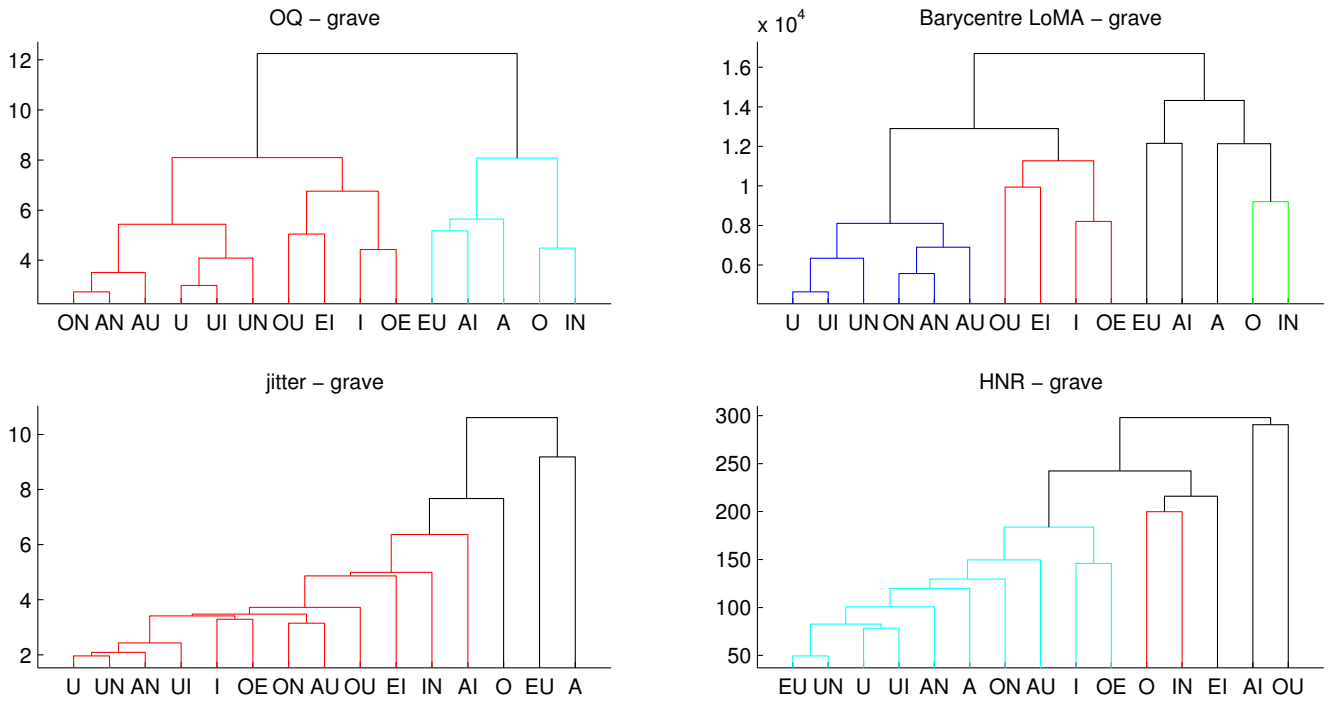


Figure A.15 – Hiérarchie des voyelles pour le style *grave*

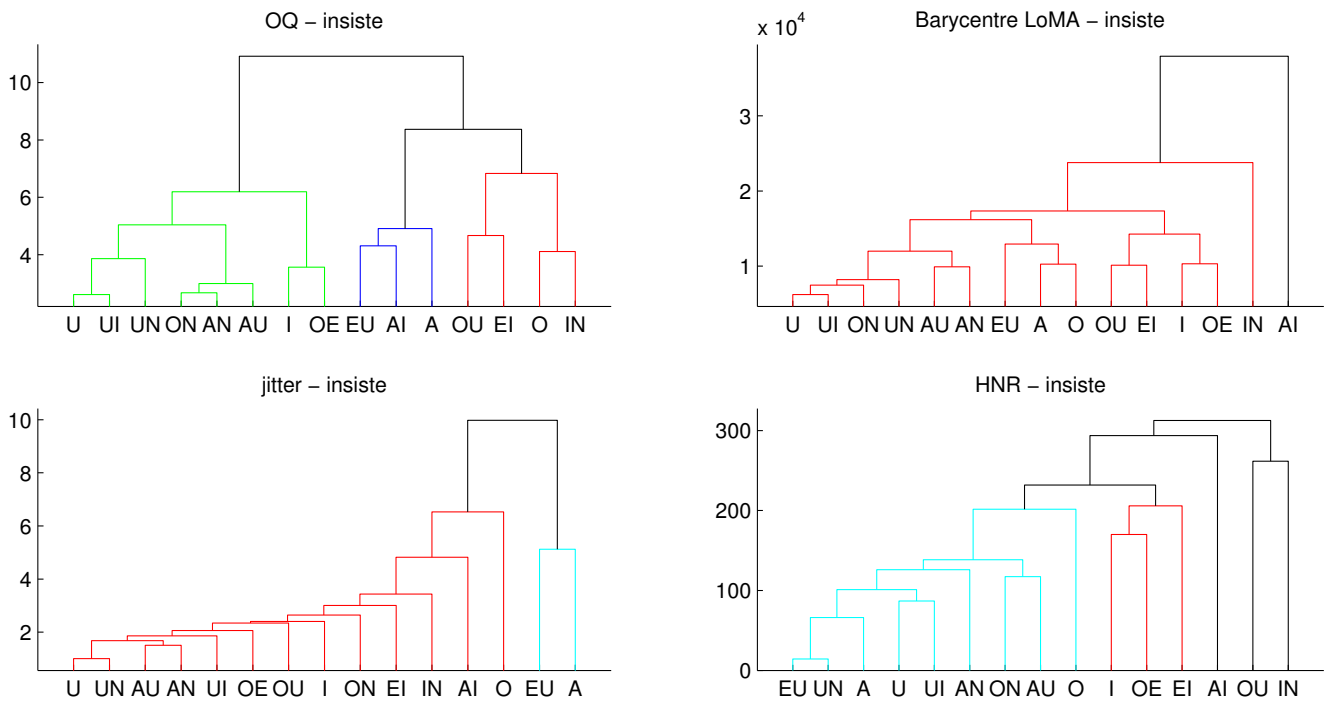


Figure A.16 – Hiérarchie des voyelles pour le style *insiste*

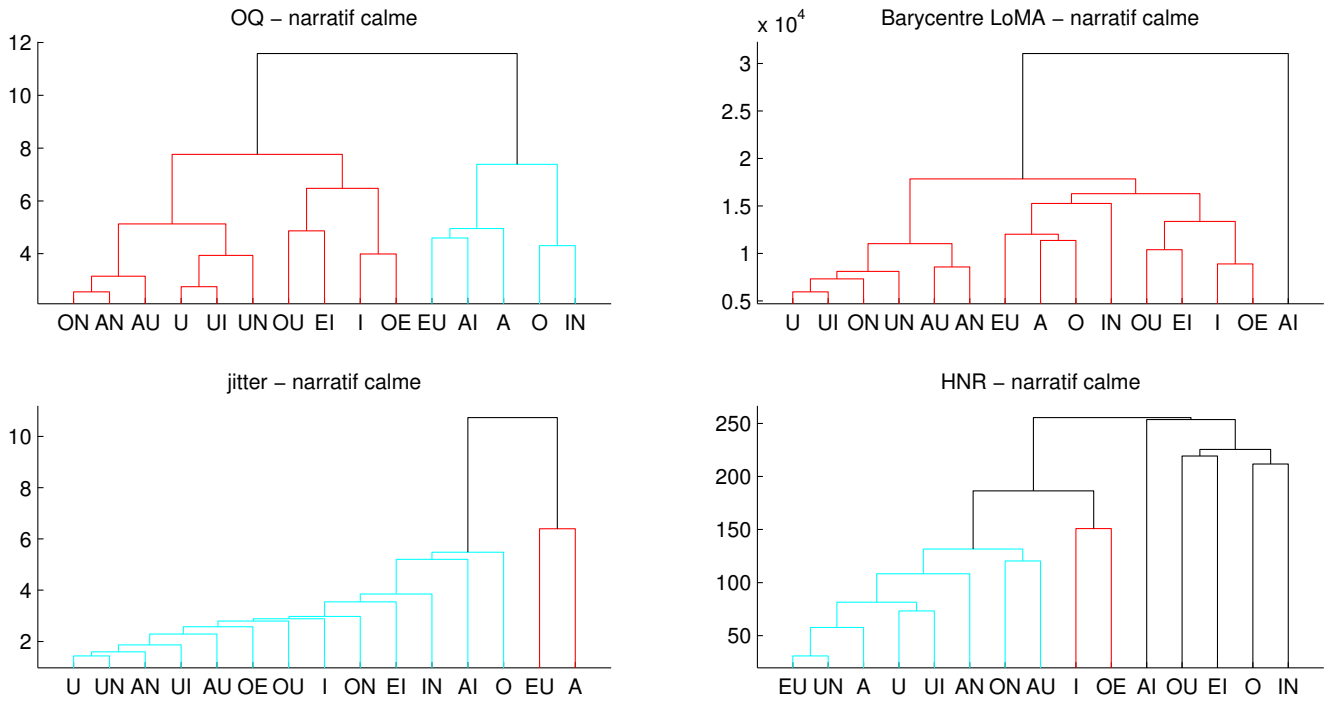


Figure A.17 – Hiérarchie des voyelles pour le style *narratif calme*

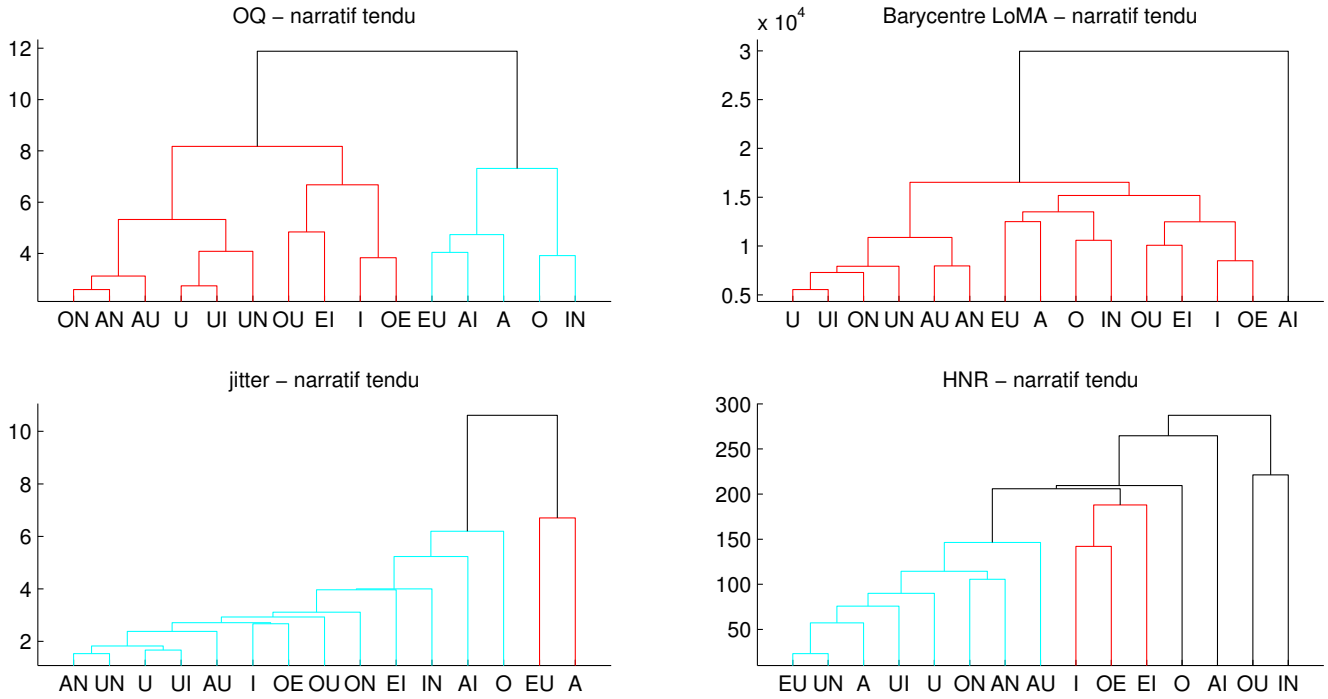


Figure A.18 – Hiérarchie des voyelles pour le style *narratif tendu*

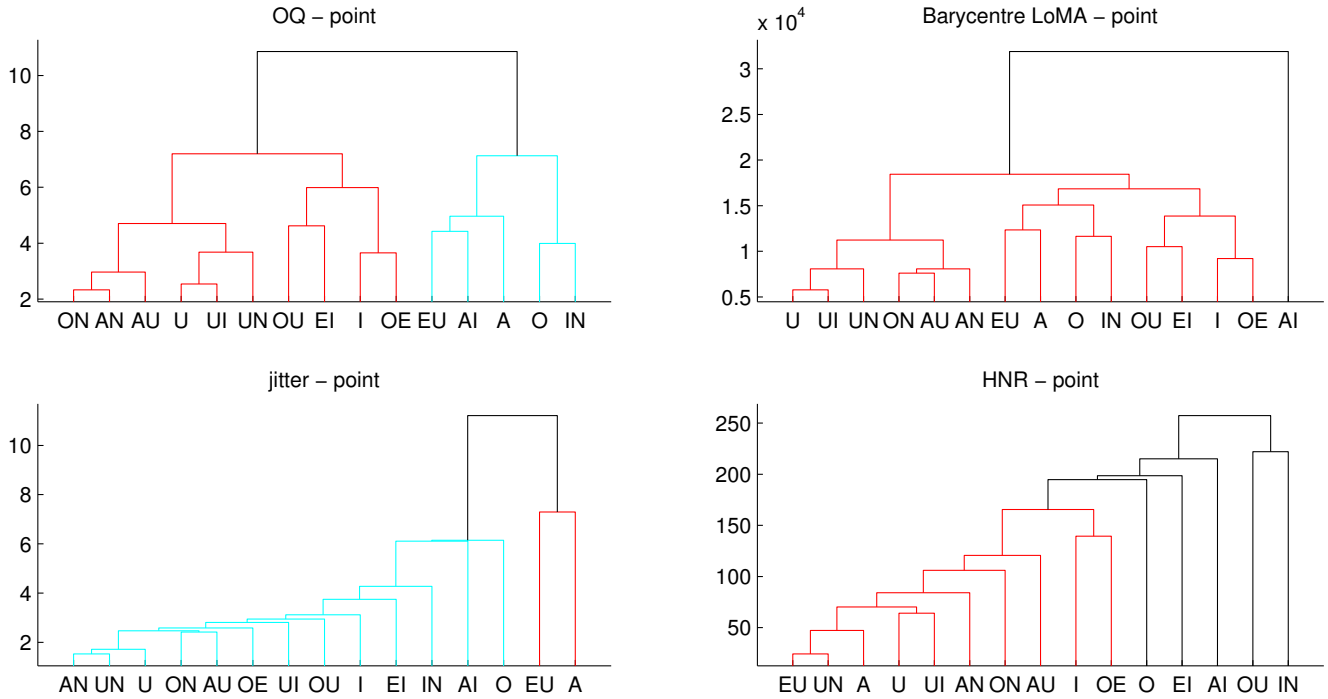


Figure A.19 – Hiérarchie des voyelles pour le style *point*

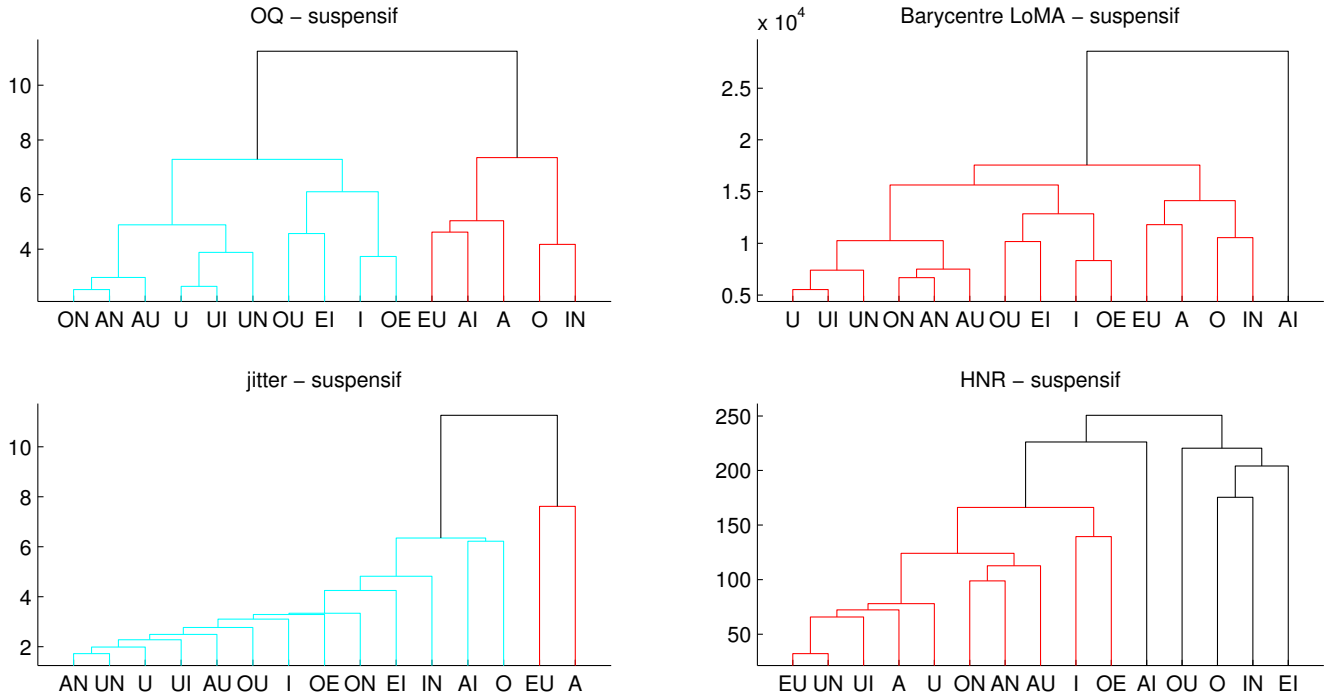


Figure A.20 – Hiérarchie des voyelles pour le style *suspensif*

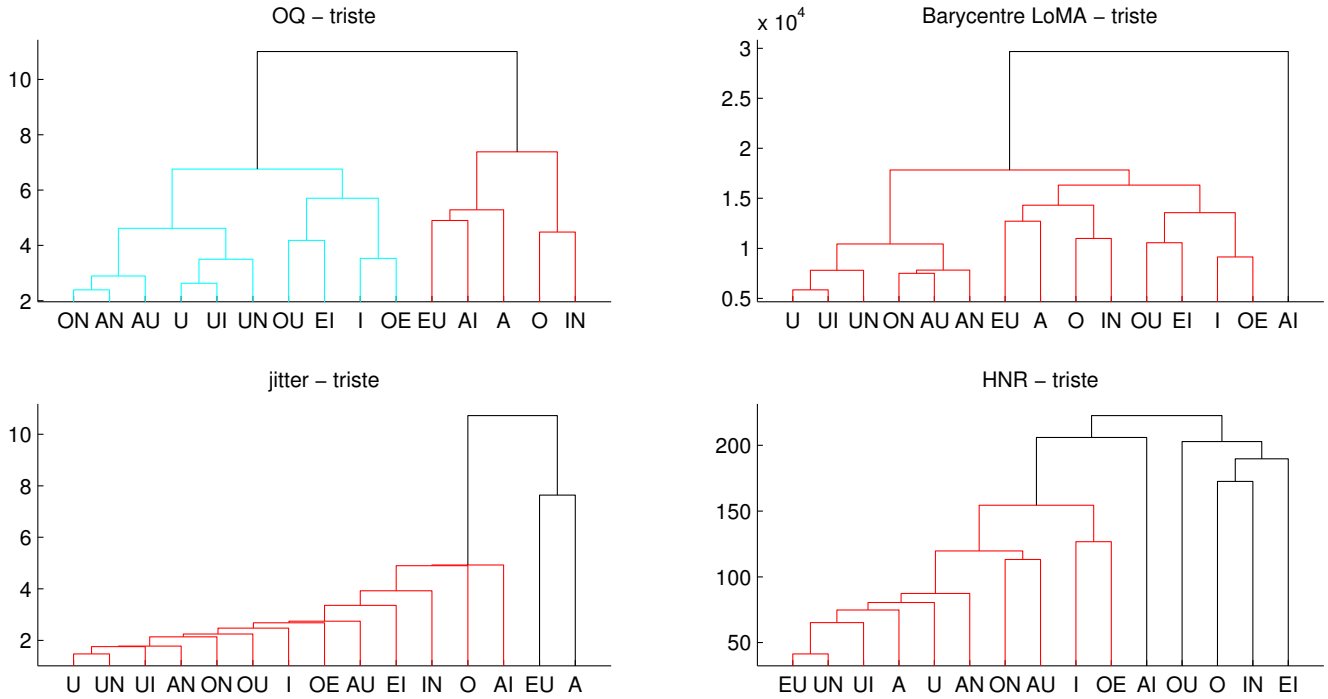


Figure A.21 – Hiérarchie des voyelles pour le style *narratif*

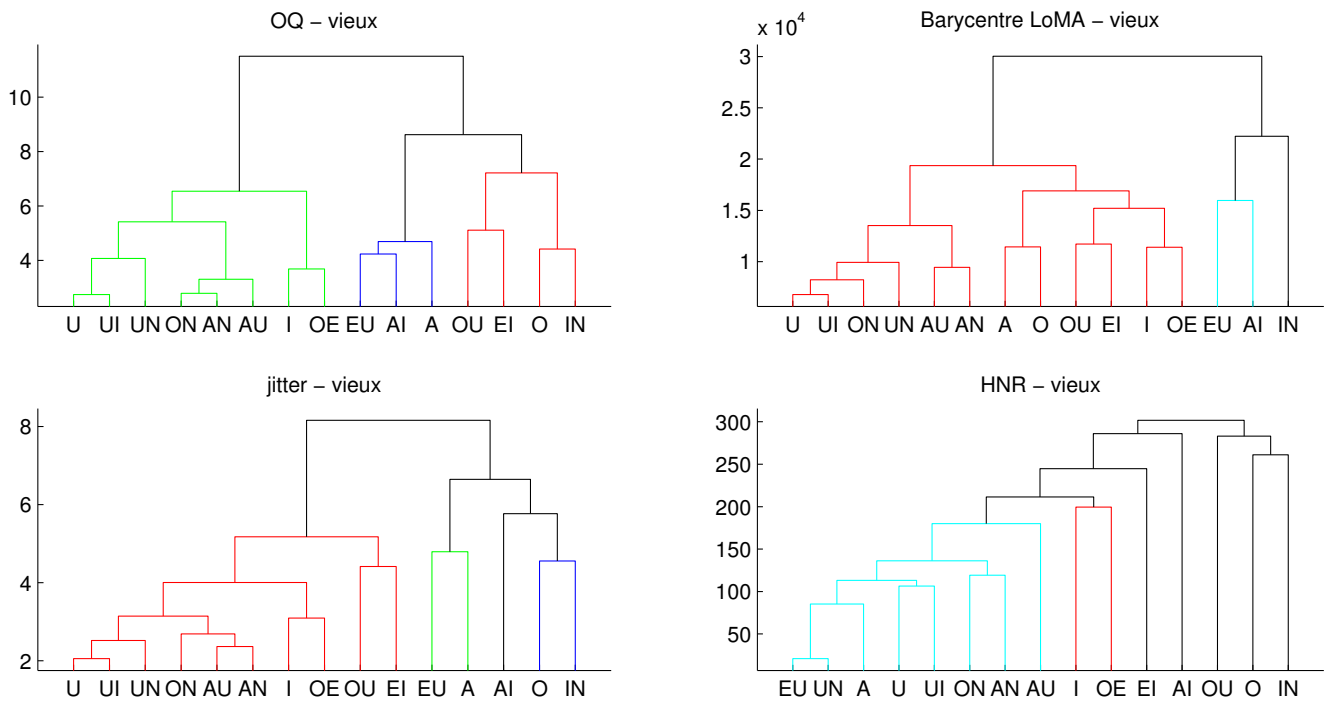


Figure A.22 – Hiérarchie des voyelles pour le style *vieux*

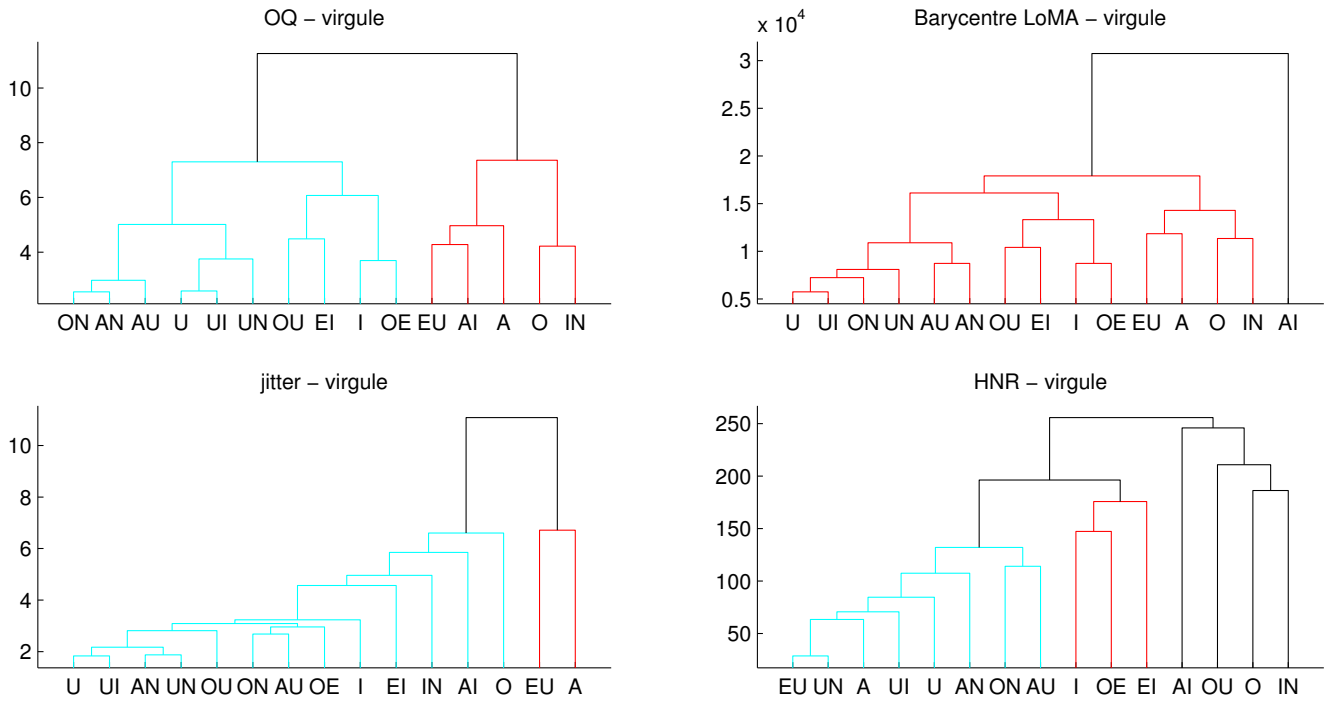


Figure A.23 – Hiérarchie des voyelles pour le style *virgule*

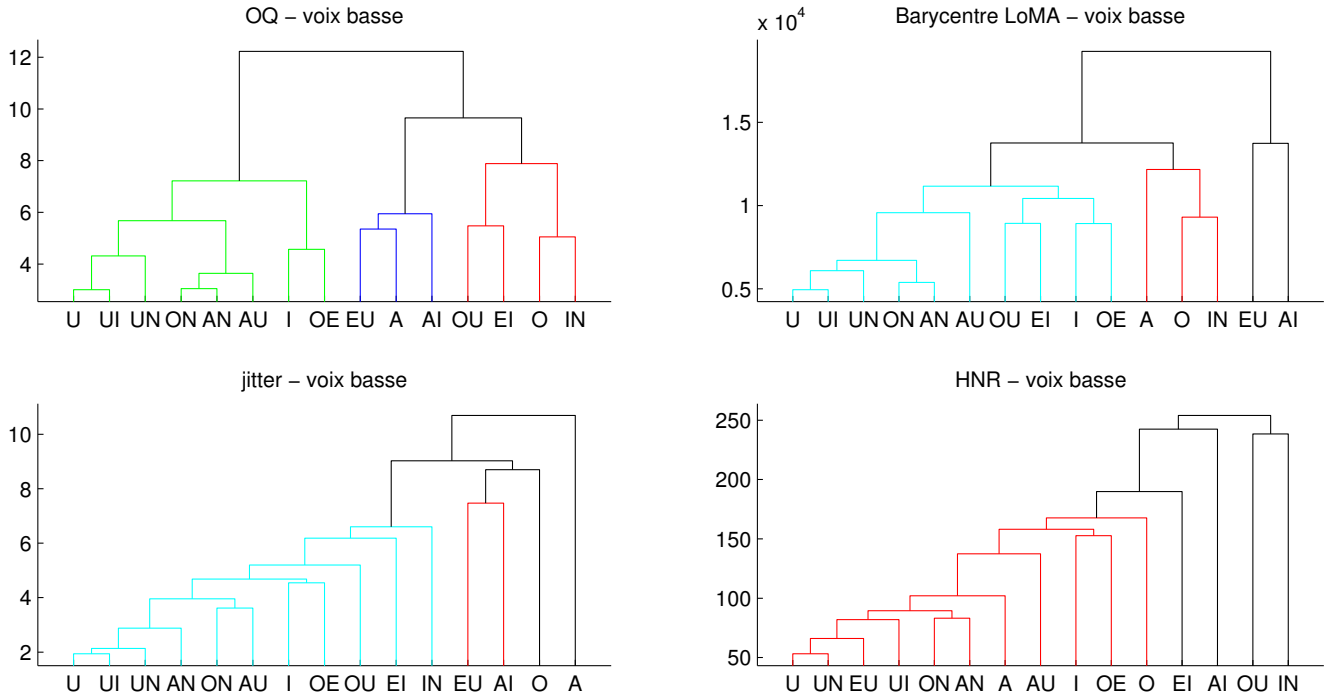


Figure A.24 – Hiérarchie des voyelles pour le style *voix basse*