



**HAL**  
open science

# Vers un système indiquant la distance d'un locuteur par transformation de sa voix

Thibaut Fux

► **To cite this version:**

Thibaut Fux. Vers un système indiquant la distance d'un locuteur par transformation de sa voix. Acoustique [physics.class-ph]. Université de Grenoble, 2012. Français. NNT : 2012GRENT120 . tel-01557936

**HAL Id: tel-01557936**

**<https://theses.hal.science/tel-01557936>**

Submitted on 6 Jul 2017

**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.

## THÈSE

Pour obtenir le grade de

## DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Signal, Image, Parole, Télécoms**

Arrêté ministériel : 7 août 2006

Présentée par

**Thibaut FUX**

Thèse dirigée par **Gang FENG** et  
Co-encadrée par **Véronique ZIMPFER**

préparée au sein de **L'institut franco-allemand de recherches  
de Saint-Louis (ISL), 5 rue du Général Cassagnou, 68300  
Saint-Louis France**  
et du **Laboratoire GIPSA-lab, département Parole et Cognition**  
dans **l'École Doctorale EEATS**

# Vers un système indiquant la distance d'un locuteur par transformation de sa voix

Thèse soutenue publiquement le **jeudi 24 mai 2012**  
devant le jury composé de :

**M. Pierre-Yves COULON**

Professeur, Université de Grenoble - GIPSA-lab, *Président*

**M. Olivier BOËFFARD**

Professeur, Université de Rennes 1 - IRISA, *Rapporteur*

**M. Jean-Sylvain LIÉNARD**

Directeur de recherches CNRS, Université Paris Sud - LIMSI, *Rapporteur*

**M. Olivier ROSEC**

Ingénieur de recherche, Orange Labs, *Examineur*

**M. Gang FENG**

Professeur, Université de Grenoble - GIPSA-lab, *Directeur de thèse*

**Mme Véronique ZIMPFER**

Chargée de recherches, ISL, *Co-directrice de thèse*





*“[...] Voice transformation requires more than just modeling the speech signal; it requires understanding the speech process in terms of production, perception, and natural language processing.”*  
— Yannis Stylianou



# REMERCIEMENTS

Cette thèse a principalement été effectuée à l'Institut *franco-allemand de recherches de Saint Louis* (ISL) au sein du groupe *Acoustique et Protection du Combattant*. Je tiens à remercier les directeurs de l'ISL, M. Christian de VILLEMAGNE et M. Wolfgang FÖRSTER, le chef de division M. Emil SPAHN ainsi que le chef du groupe (APC), M. Pierre NAZ, de m'avoir permis de réaliser cette thèse dans cet institut. Ce travail de thèse a également été effectué en collaboration avec le laboratoire *Grenoble-Image-Parole-Signal-Automatique (GIPSA-lab)*, département Parole et Cognition dont je remercie son directeur M. Jean-Marc CHASSERY et le chef de département M. Gérard BAILLY pour m'avoir permis de profiter de l'expérience de ce laboratoire de renom.

Je remercie vivement les membres de mon jury de thèse : M. Olivier BOËFFARD et M. Jean-Sylvain LIÉNARD pour avoir accepté la lourde tâche de rapporter ce travail. Merci également à M. Olivier ROSEC d'avoir accepté de participer à ce jury de thèse malgré les modifications qui ont été apportées à la composition de celui-ci. Je remercie également M. Olivier ROSEC pour l'entretien téléphonique que nous avons eu et qui m'a permis de voir différemment mon travail et ainsi de réfléchir plus profondément sur celui-ci. Merci également à M. Pierre-Yves COULON d'avoir accepté de présider ce jury.

Arrivé en stage de master en n'ayant aucune expérience ni connaissance sur la parole et sur le signal de la parole je remercie chaleureusement l'ensemble des personnes avec qui j'ai pu discuter et qui ont pris le temps de m'enseigner les différents attraits de cette discipline à part entière. En particulier, je tiens à exprimer toute ma gratitude à mon directeur de thèse, M. Gang FENG professeur à l'Institut national polytechnique de Grenoble pour son investissement, son intérêt ainsi que pour ses conseils précieux. Merci également, de m'avoir appris à dompter le monde du traitement de la parole. Mais également pour les différents échanges aussi bien pour les tâches de rédactions (articles et manuscrit) que pour les divers questionnements. Même si vous avez une vision précise du rôle de directeur de thèse, qui peut au premier abord être déroutante, j'ai grandement apprécié votre philosophie de l'apprentissage par soi-même. La liberté que vous m'avez laissée quant aux différents choix ainsi que pour l'orientation de la thèse a été pour moi une grande preuve de confiance de votre part et m'a permis de murir professionnellement, de manière radicale.

Ayant passé le plus clair de ma thèse à l'ISL, je remercie ma codirectrice de thèse Véronique ZIMPFER chargé de recherche à l'Institut franco-allemande de recherche de Saint Louis, qui a pris le temps de me conseiller et de valider mes différents choix. Je la remercie également de m'avoir fait profiter de ses connaissances aussi bien dans la parole ainsi que dans le domaine de l'acoustique en général ainsi que pour la confiance qu'elle m'a accordé. Les grandes discussions que nous avons eues, m'ont permis d'appréhender les problèmes de recherche de façons différentes mais également d'enrichir ma culture scientifique.

Un grand merci également à toute l'équipe du groupe APC de l'ISL, pour leur accueil chaleureux. Merci pour votre soutien ainsi que pour ces pauses café mythiques. Je remercie tout particulièrement M. Karl BUCK, M. Sébastien DE MEZZO, M. Pascal HAMERY pour leur aide, leurs encouragements et leurs conseils précieux.

Mes trop rares séjours au sein du GIPSA-Lab m'ont permis de discuter avec des chercheurs de renom dont les discussions (trop rares également) ont été d'une aide précieuse aussi bien pour la validation de l'approche que pour le côté théorique. Je pense particulièrement à Nathalie HENRICH et à Mäeva GARNIER que je remercie vivement. Merci également à l'ensemble des gens du laboratoire de Grenoble pour leur accueil toujours très amical et agréable.

Cette thèse s'est orientée au fil des mois vers des analyses prosodiques. Etant de formation traiteur de signaux et automatique, cette partie du travail n'a pas été, dans un premier temps, des plus attrayantes. Toutefois, le soutien, les encouragements ainsi que les explications qu'a pu me fournir Mme. Véronique AUBERGÉ ont été d'une très grande aide. Une grande partie de ce travail n'aurait pas pu être possible sans son intervention. Merci alors à toi Véro pour ces discussions souvent courtes, mais intenses (aussi bien sur le fond que par le niveau sonore)!!! Merci également à Nicolas AUDIBERT, pour ces nombreux conseils et notamment vers la fin de ce travail.

Merci aux thésards que j'ai pu côtoyer durant cette formidable expérience. Ainsi je remercie Anne VANPÉ, Krystyna GRABSKI, Rosario SIGNORELLO, pour le GIPSA-lab, et Loïc EHRHARDT, Matthieu (je sais plus ton nom de famille, oups), Jérôme DHOLLANDE de l'ISL, pour les bons moments passés ainsi que pour ces sessions de plaintes mutuelles qui ne servent à rien mais qui comme chacun le sait fait du bien ! Merci également à la stagiaire Stéphanie LOBREAU que j'ai encadré, dont son travail a permis d'apporter quelques réflexions dans ce travail ; et sans doute sa première référence. Si ça s'est pas la classe !

Je tiens également à remercier chaleureusement mon frère Elie pour son soutien, ses encouragements et le partage de son expérience de doctorant. Même si c'était plus stressant qu'encourageant la plupart du temps ça m'a été utile.

Je félicite également la patience du groupe APC et les autres personnes de l'ISL durant la période des tests perceptifs. D'une manière générale je remercie tout les personnes qui ont bien voulu participer à ces tests. Rassurez-vous, c'est fini ! Je remercie en particulier mon cousin Brice de m'avoir permis de profiter de son cabinet de kinésithérapie et d'ostéopathie et d'avoir convaincu ses patients de participer à des test perceptifs. Ça c'était une bonne idée !

Merci à toute ma famille pour la compassion et leur compréhension quand j'étais (un peu) stressé et pas forcément agréable à vivre. Là encore rassurer vous ça va aller mieux ! Je remercie tous particulièrement mes parents sans qui je n'aurais pu faire ni finir cette thèse. Merci papa et merci maman de m'avoir poussé durant les études afin que je puisse en arriver là mais également de m'avoir fait confiance au cours de celles-ci.

Merci aussi à Huguette, Phillipe et Camille pour toutes ces petites attentions qui vont droit au cœur. Notamment la machine à café, le vin et la bière, les tableaux. Mais surtout la bière !

Je salue également l'intérêt et le soutien de tous mes amis pour mon travail, même si pour la plupart ils ne se rappellent jamais ce que je fais. Merci à vous, les potes, en m'excusant pour ceux que j'oublie : Yann, Axel, David et Marie-Eve (et Aaron), Eric et Caren, Matthieu et Nelly, Xavier, Nico et Susie, Moog, Franck, Cricri.

Merci également à celle qui partage ma vie, Marion. Merci ma puce pour ta patience à toute épreuve ta compréhension et tous tes encouragements. Sans ton aide et ton soutien je ne sais pas si j'aurais réussi à finir.

Enfin, merci à tous ces gens qui ont participé indirectement à ce travail. Merci à Marco, pour ses bonnes pizzas très revigorantes et à Momo et Carine pour ses grandes discussions philosophiques très relaxantes.

Une dernière pensée pour tous ceux qui ne sont plus là et avec qui je n'ai pas pu partager ces moments de peines et de bonheur. Je pense à mes grands parents ainsi qu'à mes deux amis partis trop tôt.

# SOMMAIRE

<b>Remerciements .....</b>	<b>5</b>
<b>Sommaire .....</b>	<b>7</b>
<b>Abréviations.....</b>	<b>9</b>
<b>Conventions .....</b>	<b>10</b>
<b>CHAPITRE 1 Introduction générale .....</b>	<b>11</b>
1.1 <i>Contexte .....</i>	11
1.2 <i>La problématique .....</i>	15
1.3 <i>Orientation de la thèse et méthodologie.....</i>	16
1.4 <i>Restriction de l'étude .....</i>	20
1.5 <i>Organisation du document .....</i>	21
<b>PARTIE I : FONDAMENTAUX ET ACQUIS.....</b>	<b>23</b>
<b>CHAPITRE 2 Perception auditive de distance (PAD).....</b>	<b>25</b>
2.1 <i>PAD liés par les phénomènes acoustiques .....</i>	26
2.2 <i>Autres phénomènes intervenant dans la PAD .....</i>	35
2.3 <i>PAD par la voix du locuteur .....</i>	36
2.4 <i>Conclusion du chapitre 2 .....</i>	40
<b>CHAPITRE 3 La parole : production, représentation et variabilité.....</b>	<b>43</b>
3.1 <i>Anatomie et physiologie de l'appareil phonatoire .....</i>	43
3.2 <i>Théorie acoustique de la production de la parole.....</i>	50
3.3 <i>Modélisation acoustique de la parole .....</i>	54
3.4 <i>La parole : information et variabilité.....</i>	63
3.5 <i>Paramètres paralinguistiques .....</i>	67
<b>CHAPITRE 4 Effort vocal et effort vocal.....</b>	<b>71</b>
4.1 <i>Causes et circonstances .....</i>	71
4.2 <i>Effort vocal : définition nous concernant.....</i>	73
4.3 <i>La mesure de l'effort vocal.....</i>	76
<b>CHAPITRE 5 Caractéristiques des voix en communication à distance : état de l'art.....</b>	<b>79</b>
5.1 <i>État de l'art des connaissances actuelles sur la voix chuchotée .....</i>	80
5.2 <i>État de l'art des connaissances actuelles sur la voix criée .....</i>	85
5.3 <i>Identification du locuteur et intelligibilité.....</i>	96
5.4 <i>Conclusion du chapitre 5 .....</i>	98
<b>PARTIE II : CARACTÉRISATION DES VOIX CRIÉES .....</b>	<b>101</b>
<b>CHAPITRE 6 Élaboration des bases de données.....</b>	<b>103</b>
6.1 <i>Protocole d'enregistrement en champ libre : DB1 .....</i>	104
6.2 <i>Protocole d'enregistrement en champ libre version améliorée : DB2.....</i>	107
6.3 <i>Protocole de simulation de communication orale à distance : DB3.....</i>	112
6.4 <i>Corpus dédié à l'analyse micro-prosodique : DB4.....</i>	115
6.5 <i>Conclusion du chapitre 6 .....</i>	117
<b>CHAPITRE 7 Perception de l'effort vocal : Prosodie ou paramètres spectraux ? .....</b>	<b>119</b>
7.1 <i>État de l'art sur les paramètres pertinents de l'effort vocal .....</i>	120
7.2 <i>Le rôle de la prosodie dans la perception auditive de l'effort vocal.....</i>	127
7.3 <i>Conclusions du chapitre 7.....</i>	135

<b>CHAPITRE 8 Dynamique des paramètres et effort vocal .....</b>	<b>137</b>
8.1 <i>Analyses de l'intensité et de la F0 globales .....</i>	138
8.2 <i>Dynamique des paramètres vs effort vocal .....</i>	141
8.3 <i>Dynamique des formants.....</i>	152
8.4 <i>Conclusion du chapitre 8 .....</i>	156
<b>CHAPITRE 9 La prosodie et la micro- prosodie de l'effort vocal.....</b>	<b>159</b>
9.1 <i>Analyses de la durée.....</i>	160
9.2 <i>Analyses de l'intensité des logatomes .....</i>	169
9.3 <i>Analyses de la fréquence fondamentale des logatomes.....</i>	179
9.4 <i>Micro-mélodie des énoncés CV, CVC, VCV et CVCV.....</i>	190
9.5 <i>Interprétations hypothétiques.....</i>	210
9.6 <i>Conclusions du chapitre 9.....</i>	213
<b>PARTIE III TRANSFORMATION .....</b>	<b>215</b>
<b>CHAPITRE 10 Transformation de la voix.....</b>	<b>217</b>
10.1 <i>Choix de la méthode d'analyse-synthèse.....</i>	218
10.2 <i>Description des procédures d'analyse .....</i>	222
10.3 <i>Transformation de voix modale vers une voix chuchotée.....</i>	227
10.4 <i>Transformation de voix modale en voix criée .....</i>	236
10.5 <i>Conclusion du chapitre 10 .....</i>	251
<b>CHAPITRE 11 Évaluation des transformations.....</b>	<b>253</b>
11.1 <i>Tests d'intelligibilité.....</i>	254
11.2 <i>Perception de distance .....</i>	267
11.3 <i>Conclusions du chapitre 11.....</i>	275
<b>CHAPITRE 12 Conclusions générales et perspective .....</b>	<b>277</b>
12.1 <i>Perspectives de l'étude.....</i>	280
12.2 <i>D'autres champs d'application possibles .....</i>	281
<b>Table des matières.....</b>	<b>283</b>
<b>Références.....</b>	<b>287</b>
<b>ANNEXES .....</b>	<b>309</b>
<b>Annexe A Listes des logatomes CV, CVC, VCV ; CVCV.....</b>	<b>311</b>
<b>Annexe B Mesure du quotient ouvert pour le corpus DB3.....</b>	<b>313</b>
<b>Annexe C Matrice de confusion des tests d'intelligibilité pour les voyelles .....</b>	<b>315</b>
<b>Annexe D Matrice de confusion des tests d'intelligibilité en fonction des locuteurs .....</b>	<b>319</b>
<b>Annexe E Valeurs caractéristiques des contours de F0 des logatomes.....</b>	<b>341</b>
<b>Annexe F Résultats des tests de perception de la distance pour chaque sujet .....</b>	<b>355</b>
<b>Résumé.....</b>	<b>358</b>

# ABRÉVIATIONS

<b>Abréviations</b>	<b>Description</b>
<i>dB SPL</i>	Décibel Sound Pressure Level
<i>DLA</i>	Distance Locuteur-Auditeur
<i>PAD</i>	Perception auditive de la distance
<i>Signal EGG</i>	Signal issu de l'Electroglottographe
<i>Signal DEGG</i>	Dérivé première du signal issu de l'Electroglottographe
<i>LP</i>	Linear Prediction
<i>OLA</i>	Overlap Add
<i>CV</i>	Logatomes de type consonne-voyelle (C1-V1)
<i>VCV</i>	Logatomes de type voyelle-consonne-voyelle (V1-C1-V2)
<i>CVC</i>	Logatomes de type consonne-voyelle-consonne (C1-V1-C2)
<i>CVCV</i>	Logatomes de type consonne-voyelle-consonne-voyelle (C1-V1-C2-V2)

<b>Notations</b>	<b>Description</b>
<i>F0</i>	Fréquence fondamentale de la parole
<i>F<sub>x</sub></i>	Fréquence centrale du formant n° x
<i>B<sub>wx</sub></i>	Largeur de bande à -3 dB du formant n° x
$\Delta F0$	Variation de fréquence fondamentale
$\delta F0$	Dynamique de la fréquence fondamentale
<i>I</i>	Intensité en dB SPL
<i>F<sub>s</sub></i>	Fréquence d'échantillonnage d'un signal numérique
$\Delta I$	Variation d'intensité
$\delta I$	Dynamique de l'intensité

# CONVENTIONS

## 1- Tableau des symboles phonétiques des voyelles utilisés

VOYELLES							
ORALES				NASALES			
[a]	<b>pl</b> at	[ɔ]	<b>bo</b> l	[e]	<b>blé</b>	[ã]	<b>blanc</b>
[ɑ]	<b>mâ</b> t	[o]	<b>pô</b> t	[œ]	<b>he</b> ure	[õ]	<b>bo</b> n
[i]	<b>pi</b> le	[ə]	<b>le</b>	[u]	<b>rou</b> e	[ɛ̃]	<b>li</b> n
[y]	<b>ru</b> e	[ɛ]	<b>lai</b> t	[ø]	<b>peu</b>	[œ̃]	<b>bru</b> n

## 2- Tableau des symboles phonétiques des consonnes utilisés

CONSONNES													
LIQUIDES	NASALES	FRICATIVES		OCCLUSIVES		SEMI							
		voisées	non voisées	voisées	non voisées	voisées							
[l]	<b>lai</b> t	[m]	<b>mai</b> s	[v]	<b>vai</b> s	[f]	<b>fai</b> t	[b]	<b>bai</b> e	[p]	<b>pai</b> e	[j]	<b>hie</b> r
[R]	<b>rai</b> e	[n]	<b>nei</b> s	[z]	<b>zé</b> ros	[s]	<b>sai</b> t	[d]	<b>dai</b> s	[t]	<b>ta</b> i	[ʁ]	<b>hui</b> t
		[ɲ]	<b>gai</b> ner	[ʒ]	<b>jeu</b> s	[ʃ]	<b>che</b> z	[g]	<b>gai</b>	[k]	<b>g</b> uai	[w]	<b>oui</b>

## 3- Symboles de significativité

Les tests de significativité ont été effectués via le test de *Student*. Les notations utilisées sont les suivantes :

\*\*\* correspond à  $p < 0,001$ , \*\* correspond à  $p < 0,010$ , \* correspond à  $P < 0,05$  et n.s. à  $p > 0,05$

## 4- Référence du demi-ton

Dans certaines situations les valeurs des fréquences fondamentales seront exprimées en demi-tons. Pour ce faire nous avons choisi comme référence la valeur de 50 Hz.

# Introduction générale

---

*“I felt advancing communication would advance our quality of life.”*  
— James L. Flanagan

La vision et l’audition de l’Homme se sont développées pour percevoir les informations dans un espace en trois dimensions (3D). Cependant, les systèmes actuels de restitution d’images et/ou de sons fournissent des informations qui dans la majorité des cas ne sont qu’en deux dimensions (images, sons stéréophoniques). Certes, dans le cas du son, une restitution en stéréo permet d’apporter une représentation sommaire de l’orientation d’une source sonore, il s’agit cependant de dimension réduite car il n’est pas toujours facile de percevoir la *distance des sources*. Ainsi, la restitution en 3D autant pour les images que pour le son, fait l’objet d’études approfondies depuis plusieurs années et de nouvelles avancées technologiques font leur apparition de jour en jour. Notre thèse se situe dans le cadre de la reconstruction ou restitution de son en 3D.

## 1.1 Contexte

---

Plusieurs méthodes permettent de reconstruire, à partir d’un ou plusieurs sons monophoniques, un environnement sonore en 3D. Ceci est possible soit en utilisant plusieurs sources sonores comme pour l’Ambisonie ou encore les systèmes 5.1 (qui est une extension du « stéréo » en quelque sorte), soit en utilisant un simple casque stéréophonique comme pour les techniques basées sur les fonctions de transfert relatives à la tête (HRTFs) (Cheng et Wakefield, 2001). Les HRTFs constituent certainement l’approche la plus simple et la plus efficace permettant de se rapprocher des phénomènes physiques pour reconstruire un environnement sonore en 3D. En effet, les méthodes basées sur les HRTFs utilisent des couples de fonctions de transfert (oreille droite et oreille gauche) en fonction de la position d’une source sonore dans l’espace en trois dimensions. Elles permettent de reproduire les éléments essentiels à la perception de l’orientation d’une source sonore (azimut et élévation); à savoir

la différence interaurale du temps d'arrivée<sup>1</sup> de l'onde sonore, la différence interaurale du niveau sonore<sup>2</sup> de cette même onde ainsi que les masquages fréquentielles et interférences créés par l'oreille externe, la tête et le torse (Middlebrooks et Green, 1991). Ainsi pour reconstruire un signal en 3D et simuler une source sonore située à une position désirée, le signal monophonique est filtré par le couple de filtres HRTFs associé à la position à simuler. Il en résulte deux signaux : l'un destiné à l'oreille gauche et l'autre à l'oreille droite qui à eux deux contiennent l'ensemble des éléments permettant de percevoir le son en 3D.

Cette méthode de re-spatialisation du son offre l'avantage d'être relativement simple pour une performance acceptable. De plus, le matériel nécessaire à sa mise en œuvre est un matériel classique (un simple casque stéréophonique et un DSP (Digital Signal Processor) pour les traitements) et donc peu onéreux. Toutes ces raisons font que l'armée s'intéresse tout particulièrement à cette méthode. Des essais réalisés montrent que le temps de réaction d'un pilote d'avion face à une menace est bien plus faible quand un signal d'alarme en 3D lui indique dans quelle direction se situe la menace (Bronkhorst et al., 1996). De plus il a également été montré que ces techniques permettent d'accroître l'intelligibilité dans des situations de radio-communication multi-locuteurs; et ce en représentant en 3D la voix de chacun d'eux (Arrabito, 2000). La reconstruction en 3D de la voix permet également à un homme d'infanterie de localiser rapidement la position de son interlocuteur dans le cas où la voix de ce dernier est représentée en 3D. Ces applications ont pour but de simuler via un système de restitution sonore les phénomènes qui auraient été perçus dans le cas d'une écoute libre (i.e. sans système de restitution).

Bien que l'on parle de restitution audio 3D pour la majorité des systèmes de restitution spatiale, le terme nous semble un peu abusif. En effet, la plupart de ces systèmes et notamment les HRTFs ne permettent de reconstruire que deux dimensions : l'azimut et l'élévation. Une information majeure reste manquante : **la distance de la source sonore**. Nous pourrions imaginer étendre le concept des HRTFs à la distance en enregistrant des HRTFs pour plusieurs distances. Outre l'ampleur titanesque de la tâche, la base de données ainsi créée ne sera valide et utilisable que pour un environnement précis ; celui dans lequel les relevés auraient été effectués. En effet, contrairement aux phénomènes liés à la localisation d'une source sonore sur le plan azimutal et sur le plan vertical (l'élévation), qui peuvent s'expliquer par des phénomènes purement physiques (Middlebrooks et Green, 1991), la *perception auditive de la distance* est quant à elle essentiellement basée sur la déduction, et sur l'apprentissage passé de l'auditeur (Zahorik et al., 2005). Ainsi, la précision de l'estimation de distance dépend fortement de l'environnement d'écoute ainsi que du type de source sonore à localiser. L'estimation de distance sera bien meilleure pour une source sonore et un environnement familier. De

---

<sup>1</sup> Traduction de l'anglais : *Interaural Time Difference* (ITD)

<sup>2</sup> Traduction de l'anglais : *Interaural Level Difference* (ILD)

ce fait, une représentation précise de la distance d'une source sonore impliquerait une parfaite connaissance de l'environnement de propagation de l'onde acoustique, ainsi qu'une connaissance précise des caractéristiques de la source sonore. On comprend bien alors que cette tâche est titanesque et ne peut se limiter qu'à certaines situations particulières.

Cependant, dans le cadre d'une communication orale, ces éléments peuvent être mis au second plan car une information supplémentaire, mais essentielle, est prise en compte par l'auditeur pour estimer la distance de son interlocuteur. Il s'agit de la voix. En effet, **la voix porte une information de distance** étant donné que le niveau de production vocale est volontairement ajusté par le locuteur en fonction de la distance qui le sépare de son interlocuteur. Cet ajustement vocal fait par le locuteur, a pour but essentiel de conserver une bonne intelligibilité malgré la distance qui sépare les interlocuteurs. Ainsi, dans le cas où deux personnes communiquent à une distance (qui convient à une discussion verbale), un locuteur, ajustera sa voix dans le but d'assurer un niveau sonore acceptable au niveau des oreilles de son interlocuteur afin de se faire comprendre par celui-ci. Par conséquent pour de courtes distances de communication (de l'ordre de quelques centimètres) jusqu'à de grandes distances (de l'ordre de la centaine de mètres), la voix du locuteur varie progressivement d'un niveau de production vocale bas (le chuchotement) jusqu'à des niveaux de production vocale plus intenses (le cri / le hurlement).

Grâce à ces variations de production vocale, un auditeur est capable de juger de la distance de la personne qui parle en se basant uniquement sur la voix (Brungart et Scott, 2001; Cochran et al., 1968; Eriksson et Traunmüller, 2002; Fux et Zimpfer, 2009; Gardner, 1969; Philbeck et Mershon, 2002). En effet, grâce à sa capacité déductive, un auditeur considère qu'une personne qui utilise une voix criée se situe en général loin de lui (hors cas des cris d'énervement, d'autorité, ...), et ce de façon quasi-indépendante des effets liés à la propagation de l'onde sonore. (i.e. des variations d'intensité, du taux de réverbération ou encore des modifications spectrales). Même si ces derniers, et notamment la réverbération, peuvent éventuellement contribuer à améliorer la précision des estimations de distance (Shinn-Cunningham, 2000).

Dans le cas de la perception auditive de la distance du locuteur, le type de voix utilisé par le locuteur a une influence considérable sur l'estimation de distance. En effet, plus la voix est forcée plus la distance perçue est lointaine. Les variations de production vocale, en fonction de la distance, sont directement liées à *l'effort vocal*. L'effort vocal est d'ailleurs une notion très importante que nous préciserons par la suite.

Il est ainsi envisageable, dans le cas de la radiocommunication, d'apporter une information sur la distance d'un locuteur par transformation de la source sonore elle-même : **la voix**. Cette approche, initialement proposée par Brungart et Scott en 2000 (Brungart et Scott, 2000) constitue le point de départ de notre étude. Comme nous l'avons déjà mentionné, l'armée s'intéresse particulièrement à ce

type d'approche. L'AFLR (Air Force Research Laboratory) a d'ailleurs financé un projet sur le thème « Speech Synthesis for Distance Cueing in Audio Displays » en 2005<sup>1</sup>. Cependant, à notre connaissance, il n'existe qu'une seule étude résultant de ce projet (Cheyne et al., 2009). Nos travaux constituent la continuité logique de ces études mais avec l'originalité de tenir compte de certains éléments ignorés jusqu'à présent tel que les variations prosodiques liées à l'effort vocal.

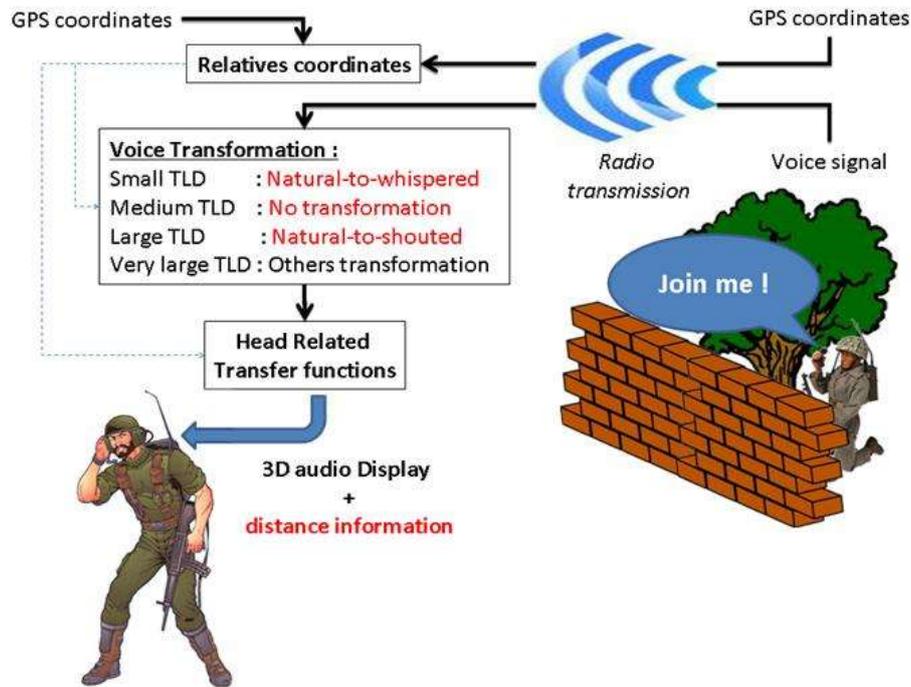


Figure 1-1: Représentation du système complet de communication radiophonique incluant la représentation 3D de la voix ainsi qu'une indication de la distance du locuteur

On peut illustrer l'esprit essentiel de cette étude à l'aide de la Figure 1-1. Comme les études précédentes, nous proposons ainsi de modifier l'effort vocal de la voix du locuteur en fonction de la distance de son interlocuteur. En d'autres termes, nous souhaitons transformer une voix modale en une voix plus ou moins chuchotée pour évoquer une distance plus ou moins courte ou en une voix plus ou moins criée pour évoquer une distance de communication plus ou moins grande. Ce principe additionné à la représentation 3D des sons (voir à une modélisation de la propagation des ondes sonores) constitue une approche de réalité virtuelle visant à recréer directement au niveau des oreilles d'un auditeur l'ensemble des phénomènes perçus.

*Ainsi, l'objectif final de cette thèse consiste à compléter les systèmes de communication radiophonique 3D en apportant une notion de distance de l'interlocuteur par transformation de sa voix.*

<sup>1</sup> Source : <http://www.sbir.gov/sbirsearch/detail/303900>

---

## 1.2 La problématique

---

Dans le but de transformer une voix modale en voix chuchotée ou criée, il est nécessaire de connaître les paramètres pertinents permettant de décrire l'une et l'autre de ces voix. En d'autres termes, il s'agit de répondre à la question :

*Quelles sont les caractéristiques de la voix permettant de différencier une voix chuchotée ou une voix criée d'une voix modale ?*

Soulignons tout de suite que les mécanismes permettant à un auditeur de percevoir l'effort vocal, et par ce biais d'estimer la distance d'un interlocuteur, sont encore mal connus. C'est pourquoi, l'identification des caractéristiques pertinentes du signal de la parole pour la perception de l'effort vocal (et donc la distance) constitue encore aujourd'hui un vrai défi. Mais notre travail se situe précisément dans ce cadre-là.

Dans la littérature, la transformation d'une voix modale en une **voix chuchotée** a déjà été réalisée par plusieurs études avec plus ou moins de succès (Agiomyrgiannakis et Rosec, 2009; Farner et al., 2009). L'étape essentielle consiste à dévoiser<sup>1</sup> une voix modale. En général, ce dévoisement seul est considéré comme étant suffisant. En effet, la caractéristique la plus pertinente pour la reconnaissance d'une voix chuchotée est l'absence de vibration des cordes vocales. Bien que la modification d'autres paramètres vocaux soit nécessaire afin d'obtenir une meilleure qualité de transformation, dévoiser le signal de la voix modale est souvent suffisant pour évoquer un niveau de production vocale faible. Les transformations ainsi réalisées restent néanmoins très acceptables. Notons que les études ayant pour but une transformation de voix chuchotée vers une voix modale évoquent des modifications supplémentaires directement liées à l'absence de vibration des cordes vocales (variation d'énergie intrinsèque, place des formants, modification spectral du au bruit de constriction) (Morris, 2003) qui dans la majorité des cas ne sont pas prises en compte pour la transformation de voix modale en voix chuchotée.

Cependant, transformer une voix modale en une **voix criée** est nettement plus complexe. A l'inverse de la voix chuchotée où l'absence de vibration des cordes vocales constitue le facteur pertinent prédominant dans la transformation et l'identification de ces types de voix, les paramètres pertinents de la voix criée ne sont pas aussi clairement identifiés. On connaît certains éléments prédominants comme l'augmentation de l'intensité de la voix lors d'un cri, mais ceci ne semble pas tout à fait pertinent. Nous reprendrons ici un exemple très parlant évoqué par Ingo Titze :

---

<sup>1</sup> Le verbe « dévoiser » évoque l'action d'éliminer dans le signal de la voix, l'influence de la vibration des cordes vocales. En d'autres termes, « dévoiser » un signal consiste à supprimer sa composante périodique.

*“The fact that we can turn the radio up and down and still tell whether a vocalist sings loud or soft is evidence that loudness involves the spectrum of the sound, not just the amount of sound”. (Titze, 1994, p.244)*

En d’autres termes, même si on augmente le volume sonore d’un poste radio, nous n’avons pas l’impression que l’animateur ou le chanteur que l’on entend est en train de crier ; le son de sa voix sera simplement plus fort. Ainsi l’intensité ne semble pas être, du moins à elle seule, un facteur pertinent pour la reconnaissance d’une voix criée. Un autre élément connu est l’augmentation de la vitesse de vibration des cordes vocales (i.e la fréquence fondamentale de vibration des cordes vocales ou fréquence fondamentale de la voix). Mais une augmentation de la fréquence fondamentale de la voix (notée F0) seule, n’aura pas non plus pour effet de transformer une voix modale en voix criée mais simplement l’effet de produire une voix plus aigüe et souvent peu naturelle. Ainsi, la voix criée relève de mécanismes plus complexes qu’une simple augmentation de l’intensité et/ou de la F0. Malheureusement, nous ne disposons pas suffisamment d’éléments dans la littérature concernant la modélisation des mécanismes de production et de perception de la voix criée permettant l’élaboration de règles et d’algorithmes de transformation directement.

### 1.3 Orientation de la thèse et méthodologie

---

L’objectif premier de cette étude est donc de déterminer quels sont les paramètres pertinents nécessaires à la transformation d’une voix modale en voix criée (ou chuchotée) et la façon dont ces paramètres varient. Pour ce faire il est nécessaire d’étudier les différences entre une voix modale et une voix criée (ou chuchotée). Étant donnée les connaissances sur la transformation en voix chuchotée qui existent dans la littérature, notre travail se concentre principalement sur la caractérisation de la voix à fort effort vocal. Nous effectuons des analyses approfondies dans le but d’élaborer des règles de transformation permettant le passage d’une voix modale en une voix criée pour refléter la distance d’un locuteur. Concernant la voix chuchotée, nous proposons, sur la base de la littérature et de nos propres réflexions et analyses, une méthode de transformation de très bonne qualité.

Pour mener à bien ce projet plusieurs étapes sont nécessaires qui peuvent se résumer comme suit :

1. Enregistrement de corpus
2. Analyse du corpus et élaboration de règles de transformation
3. Modification de la voix parlée

### 1.3.1 Enregistrements de corpus

L'objectif essentiel de cette étape est d'obtenir des signaux de parole produits par un locuteur situé à des distances différentes allant typiquement de 1 m à 100 m. Concernant cette étape, il s'agit de mettre en place un protocole d'enregistrement permettant de refléter le plus fidèlement possible la communication verbale à distance. Cette étape est cruciale pour permettre de réaliser des analyses et d'en extraire des règles de transformation.

### 1.3.2 Analyse du corpus et élaboration de règles de transformation

Rappelons d'abord que pour réaliser une transformation de la voix, plusieurs approches sont envisageables.

La première approche est dite **analytique**. Celle-ci consiste à réaliser des analyses systématiques sur les signaux de la parole du corpus. Ces analyses ont pour objectif d'identifier des corrélats entre les variations de paramètres de la voix et la distance de communication. Par la suite, à partir de ces corrélats, des règles de modifications peuvent alors être établies afin de permettre le passage d'une voix parlée en voix criée ou chuchotée. Ces règles pouvant être alors basées sur des modèles existant ou encore être empiriques.

Une deuxième approche qui peut être vue comme complémentaire à la précédente est dite **statistique**. Cette approche déjà utilisée dans la conversion vocale consiste à entraîner un modèle statistique de type modèle de Markov caché (HMM<sup>1</sup>), modèle de mélange de gaussienne (GMM<sup>2</sup>) ou réseau de neurones sur un corpus de voix donné. Sur la base de certains paramètres vocaux définis à l'avance, ces modèles déterminent des lois de passage statistiques entre les paramètres de voix sources et les paramètres de voix cibles (par exemple entre une voix parlée et une voix criée). Ces lois permettent par la suite de modifier les caractéristiques de la voix pour réaliser la conversion souhaitée. Ces méthodes statistiques possèdent l'avantage de réaliser des conversions de voix de grande qualité avec de moindres coûts de calcul. La plus grosse charge de calcul, consistant à analyser les corpus et établir les lois de passage, est réalisée au préalable. Toutefois, l'inconvénient majeur de ces méthodes est l'aspect statistique de la conversion. En effet, les lois de passages établies ne sont, en général nullement interprétables aux niveaux des mécanismes de production vocale car non basées sur des modèles de la production. Ainsi l'apport de ces méthodes d'un point de vue fondamental reste très faible. Par ailleurs, le manque de connaissance sur la voix criée est également un problème de taille pour cette méthode. En effet, comme les paramètres pertinents de la voix criée ne sont pas encore clairement établis, il est délicat de savoir sur quels paramètres entraîner ces modèles. Pour notre étude,

<sup>1</sup> De l'anglais « Hidden Markov Model »

<sup>2</sup> De l'anglais « Gaussian Mixture Model »

l'approche statistique est une option envisageable. Mais la variabilité de la voix et tout particulièrement la variabilité des modifications observées entre la voix parlée et la voix créée par exemple, ne nous semble pas permettre d'appliquer une telle méthode de manière aisée.

La troisième et dernière approche, celle de plus bas niveau, se base sur la physiologie l'appareil phonatoire. Cette approche consiste à mesurer des effets de la voix créée ou chuchotée directement sur les organes de l'appareil phonatoire ainsi que les différentes grandeurs impliquées (pression, tension des muscles, ...) mais également des signaux de la voix. Puis, à partir des théories articulo-acoustique il est alors possible de prédire les paramètres qui seront affectés par les différents ajustements de l'appareil phonatoire mais également de quantifier leurs variations. Enfin, il faut vérifier après cette étape si les résultats obtenus correspondent aux observations faites sur la voix créée directement. Les règles de modification seront alors issues d'un modèle articulo-acoustique qui prendrait en compte les éléments physiologiques identifiés. Contrairement aux deux approche précédentes, celle-ci constitue une étude qui relève plus du domaine médicale que du traitement du signal. L'avantage de cette approche est un niveau très fondamental. Mais cette approche nécessite une forte compréhension de physiologie humaine, des modèles articulatoires et de vibration des cordes vocales, qui malheureusement ne sont pour le moment pas pleinement opérationnels. De plus un frein majeur à ce type d'approche est la difficulté d'observation de certaines données physiologiques telles que la tension des muscles ou encore la masse ou la rigidité des cordes vocales qui sont primordiales pour effectuer les simulations aéromécaniques.

Il s'agit là uniquement de la première étape qui consiste à identifier les paramètres dans le signal de parole ainsi que de mettre en place des règles de transformation.

### **1.3.3 Modification de la voix parlée**

Les règles étant définis, il s'agit à présent de les appliquer sur le signal de la voix parlée. Si l'approche est une approche physiologique, un modèle articulatoire est le plus approprié. Ceux-ci permettent de modifier directement des paramètres physiologiques tel que la pression issu des poumons, la tension des cordes vocales, l'ouverture de la bouche, etc... Par la suite, ces modèles aéromécaniques permettent de reproduire le signal acoustique de la parole. D'une part, l'étude physiologique est complexe mais la mise en place de modèle aéromécanique l'est tout autant.

Si les règles de transformation sont issues d'analyse acoustique, il existe alors plusieurs techniques et modèles permettant de modifier les paramètres de la voix.

Un groupe de techniques est basée sur la théorie de la production de la parole (Fant, 1960). Ces méthodes analytiques consistent à séparer du signal de la parole, la contribution due au conduit vocal de la radiation effectué par les lèvres, et celle due à la vibration des cordes vocales. L'avantage de ces

modélisations est une décomposition paramétrique des éléments constitutants. Ces méthodes permettent une très bonne modélisation de la production vocale tout en restant dans un espace paramétrique permettant de quantifier et de modifier le signal de la parole. Notons toutefois qu'il est très délicat d'extraire de façon précise les différentes composantes de la parole ; à savoir la réponse en fréquence du conduit vocal ainsi que le signal généré par la vibration des cordes vocales. Cette décomposition dite source-filtre n'est pas des plus simples. De plus cette théorie de la production vocale fait une hypothèse forte à savoir que le signal de la source glottique et le filtre du conduit vocal sont indépendants. Or, on sait qu'il existe des interactions entre la source et le filtre.

D'autres techniques dites de modification ne se base sur aucun a priori concernant le signal et permettent de réaliser des transformations diverses qui ne sont pas uniquement dédiées au signal de la parole. Ces méthodes sont brièvement décrites dans le chapitre 10.

### 1.3.4 Approche retenue

Nous avons choisi pour notre étude l'approche analytique qui nous semble être la plus raisonnable. En effet, le but étant d'une part de mieux comprendre les effets de la distance sur le signal de la parole pour pouvoir par la suite réaliser des transformations de voix modale vers des voix criées ou chuchotées. Ainsi, une approche analytique de type source-filtre nous permettra de réaliser des relevés de paramètres, des interprétations et des transformations de façon plus abordable que les deux autres méthodes citées ci-dessus ; même si l'approche via la modélisation statistique constitue sans doute une autre façon de faire, nous n'abordons pas cette approche puisque nous avons voulu volontairement mener une étude destinée à la compréhension et la description de la voix créée. Quant à l'approche physiologique, bien qu'elle soit très intéressante, elle semble à nos yeux trop complexes pour être réalisable. Par un compromis de complexité-faisabilité, nous retenons l'approche analytique basée sur le modèle source-filtre de la parole.

Concrètement, nous cherchons à élaborer des règles de transformation à partir de certains paramètres pertinents du signal de la parole permettant de modifier le niveau d'effort vocal ressenti par un auditeur. En absence de relation claire entre la distance et l'effort vocal, nous utiliserons la distance entre un locuteur et son interlocuteur comme indicateur de l'effort vocal. Les différents niveaux d'effort vocal seront donc représentés par différentes distances de communication.

Une fois les règles de transformation définies, nous les intégrons dans un algorithme d'analyse-synthèse permettant de transformer une voix modale. Nous réalisons des tests perceptifs afin de juger de la pertinence des règles de transformation ainsi que de l'algorithme de transformation.

Il est important de préciser que nos travaux ne concernent que la voix créée liée à la communication à distance et non pas toutes les voix criées que l'on utilise dans des situations telles que l'énervement,

l'autorité, lorsqu'on s'effraye ou encore lorsqu'on est plongé dans le bruit (effet Lombard, (Lombard, 1911)).

*Nous parlerons ainsi, dans ce manuscrit de voix créée uniquement dans le sens de l'augmentation de l'effort de production dans le but d'une communication verbale à distance.*

## 1.4 Restriction de l'étude

La modélisation de la parole, quelle que soit l'approche utilisée, fait encore l'objet de nombreuses recherches. En effet, certains mécanismes sont encore mal connus voire même inconnus à ce jour. C'est en partie dû à ces inconnues qu'on arrive encore aujourd'hui, malgré les avancées extraordinaires de ces 50 dernières années dans les domaines de la modélisation et de la synthèse des voix, à reconnaître une voix de synthèse, une voix transformée ou encore une voix convertie. Ainsi, ayant parfaitement conscience de ces limitations actuelles, il n'est pas réaliste d'envisager une transformation imperceptible de voix modale en voix chuchotée et encore moins en voix créée où des modifications drastiques de paramètres doivent être appliquées. C'est pourquoi, notre objectif premier est de transformer la voix dans le but de faire ressentir un changement d'effort vocal qui à son tour évoquerait la distance d'un locuteur ; même si les qualités de ces transformations ne sont ni optimales ni parfaitement naturelles. De plus, faute de connaissances approfondies sur le sujet, l'utilisation de certaines règles empiriques nous semble, au premier abord, inévitable.

Dû au nombre considérable de difficultés et problématiques, notre étude se limite à l'objectif de reproduire, par transformation de la voix, 3 plages qualitatives de distance. A savoir, une transformation de voix modale en voix chuchotée (pour évoquer une distance courte) et une transformation de voix modale en voix créée (pour évoquer une grande distance de communication). On pourrait reprendre les suggestions de Erten (2005) dans le cadre de l'appel de l'AFRL faites dans un brevet provisoire. Ce brevet provisoire décrit un système très similaire à notre application où 4 zones de production vocal, en fonction de la distance de communication, sont mentionnées. Ainsi Erten (2005) définit :

1. une zone de 0 à 0,3 m pour les voix chuchotées,
2. une zone de 0,3 à 3 m pour les voix conversationnelles,
3. une zone de 3 m à 30 m pour les voix fortes et criées, et
4. une zone de 30 m à 100 m pour les voix hurlées.

En première approche, notre étude ne considérera que 3 zones (en fusionnant la zone 2 et 3). Toutefois, il sera évoqué des aspects et pistes permettant d'orienter les transformations de voix vers un continuum de distance de communication.

En pratique, la réalisation du système final présenté en Figure 1-1 pose plusieurs problèmes. D'abord, le signal de la parole du locuteur, risque d'être dans la majorité des cas extrêmement bruité (sur un champ de bataille par exemple). Ceci aurait un effet catastrophique pour les transformations de voix. La possibilité que le type de phonation en entrée du système soit différent d'une voix modale n'est pas à exclure. Dans certain cas, il est tout à fait probable que la voix du locuteur soit une voix chuchotée, criée ou stressée (approche furtive, situation stressante, ...). Dans une telle situation, les règles que nous tentons d'établir ne seront plus valables. Ensuite, les systèmes de radiocommunication ayant la capacité de transmettre le signal de la parole à des distances bien plus grandes que celles de la portée de la voix, une solution permettant d'indiquer que le locuteur se trouve hors de portée de voix devra être envisagée. On pourrait parfaitement imaginer une transformation résultant en une voix non naturelle permettant de faire comprendre qu'il ne s'agit plus de réalité virtuelle dans ce cas précis mais d'une réalité augmentée. Un dernier aspect problématique, et non des moindres est le fonctionnement temps réel du système. Dans le cadre d'une radiocommunication, le traitement temps réel est impératif. Toutefois, on estime qu'un retard inférieur à 200 ms ne perturbe pas la communication entre utilisateurs (ITU-T G.114, 2003).

De toute évidence, tous ces problèmes ne peuvent pas être résolus dans une seule thèse. C'est pourquoi, nos travaux se concentrent principalement sur le problème de fond lié à la réalisation de ce système **à savoir :**

*La transformation de l'effort vocal à partir d'une voix modale, sans bruit de fond, permettant d'indiquer la distance d'un locuteur et de conserver une bonne intelligibilité du message.*

## 1.5 Organisation du document

Ce manuscrit se décompose en trois parties. La première partie a pour but principal la réalisation d'un état de l'art des acquis des travaux antérieurs. Ainsi, le chapitre 2 est consacré à un état de l'art concernant la perception auditive de la distance (PAD). Nous verrons que la PAD d'un locuteur est essentiellement réalisée via la perception de l'effort vocal fourni par celui-ci. Le chapitre suivant (chapitre 3) rappelle les différents éléments concernant la production de la parole et la représentation du signal de la parole. Dans ce dernier chapitre, nous évoquons également les différents aspects liés à

la variabilité de la parole. Le chapitre 4, quant à lui est dédié à définir le terme effort vocal qui n'est pas interprété de la même manière en fonction des études dans lesquelles il est employé. La fin de cette partie est alors consacrée à l'état de l'art concernant les paramètres des voix chuchotées et des voix criées (Chapitre 5).

La deuxième partie constitue un élément central de notre étude qui décrit les différentes étapes qui ont pour objectif final de mettre en place des règles de transformation. Il s'agit d'enregistrer des voix produites pour différentes distances de communication, puis d'analyser dans le plus grand détail ces enregistrements afin d'extraire les informations pertinentes permettant de décrire une voix créée. Ainsi, le chapitre 6 décrit les enregistrements des différents corpus que nous utilisons dans cette étude. Vu le grand nombre de paramètres modifiés dans une voix créée liée à une communication à distance, le chapitre 7 est consacré à l'identification des éléments les plus pertinents pour la perception de l'effort vocal. Les chapitres 8 et 9 sont dédiés aux analyses effectuées sur les corpus décrit auparavant, dans le chapitre 6. Ces analyses nous permettent d'identifier certaines caractéristiques des voix criées, souvent négligées dans la littérature. Le chapitre 8 se concentre principalement sur une étude statistique des variations des paramètres de la voix créée en fonction de l'effort vocal. Différents éléments liés à la prosodie ont été identifiés comme variant fortement avec la distance (et donc avec l'effort vocal) dans ce chapitre. Le chapitre 9 approfondit la notion de modification prosodique, avec l'augmentation de l'effort vocal, sur la base d'un corpus dédié à cette étude. Ce chapitre se concentre sur des structures phonétiques de base permettant à partir d'analyses prosodiques approfondies de mettre en place les règles de transformation pour le passage de voix parlée en voix créée.

La dernière partie de cette thèse est consacrée à l'élaboration de règles de transformation permettant de passer d'une voix parlée à une voix créée ou chuchotée, ainsi qu'à la description des algorithmes développés pour réaliser ces transformations. Soulignons que l'élaboration des règles de transformation exige non seulement une très bonne connaissance de la voix créée (ou chuchotée) mais également une forte expérience. Quant à l'algorithme de transformation proprement dit, il joue un rôle primordial pour garantir la qualité de la voix transformée. Nous décrivons en détail ces points dans le chapitre 10. Pour valider notre démarche, nous avons réalisé des tests perceptifs comprenant notamment un test d'intelligibilité et un test concernant la perception de la distance. Nous décrivons les résultats de ces tests dans le chapitre 11. Nous concluons ce travail en mentionnant certaines perspectives ainsi que des domaines d'applications possibles de ce travail.

# **PARTIE I :**

## **FONDAMENTAUX ET ACQUIS**



# Perception auditive de distance (PAD)

---

*« On ne perçoit du monde que ce qu'on est préparé à en percevoir »*  
— Bernard Weber

La perception auditive de la distance (PAD) est une capacité de l'Homme qu'il utilise au quotidien. Par le biais des connaissances qu'il a pu acquérir au cours de sa vie, il est capable d'interpréter des phénomènes acoustiques lui permettant de juger de la distance d'une source sonore. En effet, les connaissances quant à l'environnement dans lequel il se situe et le type de source sonore qu'il perçoit, lui permettent d'associer à des phénomènes acoustiques qu'il perçoit des situations déjà rencontrées. Ainsi, pour de nouvelles situations, l'Homme fait preuve de capacités d'apprentissage pour accroître la précision de son estimation de distance.

De toute évidence, des paramètres non acoustiques permettent également d'estimer la distance d'une source sonore, comme la vision ou encore la connaissance, *a priori* de la typologie des lieux et de la provenance de la source sonore. Le paramètre non acoustique le plus important étant la familiarité que nous avons déjà évoquée. Hormis pour ce dernier paramètre, nous n'aborderons pas au cours de ce chapitre les différents points liés à l'usage des paramètres non acoustiques pour la PAD. Nous nous concentrons particulièrement sur le rôle des phénomènes acoustiques. Nous pouvons classer, d'une manière générale, les paramètres acoustiques en deux catégories : les paramètres environnementaux, et les paramètres dynamiques. Les paramètres environnementaux sont les modifications du signal acoustique directement liées à l'environnement de propagation de l'onde acoustique. Les paramètres dynamiques quant à eux, sont liés au mouvement de la source elle-même ou au mouvement de l'auditeur.

Du fait du grand nombre de références relatant l'état de l'art de la PAD (Botte et al., 1990; Cochran et al., 1968; Coleman, 1963; Middlebrooks et Green, 1991; Moore, 2004; Naguib, 2001; Ohta et Tamura, 1999; Zahorik, 1996; Zahorik et al., 2005), nous n'aborderons dans ce chapitre que brièvement le rôle de chacun des paramètres acoustiques. Dans un second temps, nous nous intéressons tout

particulièrement à la PAD d'un locuteur par le biais de sa voix. Nous renvoyons ainsi le lecteur plus intéressé par ces différents points aux états de l'art mentionnés.

## 2.1 PAD liés par les phénomènes acoustiques

Dans cette section nous nous intéressons aux phénomènes acoustiques qui influent sur la perception auditive de la distance. Les 3 facteurs les plus pertinents sont présentés : (1) les variations d'intensité, (2) les variations de rapport d'énergie entre l'onde directe et les ondes réverbérées ainsi que (3) les variations spectrales engendrées par l'absorption atmosphérique. Pour chacun de ces points nous débutons par un rappel théorique sur les causes de ces phénomènes puis, nous abordons leurs conséquences sur la perception auditive de la distance.

### 2.1.1 L'intensité

Lorsqu'une source sonore s'éloigne son intensité diminue. Ce phénomène s'explique d'ailleurs simplement.

Une source sonore ponctuelle omnidirectionnelle de puissance  $P$  (en W) diffuse sa puissance le long d'un front d'onde sphérique. L'intensité sonore à la surface de cette onde est alors :  $I = \frac{P}{4\pi r^2}$ , (en  $W/m^2$ ) où  $r$  est le rayon du front d'onde. L'intensité est donc fonction de l'inverse de la distance au carré. D'où son nom : la loi du carré inverse. Ainsi, si la distance  $r$  est doublée,  $I$  obéit à cette loi et est alors diminuée d'un facteur quatre. Pour un doublement de distance on peut alors calculer la différence d'intensité en décibel (dB):

$$\begin{aligned} \Delta I_{d1,d2} &= 10 \log_{10}(I1) - 10 \log_{10}(I2) \\ &= 10 \log_{10}\left(\frac{P}{4\pi r^2}\right) - 10 \log_{10}\left(\frac{P}{4\pi (2r)^2}\right) \quad (\text{Eq. 2-1}) \\ &= 10 \log_{10}(4) = 6.02 \text{ dB} \end{aligned}$$

Où  $I1$  est l'intensité à une distance  $r$  de la source et  $I2$  l'intensité à une distance doublée ( $2r$ ). Le niveau d'intensité sonore diminue alors de 6 dB pour chaque doublement de distance.

Il s'agit d'un cas idéal. En effet, plusieurs facteurs ne sont pas pris en compte dans cette loi. On pourrait citer notamment le cas d'un environnement réverbérant (type église) où en chaque point du front d'ondes direct viendra s'ajouter les ondes réverbérées. Il n'est donc pas rare d'observer des diminutions d'intensité moins forte, comme décrit par exemple par Zahorik et Wightman (2001) qui

mesure -4,25 dB par doublement de distance en environnement réverbérant. D'autre part, cette loi ne tient pas compte des phénomènes d'absorption atmosphérique, qui génèrent également des modifications d'intensité décrites par la suite.

L'intensité est depuis longtemps considérée et reconnue comme un facteur utile à la perception auditive de la distance. En 1882, Thomson évoque déjà le rôle de l'intensité dans la perception de la distance :

*“In the case of known sounds we doubtless judge chiefly of their distance by their relative loudness, the intensity decrease inversely as the square of the distance.”*  
(Thompson, 1882).

Ainsi, en diminuant l'intensité d'une source sonore, on peut créer une sensation d'éloignement de celle-ci. De la même manière, une augmentation de l'intensité d'une source sonore, engendre une sensation de rapprochement de cette source (Gardner, 1969).

Tout en évoquant le fait que l'intensité est un facteur pertinent pour la perception de la distance, Thomson suggère que ceci est vrai uniquement dans le cas de connaissances *a priori* sur l'intensité de la source. L'intensité constitue donc un facteur relatif dans la perception auditive de la distance. En effet, Mershon et King (1975) montrent dans une expérience que des sons présentés pour la première fois aux auditeurs à des intensités différentes (pour une plage de variation de 20 dB), ne permettent pas d'estimer la distance de la source. Toutefois, après plusieurs présentations du son, l'estimation de la distance devient plus précise du fait de l'apprentissage des caractéristiques de la source sonore par les locuteurs.

D'après la théorie, une diminution de 6 dB de l'intensité d'une source sonore correspond à une distance doublée. Ainsi, 6 dB de diminution en intensité devraient évoquer une distance doublée. Or il a été montré à plusieurs reprises qu'il est nécessaire que la diminution d'intensité d'une source sonore soit supérieure à 6 dB pour percevoir un doublement de la distance. Le Tableau 2-1 résume les valeurs de diminution d'intensité, nécessaires à la perception d'un doublement de distance estimées par plusieurs auteurs. Hormis la valeur obtenue par Warren (1968), toutes les valeurs reportées par ces études sont supérieures à la diminution théorique de 6 dB.

Du fait de ces écarts d'intensité nécessaires pour percevoir un doublement de distance, plusieurs auteurs ont constaté que la distance perçue augmente moins rapidement que la distance réelle, pour des distances supérieures à 1 mètre (Von Békésy, 1949; Cochran et al., 1968; Haustein, 1969; Simpson et Stanton, 1973). Pour Cochran et al. (1968), on observe par exemple une sous-estimation de l'ordre de 30 % pour une distance de 29 m.

Tableau 2-1 : Variations d'intensité nécessaire pour percevoir un doublement de la distance en environnement anéchoïque (selon (Zahorik, 1996))

(Stevens et Guirao, 1962)	10 dB
(Gardner, 1969)	20 – 30 dB
(Mershon et King, 1975)	15 dB
(Warren, 1968)	6 dB
(Sheeline, 1983)	10 – 21 dB
(Petersen, 1990)	21 dB
(Begault, 1991)	9 dB

Si les écarts sont plus grands que la théorie en termes de perception de distance doublée, qu'en est-il concernant les niveaux juste perceptibles de distance, via le facteur intensité ? En théorie, la différence d'intensité juste perceptible, est de l'ordre de 1 dB ((Jesteadt et al., 1977; Miller, 1947; Riesz, 1933) d'après (Zahorik et al., 2005)). Ainsi, sur la base de la diminution théorique juste perceptible de l'intensité, on peut en déduire le niveau de différence de distance juste perceptible comme suit :

$$\Delta I_{d_1, d_2} = 1 \text{ dB} = 20 \cdot \log_{10} \frac{d_1}{d_2} \quad (\text{Eq. 2-2})$$

$$\frac{d_1}{d_2} = 10^{1/20} = 0,122 = 12,20\%$$

Ainsi, en théorie une variation de distance de l'ordre de 12.2 % serait juste perceptible. Or, on trouve dans la littérature, des niveaux juste perceptibles très variés. Ce niveau de perception est de l'ordre de 20% pour Edwards (1955) et Gramble (1909) . Simpson et Stanton (1973) dans une chambre anéchoïque, trouvent des valeurs de 13 % à 48 % pour des distances variant de 0,61 m à 2,13 m. Ils démontrent également que les niveaux sont plus faibles pour des stimuli qui s'approchent. Strybel et Perott (1984), quant à eux, au cours d'expériences en extérieur, mesurent des valeurs variant de 9 % à 20 % pour des distances variant de 0,5 m à 3 m et des valeurs de 3 % à 7 % pour des distances variant de 6 à 49 m. De la même manière, pour des distances proches variant de 1 à 2 mètres en environnement anéchoïque, Ashmead et al. (1990) relèvent un niveau de discrimination de distance de 6 %.

Nous retrouvons donc des niveaux supérieurs et inférieurs à la valeur théorique. En général, les niveaux largement surestimés, sont mesurés pour des distances proches car la loi du carré inverse n'est plus applicable (Brungart et al., 1999). En revanche, on retrouve dans les études de Ashmead et al. (1990) et Strybel et Perott (1984), des niveaux bien inférieurs qui pourraient s'interpréter, d'après

Zahorik et al. (2005), par le fait que d'autres facteurs, en plus de la variation d'intensité, sont pris en compte lors de la discrimination de distance, et notamment dans le cas de l'étude de Strybel et Perott (1984) qui se déroule en environnement réverbérant.

Au regard de ces études, l'intensité constitue donc un facteur **relatif** pour la perception de la distance, mais reste assez peu précis pour la discrimination de distance. Une explication directe est que notre oreille ne constitue pas un récepteur linéaire parfait. Ainsi, on ne juge pas la force d'un son par son intensité physique mesurée, mais par sa sonie (i.e. la sensation d'intensité). Dans ce sens Stevens et Guirao (1962) ont étudié le rôle de la sonie dans la perception de distance. Ils en concluent que la perception de sonie est exactement inversement proportionnelle à la perception de la distance. Ce qui va dans le sens de la théorie de Warren (1977), qui énonce le fait que pour une sonie (et non pas l'intensité) diminuée de moitié, engendrera une distance perçue doublée.

## 2.1.2 Les variations spectrales

### 2.1.2.1 Influence directe

Lors de sa propagation dans l'air, une onde sonore subit des modifications spectrales dues aux propriétés acoustiques de l'air. L'onde subit une atténuation atmosphérique en fonction de l'humidité de l'air, de la pression atmosphérique, de la température, de la distance à parcourir et de sa fréquence. En particulier, la propagation du son dans l'air engendre une absorption sélective des hautes fréquences. Pour des distances supérieures à 15 m, les propriétés de l'air modifient le spectre d'un son de façon significative (Blauert, 1997). Typiquement, les variations spectrales apparaissent au delà de 2 kHz et sont de l'ordre de -3 dB à -4 dB pour un son de fréquence 4 kHz, un taux d'humidité de 40 % et une distance de 100 m (Ingard, 1953). La Figure 2-1 représente les courbes d'absorption en fonction du rapport fréquence-pression atmosphérique pour plusieurs taux d'humidité.

Cette absorption des hautes fréquences résulte de trois contributions liées à des pertes d'énergie causées par les vibrations des molécules d'azote et d'oxygène. Jusqu'à 500 Hz ce sont les effets de l'azote (ou Nitrogène en anglais) qui prédominent. De 500 Hz à 30 kHz c'est l'oxygène qui prédomine et au-delà de 30 kHz, c'est l'absorption dite classique (absorption générée par des effets de viscosité et de conduction thermique de l'air (Piercy, 1977)) qui engendre une atténuation (cf. Figure 2-2). Notons que pour des sons audibles (de 20 Hz à 20 kHz), l'effet de l'absorption classique (qui se manifeste pour des fréquences supérieures à 30 kHz) n'est pas perceptible.

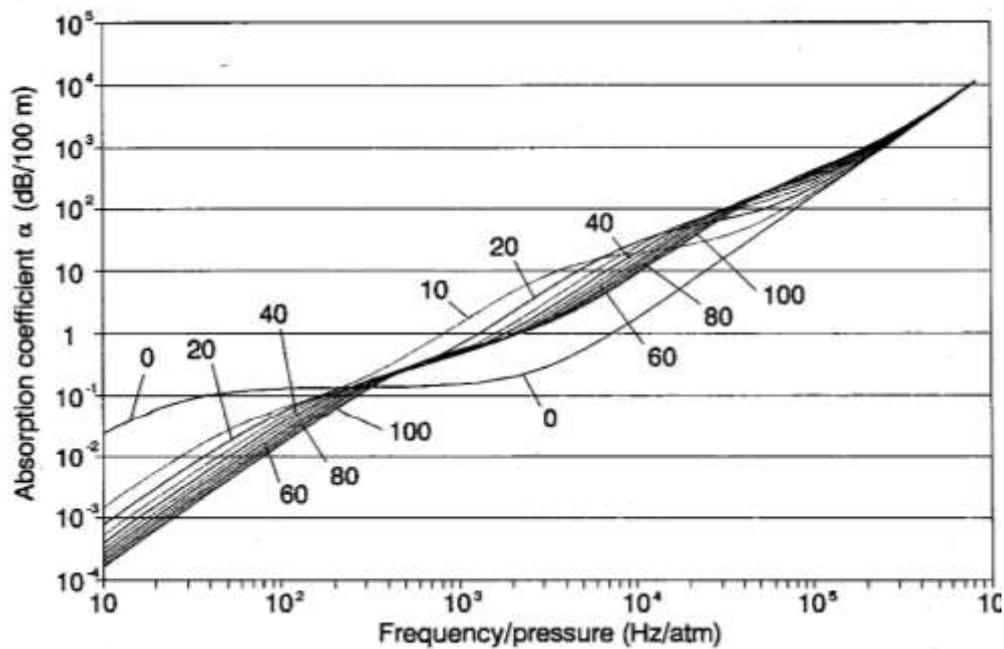


Figure 2-1 : Coefficient d'absorption dans l'air (en dB/100m) en fonction du rapport fréquence pression atmosphérique pour des différents valeurs d'humidité et une température de 20°C. (d'après (Bass, 1990))

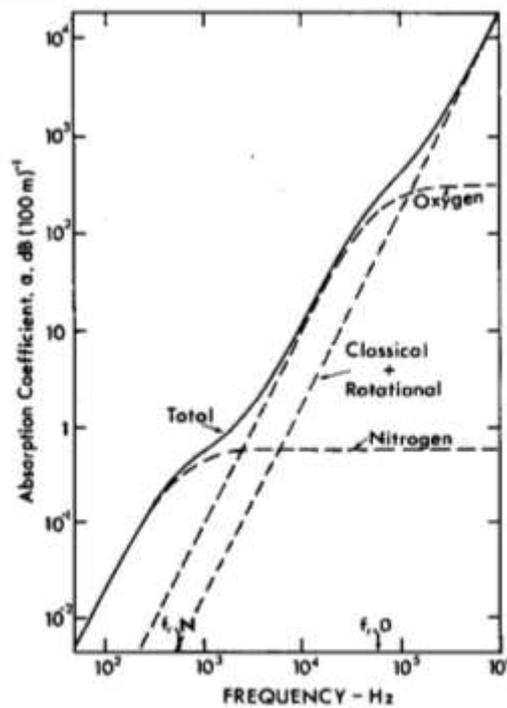


Figure 2-2 : Coefficient d'absorption atmosphérique estimé en fonction de la fréquence, en dB/100m pour une pression atmosphérique de 1 atm, une température de 20°C et un taux d'humidité de 70% (issu de (Piercy, 1977))

Malgré la faible ampleur de ces modifications, plusieurs auteurs ont démontré que ce facteur joue un rôle non négligeable pour percevoir la distance. Des sons dont les hautes fréquences auraient été filtrées, sont systématiquement perçus comme plus éloignés que des sons ayant une composition en

haute fréquence plus énergétique (Butler et al., 1980; Coleman, 1968; Little et al., 1992; Lounsbury et Butler, 1979). Toutefois, la majorité de ces études, appliquent uniquement un filtrage des hautes fréquences, sans tenir compte des ordres de grandeur de l'atténuation atmosphérique causée par l'air. Ainsi, le filtrage des hautes fréquences, dans ces études, est plus intense que celui que la propagation aurait engendrée (Zahorik, 1996).

De la même manière que pour l'intensité, l'utilisation de ce facteur demande d'être familier avec la source sonore (Coleman, 1962, 1963). Ainsi, la nécessité d'utilisation d'un son de référence que l'on entend, ou que l'on connaît, fait de ce facteur un facteur relatif et non pas absolu pour la perception de la distance. Pour un son non familier, il a été démontré que plusieurs expositions permettent d'améliorer la perception relative de la distance dû à l'apprentissage que le locuteur fait au fil des tests (Little et al., 1992)

### *2.1.2.2 Influence indirecte*

Hormis l'influence directe de la propagation de l'air sur le spectre d'un signal, un deuxième phénomène peut également engendrer des variations spectrales. En effet, dans un environnement réverbérant, chaque onde réfléchi par les parois subie également des modifications spectrales. En effet, en fonction des propriétés acoustiques des parois, celle-ci vont absorber certaines fréquences du son incident, pour finalement ne réfléchir qu'une partie du signal. De plus, comme nous le verrons dans la section suivante, plus la distance augmente, plus la part des ondes réfléchies est importante au niveau des oreilles de l'auditeur. Ceci peut potentiellement engendrer des modifications spectrales du signal. Ainsi, chaque onde réfléchi subie une modification spectrale due aux parois et la superposition des ondes au niveau de l'oreille, peut à son tour engendrer des modifications spectrales. Ces changements peuvent toutefois être assimilés à des effets indirects de rapport d'énergie (Zahorik et al., 2005).

Pour des sons proches, par contre, aucune absorption des hautes fréquences n'est constatée due à la faible distance parcourue par l'onde. Cependant, dans le cas de sons proches, ce sont les basses fréquences qui sont modifiées. En effet, par masquage de la tête, du torse, ainsi que les réflexions associées, le son arrivant aux oreilles de l'auditeur a une composition spectrale différente dans les basses fréquences (Brungart, 1999; Brungart et al., 1999). On constate également que la différence interaurale d'intensité (ILD) est plus grande que la normale pour des sons proches, ce qui permet également d'estimer la distance d'une source (Brungart, 1999; Brungart et al., 1999).

### 2.1.2.3 Rôle de l'absorption atmosphérique

Nous avons précisé que la majorité des études, sont en accord avec le fait qu'un filtrage des hautes fréquences d'un signal permet d'apporter une sensation d'éloignement de la source. Les valeurs utilisées pour ces filtrages ne sont toutefois pas en accord avec les valeurs théoriques de l'absorption atmosphérique. Nous avons donc décidé de mener une étude concernant le problème de la pertinence des effets liés à la propagation du son de la parole sur la perception de distance (Lobréau, 2010). Cette étude a utilisé un modèle de propagation acoustique (Modèle de Chien et Soroka (Chien, 1975)) afin de simuler la propagation d'un signal de parole criée, sur des distances de 5 m, 50 m et 100 m, sur deux types de sols différents (herbe et béton). Les voix criées utilisées sont issues d'une situation simulée de communication à une distance de 120 m (cf. §6.3 pour des précisions sur le protocole de simulation).

Dans cette étude nous ne considérons qu'un cadre parfait. En effet, elle prend en compte l'absorption atmosphérique réelle des hautes fréquences faite par l'air et une seule réflexion faite par le sol (i.e. milieu quasi anéchoïque). La simulation a été réalisée dans les conditions atmosphériques suivantes : une température de 20°C, un taux d'humidité de 50% et une pression atmosphérique de 1 atm. Ce qui correspond à une atténuation de -3 dB/100m à 4 kHz. Sur la Figure 2-3 on peut voir le schéma de principe de la simulation. L'onde directe est affectée par l'absorption atmosphérique ainsi qu'un délai de propagation. L'onde réfléchie est quant à elle, affectée par l'absorption atmosphérique, puis réfléchi par le sol et est modifiée en composition spectrale et en phase par les propriétés du sol. Cette onde subit alors à nouveau les effets de l'absorption atmosphérique et d'un délai de propagation avant de se superposer à l'onde directe au niveau de l'auditeur. Notons que la diminution d'intensité due au rayonnement n'est pas prise en compte dans notre étude.

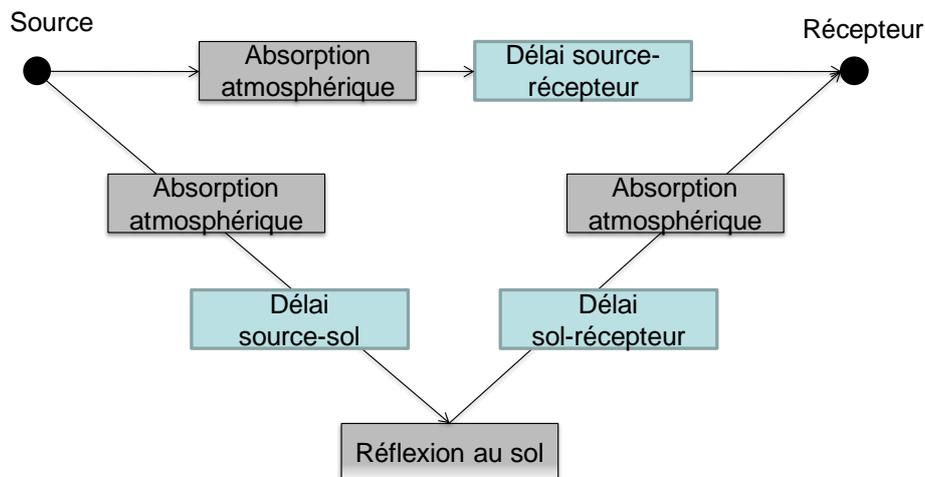


Figure 2-3: Schéma de principe de la simulation de distance utilisé dans l'étude (Lobréau, 2010).

En effet, la sonie du signal au niveau du récepteur est maintenue constante. Ceci dans le but de réaliser un test perceptif uniquement sur la base de l'absorption atmosphérique et de la réflexion au sol. Par la suite un test perceptif sur la base de comparaisons entre les différentes simulations a été effectué. Le but du test perceptif était de comparer deux voix criées simulées, à des distances différentes et d'indiquer laquelle paraissait provenir de plus loin. L'ensemble des comparaisons a été réalisé lors de ces tests (5 m-50 m, 50 m-100 m et 5 m-100 m).

Pour les simulations sur le béton et sur l'herbe, les résultats sont similaires : les simulations plus proches sont perçues comme tel dans 64 % des cas. Toutefois, dans le cas où l'intensité est également simulée, les résultats sont dans ce cas sans appel et les simulations les plus lointaines sont dans 97 % des cas, perçues comme provenant d'une distance plus éloignée (cf. Figure 2-4). Notons que cette expérience ne tient également compte que d'une seule réflexion faite par le sol.

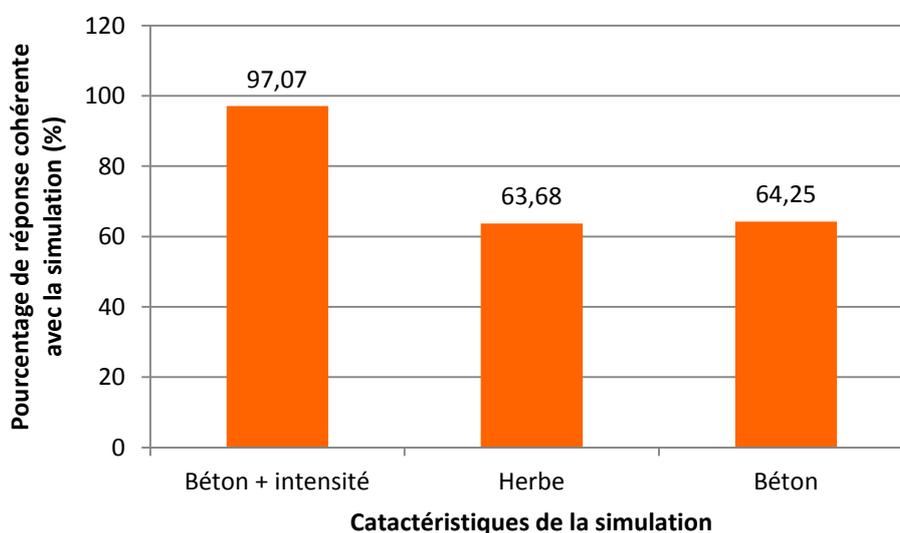


Figure 2-4 : Résultats des tests perceptifs de l'étude (perception avec radiation d'intensité, perception à sonie égale sur du béton et sur de l'herbe). Parmi l'ensemble des situations proposées, les réponses cohérentes (i.e. réponse en accord avec la théorie) sont reportées sur le graphique. (Lobréau, 2010).

D'après les résultats de cette étude nous tirons plusieurs conclusions. Tout d'abord l'apport d'informations (i.e. réflexion au sol plus absorption atmosphérique) ne permet pas de différencier de façon significative deux distances (ici 5 m, 50 m et 100 m). D'autre part, le type de sol n'a aucune influence sur les résultats. Enfin, l'intensité constitue un facteur non négligeable pour la perception de la distance; sans doute car aucune autre information n'est exploitable lors de ces tests perceptifs. (La réflexion au sol et l'absorption atmosphérique étant imperceptible d'après nos résultats).

Ainsi, par cette étude nous émettons certains doutes sur l'intérêt perceptif de l'absorption atmosphérique pour la perception auditive de voix. Rappelons cependant que cette étude ne simule qu'une distance maximale de 100 m, en utilisant un signal de la parole en voix criée. Le signal de la

parole possède d'une manière générale un spectre qui diminue de 6 dB/octave. Ainsi, la déperdition des hautes fréquences qui apparaît au-delà de 2 kHz de façon très faible, n'est sans doute pas perceptible pour le signal de la voix. Toutefois, cette étude ne permet pas de réaliser des conclusions plus générales, étant donné que la voix utilisée est une voix criée. Pour les voix criées, l'absorption atmosphérique ne semble pas être pertinente pour la perception de distance (du moins dans le cas que nous avons simulé).

Les résultats de cette étude montrent qu'à eux seuls les paramètres liés à la propagation de l'onde sonore (sans prise en compte de l'intensité), dans un cas parfait, ne permettent pas d'apporter d'informations sur la distance d'un locuteur (pour des conditions atmosphériques de 20 °C, 50 % d'humidité et 1 atm).

### 2.1.3 Rapport de l'énergie direct à l'énergie réverbérée

Le rapport entre l'énergie directe et l'énergie réverbérée constitue un facteur important dans la perception auditive de la distance. Pour illustrer ces phénomènes prenons l'exemple de la Figure 2-5 où une source sonore est placée dans une pièce fermée. L'auditeur, perçoit dans un premier temps l'onde directe issue de la source sonore. Dans le cas où cette source sonore n'est pas purement directive, le front d'onde va également venir percuter les parois, le sol et le plafond de la pièce qui réfléchissent ces ondes sonores. Chacune de ces ondes réverbérées à une intensité sonore plus faible que l'onde directe (d'après la loi du carré inverse). En effet, la distance parcourue par les ondes réverbérées est plus grande que celle parcourue par l'onde directe. D'autre part, les parois, en fonction de leurs propriétés acoustiques, absorbent une partie de leurs énergies. Ainsi l'auditeur perçoit une onde directe, plus un nombre infini d'ondes réverbérées d'intensité, de délais et de compositions spectrales différentes. On comprend alors que plus la distance entre la source et l'auditeur augmente plus la part des ondes réverbérées sera forte. En effet, plus la source s'éloigne plus le trajet effectué par les ondes réverbérées sera faible par rapport au trajet de l'onde directe. Le rapport entre l'énergie de l'onde directe par rapport aux ondes réverbérées est donc inversement proportionnel à la distance. Ce rapport diminue à mesure que la distance source/récepteur augmente

Von Békésy (1938) est sans doute le premier à avoir démontré que la diminution du rapport d'énergie permet de créer une sensation d'éloignement. Par la suite, Mershon et King (1975) réalisent une expérience de perception de distance apparente, utilisant des séquences de bruit blanc à deux niveaux d'intensité (fort et faible), présentés à des distances de 2,74 m et 5,49 m en environnement anéchoïque et en environnement réverbérant. Suite à cette expérience ces auteurs observent que l'estimation de distance est plus précise en environnement réverbérant qu'en environnement anéchoïque où le rapport d'énergie est absent.

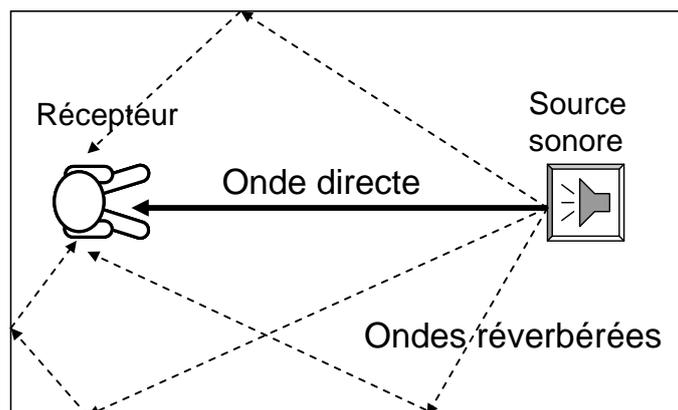


Figure 2-5: Schéma de réflexion

L'étude de Mershon et King (1975) montre également qu'après plusieurs présentations du son, ils observent des résultats constants pour un environnement réverbérant mais des estimations de plus en plus précises dans des environnements anéchoïques. Cette conclusion va également dans le sens de Mershon et Bowers (1979) qui montrent une estimation précise de la distance en environnement réverbérant dès la première écoute. De plus, même si l'environnement réverbérant n'est pas connu par l'auditeur, celui-ci fait preuve d'apprentissage afin de réaliser une meilleure estimation de distance (Kopčo et al., 2004; Shinn-Cunningham, 2000). Ces résultats démontrent que le rapport d'énergie peut être utilisé comme facteur absolu, alors que l'intensité (seul facteur présent dans l'environnement anéchoïque) demande un apprentissage préalable et ne peut donc être utilisé comme facteur absolu.

Les changements de rapport d'énergie sont toutefois peu perceptibles. Zahorik (2002) mesure des niveaux juste perceptibles de l'ordre de 5 à 6 dB. Le rapport d'énergie est un facteur relatif assez pauvre et peu précis. Le rôle du rapport est peut être destiné à donner une information absolue de la distance, plutôt que de discriminer de faibles variations de distance plus faciles à réaliser sur la base de l'intensité (Zahorik et al., 2005).

## 2.2 Autres phénomènes intervenant dans la PAD

Si le son n'est pas présenté en face de la personne les sujets utilisent également la binauralité (Coleman, 1963). L'utilisation de la binauralité n'est vrai que pour des distances inférieures à 1 mètre, car on peut constater que les filtres HRTFs ne sont pas influencés pour des distances supérieures à 1 mètre (Brungart et Rabinowitz, 1999; Brungart, 1999; Brungart et al., 1999) (cf. (Zahorik et al., 2005) pour un état de l'art et une discussion sur le sujet).

Dans les cas précédents, la source et l'auditeur sont fixes. Or, si la source est mobile, trois facteurs appelés facteurs dynamiques interviennent dans la perception de distance :

- (1) Le parallaxe de mouvement : induit par un changement de direction de la source et provoquant des modifications des indices acoustiques.
- (2) Le  $\tau$  acoustique : La durée que va mettre un objet en mouvement émettant un signal sonore pour arriver jusqu'à l'auditeur.
- (3) L'effet Doppler : Effet acoustique qui engendre une augmentation de la fréquence d'une source qui se rapproche et une diminution de la fréquence d'une source qui s'éloigne.

Dans notre étude, la source et le récepteur sont considérés constant et par conséquent nous ne développerons pas cet aspect de la perception auditive de distance. Nous renvoyons le lecteur plus intéressé à (Zahorik et al., 2005).

Par ailleurs, la présence de bruits de fond affecte également la perception de la distance (du moins pour des environnements réverbérants) (Mershon et al. 1989). Plus le niveau de bruit est élevé par rapport au niveau de la voix, plus la distance perçue décroît. En effet, le fait d'augmenter l'intensité du bruit de fond revient à diminuer le rapport entre l'onde directe et l'onde réverbérée ce qui procure une distance perçue moins grande. Néanmoins, ce constat n'est pas en accord avec la théorie qui se contente d'une explication directement liée au niveau acoustique. En effet, augmenter le niveau de bruit de fond sera équivalent à une diminution relative du niveau de la source et donc à une distance perçue plus grande. On peut alors penser que dans un milieu anéchoïque cette explication prévaudrait sur la première. C'est par exemple ce qui se passe dans un cas concret en extérieur. En effet, lorsqu'un locuteur nous parle à une distance relativement proche, le bruit de fond n'est pas très intense. Toutefois, si le locuteur s'éloigne et que l'on perçoit des chants d'oiseaux qui deviennent plus forts par rapport au niveau de la voix du locuteur nous percevons une sensation naturelle d'éloignement du locuteur.

## 2.3 PAD par la voix du locuteur

---

Comme nous l'avons déjà évoqué, la PAD dépend notamment des connaissances *a priori* de l'auditeur sur la source sonore. La voix est sans doute le signal sonore le plus familier. Ainsi, les études de la perception auditive de la distance se sont intéressées à la PAD d'un locuteur grâce à sa voix.

Parmi les premiers types d'expériences réalisées sur la perception auditive par la voix, on peut citer entre autre l'étude de Von Békésy (1949). Dans une chambre anéchoïque, il a placé un locuteur proche d'une des parois. Il fait ensuite varier la position de l'auditeur d'une distance inférieure à 1 mètre jusqu'à 10 mètres de la source. Dans cette expérience il a utilisé 5 auditeurs seulement. Le locuteur prononce des syllabes à un niveau d'intensité moyen (non précisé) et donc *a priori* constant. Puis l'auditeur, les yeux bandés, doit estimer la distance qui le sépare de son interlocuteur. Dans cette étude, il montre qu'à mesure que la distance augmente, la distance perçue augmente. On note toutefois, que l'environnement est anéchoïque et que l'intensité de la voix du locuteur est maintenue constante. Ainsi, on se trouve plutôt dans une situation dans laquelle l'auditeur doit estimer la position d'un locuteur mais qui ne s'adresse pas forcément à lui. De ce fait, le seul facteur utilisable pour la perception de la distance est la diminution de l'intensité. Von Békésy (1949) a également montré que la distance perçue augmente moins rapidement que la distance physique. Il montre qu'au-delà d'une distance de 3 m, la distance perçue tend vers un maximum asymptotique dépendant de l'auditeur. En moyenne, ce maximum se situe vers 7-8 mètres, ce qui peut se traduire par un phénomène de limitation faite par les auditeurs. Se sachant dans une pièce fermée, les auditeurs ont conscience d'une certaine limite physique. L'auteur reproduit l'expérience avec une distance source-auditeur fixe de 10 m en remplaçant la source par un haut-parleur tout en faisant varier l'intensité de ladite source. Naturellement il observe que la distance perçue augmente à mesure que l'intensité de la source diminue. On retrouve le même type d'expérience et de conclusion chez Cochran et al. (1968).

Gardner (1969), a également effectué une expérience de perception de la distance d'un locuteur mais avec des types de phonation différente. Un locuteur dans une pièce anéchoïque était placé à quatre distances différentes d'un auditeur 0,91 m, 3 m, 6,1 m et 9,1 m (originellement 3, 10, 20 et 30 pieds). A chacune de ces distances le locuteur utilise quatre types de voix : une voix chuchotée, une voix basse (voisé) une voix conversationnelle ou une voix criée. Dix auditeurs placés sur un siège avaient les yeux bandés et devaient estimer la position du locuteur. Les résultats obtenus sont stupéfiants (cf. Figure 2-6). Les voix basses et les voix conversationnelles sont perçues à des distances très proches de la réalité (représentées par la ligne à 45° en pointillé, la ligne courbe en pointillé sur les autres graphiques représente l'incertitude des estimations). Toutefois, la distance apparente des voix chuchotées, est largement sous estimée à mesure que la distance réelle augmente, alors que la voix criée est systématiquement surestimée. Ces résultats montrent alors que la PAD est influencée par le type de voix utilisé par les locuteurs. Les auditeurs semblent alors se baser sur des situations connues et mémorisées. A savoir qu'une voix chuchotée est utilisée pour des distances proches et une voix criée pour des distances éloignées. Ce phénomène de sous estimation des voix proches et surestimation des voix lointaines a également été observé par Traunmüller (1997). Cet auteur montre également que l'intensité au niveau des oreilles de l'auditeur n'influe que peu (mais significativement) sur la perception de distance.

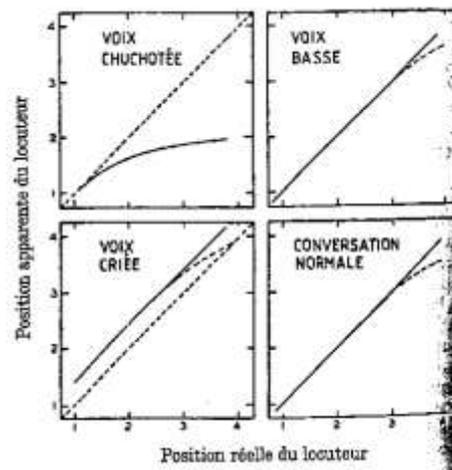


Figure 2-6 : Résultats de la perception de la distance d'un locuteur en fonction de quatre types de voix et de la distance d'un locuteur fait par (Gardner, 1969)

Philbeck et Mershon (2002) ont réalisé une étude destinée à déterminer si la capacité de jugement de distance grâce à l'effort vocal est due à une longue expérience passée ou à un apprentissage court acquis lors de l'étude. Pour ce faire, ces auteurs utilisent une seule phrase (« *How far away from you does my voice seem ?* ») prononcée par un homme et une femme en voix chuchotée, parlée et criée le plus fort possible. Les niveaux moyens et pics d'intensité ont été égalisés pour être quasiment identiques au niveau des oreilles de l'auditeur. L'expérience se déroulait dans une pièce où les voix étaient reproduites par un haut-parleur placé à 2,5 m. Les auditeurs avaient les yeux bandés et ne connaissaient pas la pièce dans laquelle l'expérience se réalisait. Ces derniers devaient durant ce test, estimer la distance de la source sonore (i.e. du haut-parleur). Dès la première présentation, les voix chuchotées sont perçues plus proches que les voix modales et que les voix criées plus loin. Leurs résultats montrent clairement que cette capacité de l'homme est due à une expérience passée et à un long apprentissage que nous avons effectué et que nous effectuons sûrement encore tous les jours.

Nous avons mené une expérience similaire qui a abouti aux mêmes résultats (Fux et Zimpfer, 2009). Dans cette expérience, quatre niveaux de voix ont été choisis, normal, fort, crié et hurlé, ainsi que cinq phrases courtes. L'ensemble des voix a été égalisé en terme de sonie et non pas d'intensité physique. Dans un test de comparaisons par paires, nous demandons ensuite aux auditeurs de comparer les distances de provenance. Au cours de ce test, les sujets avaient alors à comparer deux phrases, prononcées par le même locuteur mais d'efforts vocaux différents. Les sujets devaient répondre à la question suivante : Lequel des deux sons présentés paraît provenir de plus près ? Les résultats résumés dans le Tableau 2-2 de cette étude sont très satisfaisants. La première colonne de ce tableau correspond au niveau d'effort présenté en première position lors de la comparaison et la deuxième colonne au niveau d'effort présenté en second. La dernière colonne quant à elle, illustre le taux d'erreur effectué par les sujets en rapport aux réponses idéales, à savoir : que la voix modale doit être

perçue plus proche que la voix forte, qui elle doit être perçue plus proche que la voix criée elle-même perçue plus proche que la voix hurlée.

Tableau 2-2: Résultats de l'étude (Fux et Zimpfer, 2009)

Position N° 1	Position N° 2	Erreur
<b>Modale</b>	<b>Fort</b>	27.5%
<b>Modale</b>	<b>Criée</b>	6.5%
<b>Modale</b>	<b>Hurlée</b>	0%
<b>Fort</b>	<b>Modale</b>	5%
<b>Fort</b>	<b>Criée</b>	15.5%
<b>Fort</b>	<b>Hurlée</b>	2%
<b>Criée</b>	<b>Modale</b>	2%
<b>Criée</b>	<b>Fort</b>	10%
<b>Criée</b>	<b>Hurlée</b>	1.5%
<b>Hurlée</b>	<b>Modale</b>	2.5%
<b>Hurlée</b>	<b>Fort</b>	2%
<b>Hurlée</b>	<b>Criée</b>	8%

Malgré des erreurs plus ou moins grandes pour des niveaux d'effort relativement proches, cette étude donne de bons résultats en ce qui concerne la discrimination de la distance d'un locuteur en l'absence d'information d'intensité perçue ou de phénomène de propagation. Ainsi, plus l'effort vocal est grand plus la distance perçue est grande.

Citons enfin l'une des études les plus connues actuellement, celle de Brungart et Scott (2001). Dans cette étude, consacrée à déterminer les effets du niveau de production vocale et du niveau de présentation des voix dans la perception de distance, ces auteurs ont mené trois expériences. La première expérience utilisait des voix numériquement traitées, pour correspondre à des distances réelles de communication allant de 0,25 à 64 m. La deuxième expérience était une perception de distance en faisant varier les niveaux de production vocale de 36 dB SPL à 96 dB SPL, tout en faisant fluctuer le niveau de présentation de ces sons sur une plage de 34 dB (48 dB SPL à 82 dB SPL). La dernière expérience utilisait des voix inversées dans le temps. Cette dernière expérience avait pour but de déterminer si l'information linguistique pouvait avoir un rôle dans la perception de la distance.

Les résultats de ces deux premières expériences, montrent que l'auditeur se base sur le niveau de production et le niveau de présentation pour juger de la distance, mais de façon différente pour les trois catégories de niveau de production utilisée (chuchotée, niveau faible de parole et niveau fort de parole). Pour des voix fortes (supérieures à 66 dB à 1 mètre de la bouche), la distance perçue double quand le niveau de production augmente de 8 dB ou quand le niveau de présentation diminue de 12

dB. Pour des voix de niveau faible (<66 dB), la distance double pour chaque augmentation du niveau de production de 15 dB mais, est peu affectée par le niveau sonore de présentation. Enfin, les voix chuchotées ne sont pas affectées par le niveau de production mais diminuent si le niveau de présentation augmente. En résumé **la parole présentée avec un fort niveau sonore mais avec un faible niveau de production (chuchotement fort) indique un locuteur proche, alors que la parole avec un faible niveau de présentation et un fort niveau de production (crié de faible intensité sonore) indique un locuteur plus lointain.**

La troisième expérience a montré que les voix faibles ne sont pas affectées par cette inversion mais affecte considérablement la perception des voix fortes. En effet, pour le niveau de production le plus élevé utilisé (96 dB) la distance perçue est réduite de moitié. Les auteurs en concluent alors que les voix inversées contiennent la majorité des facteurs acoustiques pour les voix faibles, mais que dans le cas des voix fortes les informations utilisées pour percevoir la distance ne sont plus présentes. Les auteurs interprètent ces résultats en mentionnant que les informations phonétiques sont indispensables pour bien, percevoir la distance d'un locuteur. En effet dans ces sons, présentés à « l'envers », l'ensemble des éléments caractéristiques des voix criées sont toujours présents (augmentation de F0, des formants et le ratio d'énergie). Ainsi, la seule information manquante est l'information linguistique. Une seconde hypothèse est que les auditeurs n'arrivent pas à identifier une voix créée et qu'ils se basent donc uniquement sur le niveau de présentation des stimuli. Ceci va dans le sens de McGregor et al. (1985) qui montrent que des voix inversées dans le temps (qu'ils considèrent comme non-familiales), sont plus mal perçues que lorsqu'elles sont présentées normalement. Malgré tout, plus les voix d'origine sont produites fortes plus la distance perçue est grande. Comme on peut le voir plus loin, nous rajouterons à cette étude, une hypothèse qui consiste à dire que les voix fortes sont mal perçues car la structure prosodique a été inversée.

## 2.4 Conclusion du chapitre 2

---

Dans ce chapitre nous avons discuté des différents phénomènes acoustiques et non-acoustiques permettant à un auditeur d'estimer la distance qui le sépare d'une source sonore (i.e. distance égocentrique).

Lorsqu'il s'agit de sources sonores quelconques, des variations d'intensité sont perçues comme des variations de distance. Plus l'intensité est faible, plus la distance est lointaine. Cette estimation de distance basée sur l'appréciation de l'intensité, n'est cependant possible que si l'auditeur a conscience d'une intensité de référence. L'intensité joue donc un rôle relatif dans la perception auditive de la

distance. L'absorption atmosphérique due à la propagation du son dans l'air peut également être considérée comme un facteur important dans la PAD. En effet, une variation du taux de hautes fréquences d'un signal, peut être interprétée comme une variation de distance de la source. Toutefois, il semble que ceci ne soit vrai que pour des valeurs élevées de filtrage des hautes fréquences qui ne correspondent pas à la réalité. Ce facteur est également considéré comme relatif pour la PAD. Il est primordial d'avoir une connaissance de la composition spectrale du signal sonore (i.e. son timbre) pour utiliser ce facteur. Enfin, le seul paramètre absolu dans la PAD, semble être le rapport d'énergie entre l'onde directe et les ondes réverbérées. En effet, un fort rapport est associé à des distances courtes et inversement. Toutefois, le rapport d'énergie ne permet d'estimer la distance d'une source que de façon très grossière. En effet, le seuil de perception lié au rapport d'énergie ramène à une précision de l'estimation de l'ordre du doublement de distance. Ainsi, il semble qu'une première estimation de distance grossière puisse être effectuée sur la base du rapport d'intensité et que la précision de l'estimation soit améliorée en utilisant l'intensité sur lequel nous sommes plus sensibles.

Dans le cas de la PAD du locuteur, où le signal de la source correspond à la voix, le facteur le plus important semble être le type de voix utilisé par le locuteur. En effet, la voix de ce dernier sera modifiée en fonction de la distance (voix chuchotée, parlée ou criée) qui le sépare de son interlocuteur, dans le but d'assurer une bonne situation de communication. Le type de voix utilisé par le locuteur a une influence considérable sur l'estimation de distance. En effet, plus la voix est forcée plus la distance perçue est lointaine. L'intensité perçue par l'auditeur de la voix semble être mis au second plan (mais contribue toutefois à la perception de la distance). Au regard de plusieurs études (Eriksson et Traunmüller, 2002; Fux et Zimpfer, 2009; Philbeck et Mershon, 2002; Traunmüller, 1997), l'intensité ne semble pas être un facteur primordial pour la perception de la distance d'un locuteur dans une situation réelle. Ceci s'explique par le fait qu'un locuteur ajuste son niveau de production vocale en fonction de la distance et de l'environnement. **Il en résulte que le niveau d'intensité sonore perçue au niveau de l'auditeur est quasiment constant** (Zahorik et Kelly, 2007). De plus Eriksson et Traunmüller (1999) et Wilkens et Bartel (1977) montrent qu'un auditeur est capable d'estimer ou de reconstruire l'intensité d'une voix indépendamment de l'intensité sonore qu'il perçoit. Un auditeur parvient donc à se représenter l'intensité de la voix indépendamment des variations d'intensité causées par la propagation de l'onde sonore. Dans le cas de la voix criée, l'absorption atmosphérique ne semble pas jouer un rôle important dans la PAD du locuteur car les phénomènes acoustiques liés ne sont pas perceptibles (dans notre étude).

Ainsi dans le but d'apporter une notion de distance d'un locuteur dans les systèmes de radiocommunication, la modification de voix semble être la solution la plus appropriée car pour la PAD, le type de voix semble prévaloir sur l'intensité ou l'absorption atmosphérique. En effet, un locuteur ajuste le niveau de sa voix de telle sorte que l'intensité sonore perçue par l'auditeur reste à

peu près constante indépendamment de la distance de communication. Or, cet ajustement de niveau de production de la parole relève d'une modification de l'effort vocal, qui à son tour modifie les différentes caractéristiques de la voix, permettant à l'auditeur d'estimer la distance. L'absorption atmosphérique des hautes fréquences n'étant *a priori* pas perceptible, le seul facteur acoustique encore exploitable est le rapport d'énergie et le type de voix utilisé. Malgré tout, nous n'aborderons dans ce mémoire que l'aspect lié à la transformation de la voix. En effet, le rapport d'énergie reste encore problématique car il dépend fortement de l'environnement. De plus l'utilisation d'un modèle de réverbération quelconque pouvant être incohérent avec les propriétés acoustiques de l'environnement pourrait fortement perturber les utilisateurs.

# La parole : production, représentation et variabilité

---

*« La parole n'a pas été donnée à l'homme : il l'a prise. »*  
— Louis Aragon, *Le Libertinage*.

Dans le chapitre précédent, nous avons pu voir que la nature de la parole, par opposition aux facteurs acoustiques liés à la propagation de l'onde sonore, est privilégiée dans la perception auditive de la distance d'un locuteur. Dans le but d'une meilleure compréhension des différents aspects que nous abordons dans les chapitres qui suivent, nous consacrons ce chapitre aux mécanismes de production de la parole ainsi qu'à sa représentation. Nous aborderons par la suite, la grande variabilité de la production vocale par le biais de la théorie de l'information, ainsi que les causes de cette variabilité. La variabilité de la parole étant principalement liée aux différents paramètres paralinguistiques, nous évoquons enfin deux descripteurs de la variabilité qui sont : la qualité de voix et la prosodie.

## 3.1 Anatomie et physiologie de l'appareil phonatoire

---

Le terme « appareil phonatoire » englobe tous les organes qui sont nécessaires à l'émission d'un son par l'Homme. Toutefois, comme le souligne le Dr ORL Peri Fontaa :

*« Cette terminologie peut donner l'impression qu'il s'agit d'un appareil particulier, dont la seule et unique fonction serait la production sonore. En fait, la phonation est apparue chez l'animal et en particulier chez l'homme comme une adaptation fonctionnelle secondaire, utilisant des structures qui en soi n'ont rien de particulièrement orienté vers une fonction phonatoire. L'appareil vocal n'existe pas en tant qu'organe: il n'existe que sous la forme d'une entité fonctionnelle.», Dr Perri Fontaa<sup>1</sup>*

---

<sup>1</sup> Source : [http://web.me.com/polysons/Site\\_6/Introduction\\_à\\_la\\_phonation.html](http://web.me.com/polysons/Site_6/Introduction_à_la_phonation.html)

En effet, l'appareil phonatoire est constitué de 3 parties, qui ont toutes une autre fonction que celles destinées à la production de la parole (cf. Figure 3-1):

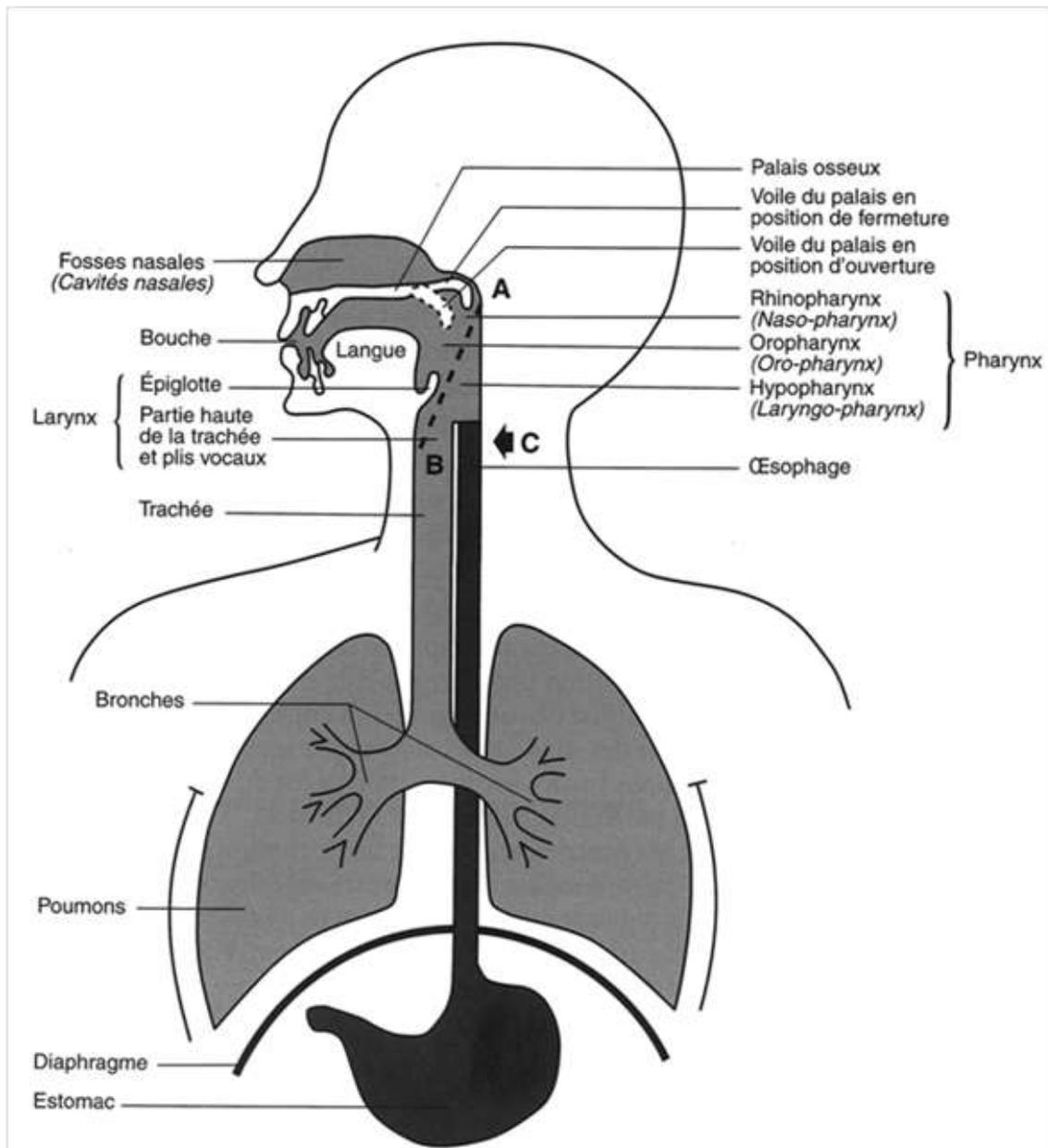


Figure 3-1 : Vue d'ensemble des organes de l'appareil phonatoire (d'après (Lehuche et Allali, 2010))

- (1) **L'appareil respiratoire** : qui a pour but premier de permettre la respiration. Dans l'acte phonatoire, cet appareil joue le rôle de soufflerie afin de fournir le flux d'air nécessaire à la vibration des cordes vocales.
- (2) **Le larynx** : qui a pour but premier de jouer un rôle dans la respiration et de protéger les voix respiratoires inférieures pendant la déglutition (notamment grâce à l'épiglotte

et aux cordes vocales). Pour la production sonore, le larynx joue le rôle de vibreur. Par le biais des cordes vocales, c'est le larynx qui produit le son de base (l'excitation).

- (3) **La sphère pharyngo-buco-nasales** : qui a pour but premier de permettre de mâcher (mâchoire, langue), de respirer tout en mangeant (cavités nasales), *etc...* Pour la phonation, ces cavités jouent le rôle de résonateur, et modulent le son de base produit par le larynx. L'ensemble du pavillon et des cavités est également appelé conduit vocal.

Ainsi, le son est produit par un ensemble d'organes de formes, de fonctionnements et de réactions complexes car il ne s'agit pas là de leur but premier.

### 3.1.1 L'appareil respiratoire

**L'appareil respiratoire ou soufflerie**, lors de la phonation, a pour rôle d'expulser l'air des poumons à travers le larynx puis le conduit vocal. Il se situe au niveau de la cage thoracique et est constitué, du diaphragme, des abdominaux (dans le cas de certaine production vocale) (Lehuche et Allali, 2010), des poumons et de la trachée.

### 3.1.2 Le larynx

Le rôle premier du larynx et notamment des cordes vocales est de protéger la trachée. Toutefois, dans la production vocale, le larynx joue un rôle secondaire et fait alors office de source sonore par mise en oscillation des cordes vocales (du moins pour les sons voisés). Le **larynx**, essentiellement constitué de tissus mous (muscles, cartilages, ligaments, muqueuses), joue alors le rôle du vibreur lors de la production vocale.

Il peut se décomposer en 3 espaces (cf. Figure 3-2):

- (1) **L'espace subglottique** : espace situé sous les cordes vocales qui se poursuit dans la trachée.
- (2) **L'espace glottique** : espace situé au niveau des cordes vocales.
- (3) **L'espace sus-glottique** : encore appelé étage vestibulaire, situé au dessus des cordes vocales. Il comprend deux bourrelets au dessus des cordes vocales appelées bandes ventriculaires ou fausses cordes vocales. Entre les cordes vocales et les bandes ventriculaires se situent deux cavités appelées ventricules de Morgagni.

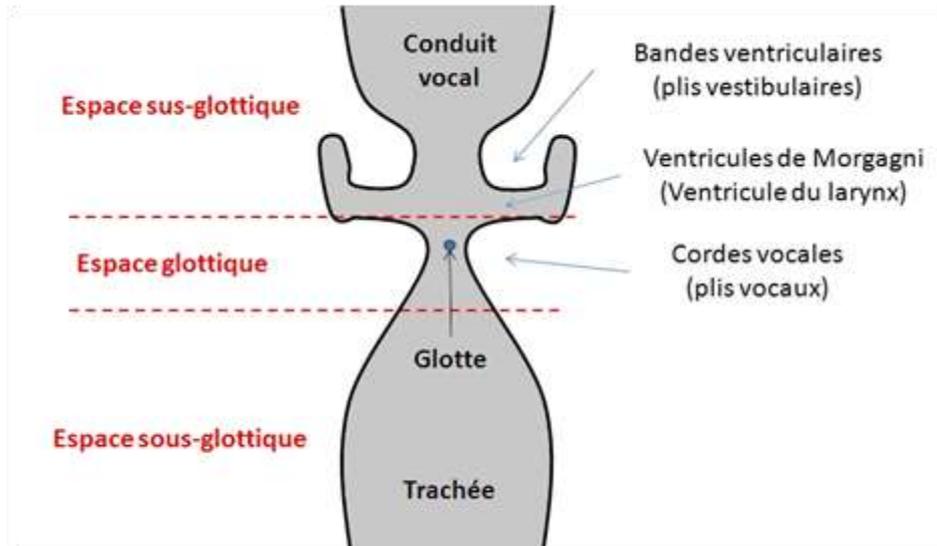


Figure 3-2 : Vue schématique antérieure du larynx

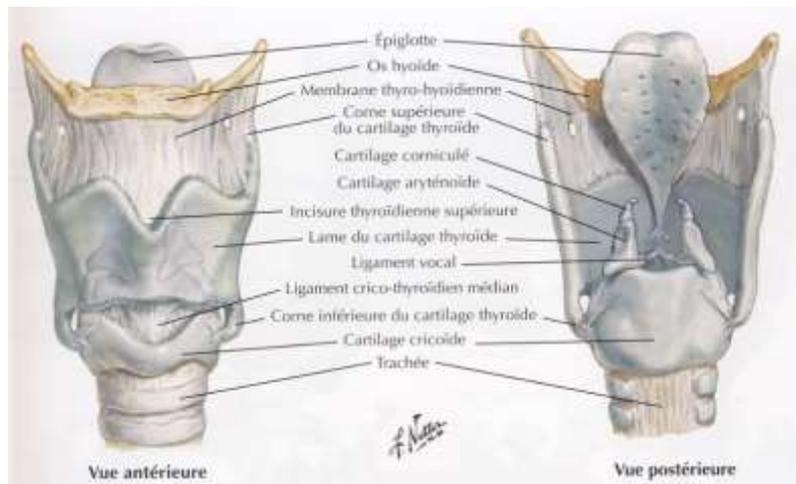


Figure 3-3 : Représentation du larynx (issu de (Netter, 2007))

Le larynx est constitué de 4 principaux cartilages : le cartilage thyroïde, le cartilage cricoïde, le cartilage aryténoïde, et l'épiglotte ainsi qu'un os : l'os hyoïde (cf. Figure 3-3). Ces cartilages et os sont reliés entre eux et au reste du corps humain par des muscles qui se classent en deux catégories :

- (1) **les muscles extrinsèques** (sous et sus-hyoïdes) : qui permettent une élévation ou un abaissement du larynx,
- (2) **les muscles intrinsèques** du larynx : qui permettent de modifier la configuration des cordes vocales aussi bien en longueur, en tension ou en épaisseur.

Il existe plusieurs explications concernant la physiologie du larynx, et notamment les mécanismes d'oscillation des cordes vocales (cf. (Lehuche et Allali, 2010) pour une liste exhaustive et un historique). Toutefois, la théorie aérodynamique myo-élastique (Van Den Berg, 1958) ou théorie myo-élastique complétée est la plus répandue. Cette théorie explique la vibration des cordes vocales de la manière suivante. Glotte fermée, l'appareil respiratoire expulse l'air des poumons à travers la trachée. Si la glotte est fermée, la pression subglottique augmente alors provoquant un déséquilibre des pressions entre l'espace subglottique et l'espace sus-glottique. Quand la pression subglottique atteint une valeur suffisamment élevée, les cordes vocales se décollent et la glotte s'ouvre faisant ainsi passer l'air dans le conduit vocal. A cet instant, les pressions sub- et sus-glottiques sont rééquilibrées. Le flux d'air traversant la glotte, engendre l'effet Bernoulli, qui explique un phénomène observé dans la physique des fluides : si le long d'un tube la section est diminuée en un endroit, le débit sera plus important à cet endroit. Si le débit est plus important, cela implique une pression plus faible à cet endroit. Ainsi, si le flux d'air est suffisant, la dépression causée par l'effet Bernoulli au niveau de la glotte, fait baisser la pression au niveau de celle-ci. Cette dépression tend alors à rapprocher les cordes vocales. Lors de ce rapprochement, les forces élastiques des cordes interviennent également pour refermer entièrement la glotte. Malgré le caractère auto-entretenu de la vibration des cordes vocales, des actions musculaires sont nécessaires pour débiter efficacement l'oscillation.

Pour la mise en vibration, les cordes vocales doivent être rapprochées. Cette adduction de la glotte est réalisée par **les muscles inter-aryténoïdiens** (transverses et obliques) qui font se rapprocher les cartilages aryténoïdes (cf. Figure 3-4(a)) et par les **muscles crico-aryténoïdiens latéraux** qui font pivoter les cartilages aryténoïdes vers l'intérieur (cf. Figure 3-4(b)). L'ouverture de la glotte est, quant à elle, assurée par **les muscles crico-aryténoïdiens postérieurs** réalisant l'action inverse des crico-aryténoïdiens latéraux, à savoir une rotation des aryténoïdes vers l'extérieur (cf. Figure 3-4(c)). Selon la théorie de Ganz, exposée dans (Lehuche et Allali, 2010), les muscles **crico-aryténoïdiens latéraux** peuvent également jouer un rôle d'abducteur dans le cas d'ouverture forcée de la glotte. La mise en vibration des cordes vocales nécessite également une tension sur ces dernières. Cette mise en tension est réalisée principalement par les muscles **thyro-aryténoïdiens** (aussi appelés *vocalis*) qui sont les muscles des cordes vocales. La contraction du *vocalis* engendre un rétrécissement de celles-ci et une augmentation de la raideur, pouvant engendrer des modes de vibrations différents (cf. Figure 3-4(d)). Dans le cas de certaines phonations comme pour les voix de têtes (où l'on souhaite atteindre des F0 élevée), une élongation des cordes vocales est nécessaire. Cette élongation est réalisée par les muscles **crico-thyroïdiens**. Leurs contractions entraînent le cartilage thyroïde vers le bas, engendrant ainsi une élongation des cordes vocales (aussi appelée bascule thyroïdienne). Les cordes vocales ayant une masse fixe, cette action engendre également un affinement et un raidissement de celles-ci, d'où une augmentation de la fréquence d'oscillations (cf. Figure 3-4(e)).

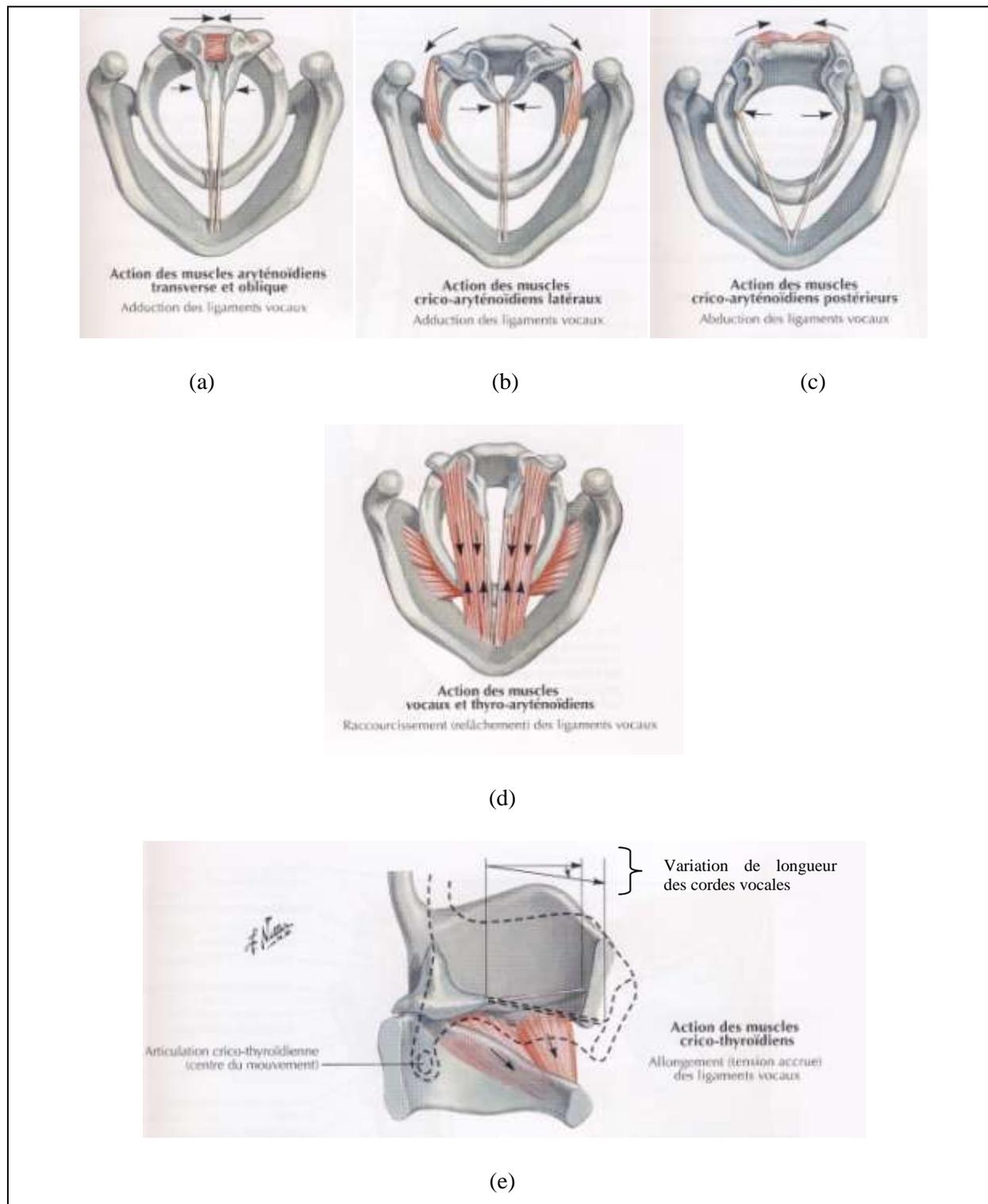


Figure 3-4: Physiologie du larynx (issu de(Netter, 2007))

### 3.1.2.1 Mécanismes glottiques

L'ensemble de ces contrôles laryngés permettent notamment de contrôler les mécanismes de vibration des cordes vocales. La production vocale non-pathologique se caractérise par 4 mécanismes (Henrich, 2006):

- (1) **Mécanisme 0** : aussi appelé voix *fry*, se caractérise par des cordes vocales courtes très épaisses et peu tendues. La vibration se caractérise par une durée d'ouverture très courte par rapport à la F0 plus faible. Ce mécanisme permet d'atteindre les fréquences fondamentales les plus basses.
- (2) **Mécanisme I** : Les cordes vocales sont épaisses et vibrent sur toute leur longueur. La masse vibrante est importante, ainsi que les amplitudes de vibration. Ce mécanisme est le mécanisme le plus utilisé dans la phonation par les hommes et utilisé par les femmes pour la réalisation de fréquence basse.
- (3) **Mécanisme II** : Dans ce cas, les cordes vocales sont fines et ne vibrent que sur les 2/3 de leur longueur. La phase ouverte est plus importante. Ce mécanisme est utilisé par les femmes pour la phonation et occasionnellement par les hommes.
- (4) **Mécanisme III** : aussi appelé voix de sifflet. Les cordes vocales sont très tendues, très fines, l'amplitude de vibration est faible et il n'y a quasiment pas de fermeture glottique. Ce mécanisme permet d'atteindre les fréquences les plus aiguës.

En fonction de la voix utilisée et donc du type de fréquence fondamentale désirée, ainsi qu'en fonction du sexe du locuteur, l'un de ces 4 mécanismes est employés. De plus dans le cas d'un glissando<sup>1</sup>, le locuteur passera d'un mécanisme à l'autre provoquant, une discontinuité perceptible dans la montée de la F0 pour les chanteurs non expérimentés.

### 3.1.3 Le conduit vocal

Créé par l'expulsion de l'air à travers le larynx faisant vibrer les cordes vocales, le flux d'air est alors propagé dans le conduit vocal. Ce dernier, ayant le rôle de résonateur, se compose d'organes mobiles qui constituent des cavités supra-laryngées au nombre de 3 (cf. Figure 3-1):

- (1) Le **pharynx** : La sphère pharyngale relie le larynx à la cavité buccale et nasale. Elle est constituée de 3 sous cavités : (a) l'hypo-pharynx ou laryngo-pharynx, située à l'arrière de l'épiglotte et allant jusqu'à l'œsophage, (b) l'oro-pharynx, allant de l'épiglotte jusqu'à la

---

<sup>1</sup> Un glissando ou glissato (du français « glisser ») est un terme musical générique d'origine italienne qui désigne soit un glissement continu d'une note à une autre, soit le passage d'une note à l'autre par un groupe de notes intermédiaires. (wikipédia)

cavité buccale et (c) le rhino-pharynx ou naso-pharynx, du voile du palais jusqu'aux fosses nasales.

(2) Les **fosses nasales** allant de l'extrémité supérieure du rhino-pharynx jusqu'aux narines.

(3) La **cavité buccale** allant de l'extrémité supérieure de l'oro-pharynx jusqu'aux lèvres.

## 3.2 Théorie acoustique de la production de la parole

---

Nous avons vu jusqu'ici les différents acteurs de la production vocale. A présent nous nous intéressons à la manière dont la parole est produite d'un point de vue acoustique.

La vibration des cordes vocales génère une onde qui se propage dans le larynx, le conduit buccale et le conduit nasale avant de se propager vers l'extérieur. Dans le cas de la production d'un son non voisé, ce sont les turbulences de l'air passant à travers la glotte qui génèrent cette onde. La théorie acoustique de la production de la parole développé par (Fant, 1960) explique comment le son est produit mais surtout comment il est modifié en fonction de la configuration du conduit vocal. Sur la base de mesures radiographiques du conduit vocal, la forme de ce dernier peut être relevée. Toutefois, la forme du conduit vocal est bien trop complexe pour réaliser des études théoriques de propagation acoustique. C'est pourquoi, en règle générale, on considère le conduit vocal comme une succession de tubes cylindriques de sections différentes. L'air de chaque cylindre correspond à l'air de la partie du conduit vocal qu'il modélise. Sur la base de cette représentation, et grâce aux connaissances sur la propagation de sondes dans un cylindre, il est alors possible de modéliser les différents effets du conduit vocal lors de la production. Toutefois, pour utiliser cette approche plusieurs considérations doivent être effectuées (Calliope (Firm), 1989; Mariani, 2002).

- Les parois des cylindres sont considérées comme rigides (pas de terme de vibration aux parois)
- La propagation est en ondes plane. Cette considération est vraie si la longueur d'onde est plus grande que la dimension transverse du cylindre. Soit pour des fréquences inférieures à 4 kHz
- Les pertes dans l'air (viscosité, conductivité thermique) sont négligées
- Hypothèse des petits mouvements, (on néglige les termes du second ordre)

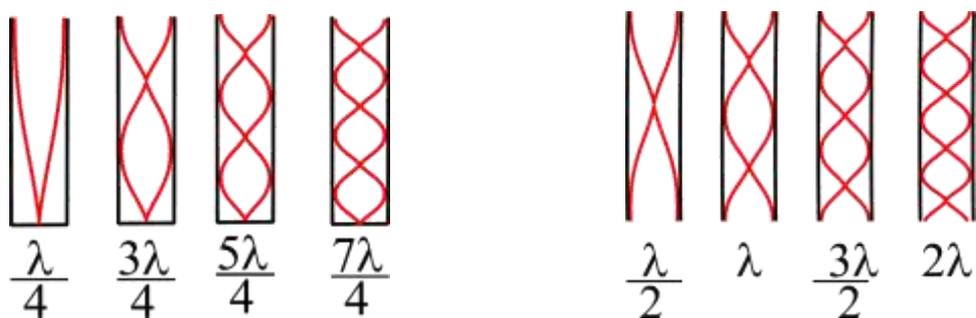


Figure 3-5 : Résonances acoustiques de cylindre pour des extrémités fermées (à gauche) et ouvertes (à droite)

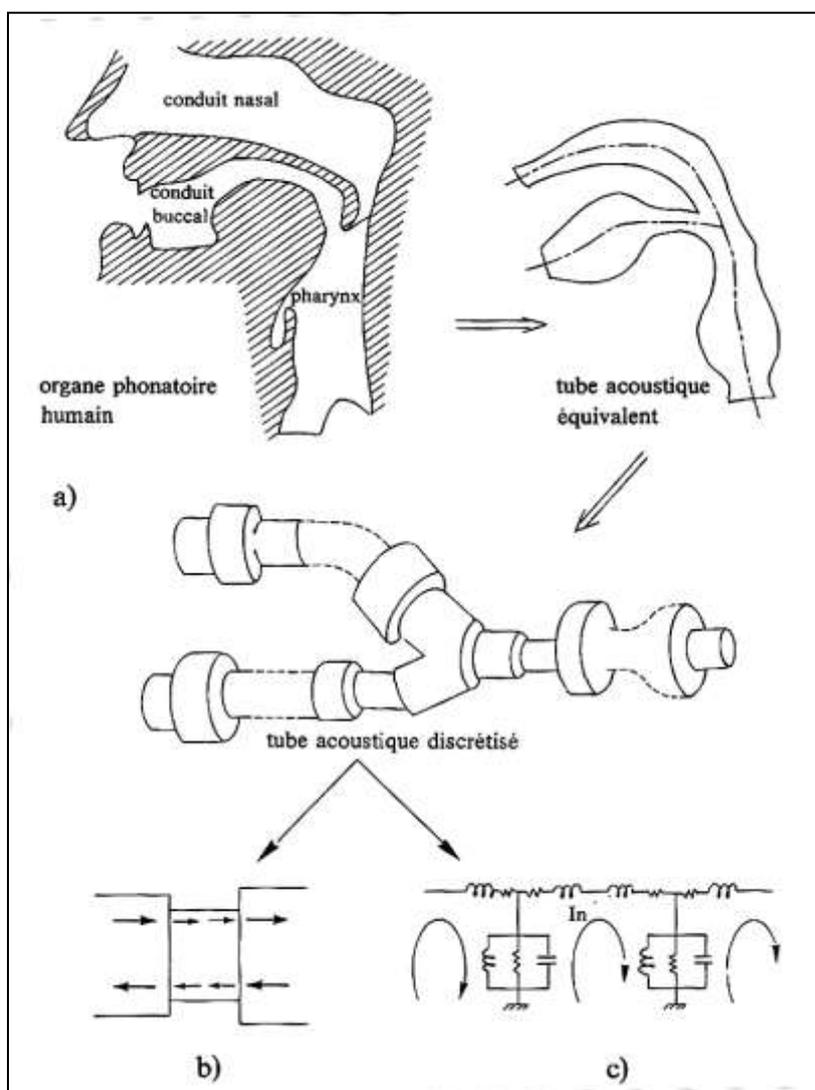


Figure 3-6: Modélisation du conduit vocal : a) conduit vocal, b) modèle acoustique de réflexion, c) Modélisation par lignes électriques à constantes localisées (issu de (Calliope (Firm), 1989))

A partir de ces considérations, on peut en déduire pour un seul cylindre les fréquences de résonance en fonction de la longueur du tube ainsi que des conditions aux limites (cf. Figure 3-5) :

$$f_x = \frac{(2n+1)c}{4l}, \text{ pour une extrémité fermée} \quad (\text{Eq. 3-1})$$

$$f_x = \frac{nc}{2l}, \text{ pour une extrémité ouverte}$$

où,  $c$  est la célérité du son dans l'air (env. 340 m/s),  $l$  la longueur du tube considéré, et  $n$  un entier représentant le mode de vibration. Remarquons que les conditions aux limites d'un tel système modifient drastiquement la position des résonances. Ceci signifie que dans une certaine mesure la source glottique influe sur la résonance du conduit vocal.

Il est toutefois nécessaire, dans le cas où deux cylindres sont mis bout à bout, de considérer également les phénomènes de réflexion entre ces deux tubes. Enfin, les parois du conduit vocal ne sont pas rigides et il est également nécessaire de considérer ce fait. A partir de l'ensemble de ces considérations et de ces simplifications, l'analogie entre la propagation acoustique d'une onde sonore dans une succession de tubes et un circuit électrique peut être effectuée. C'est d'ailleurs sur la base de cette analogie que la théorie acoustique de la production de la parole est fondée. La Figure 3-6 illustre la démarche que nous venons d'évoquer. A la lumière de cette théorie, que nous ne développons pas plus ici, la production de la parole en fonction des différents articulateurs peut être éclaircie.

Le conduit vocal se compose de plusieurs articulateurs permettant la production des voyelles et des consonnes. La production des voyelles s'effectue principalement à partir des mouvements de la langue (la position avant ou arrière), ainsi que par les mouvements de la mâchoire (la position basse ou haute) et des lèvres (l'arrondissement, la protrusion<sup>1</sup> et l'étirement). Le passage d'une voyelle orale à une autre en fonction de l'articulation peut se représenter comme sur la Figure 3-7. Le passage d'un /i/ à un /y/ s'effectue par arrondissement des lèvres. Le passage de /y/ à /u/ en avançant la langue. Notons que le larynx peut également être vu au sens d'un articulateur car sa position en hauteur est modifiable par le biais des muscles suspenseurs du larynx (i.e. muscles extrinsèques). Une position haute du larynx aura pour conséquence un conduit vocal plus court. Il s'agit cependant d'une contribution secondaire dans le sens où l'élévation du larynx n'est pas un articulateur permettant de produire des phonèmes mais permet de moduler le timbre et la richesse en harmonique de la parole (Bailly et al., 2008; Lindblom et Sundberg, 1971).

<sup>1</sup> La protrusion est une position avancée ou mouvement vers l'avant d'un organe.

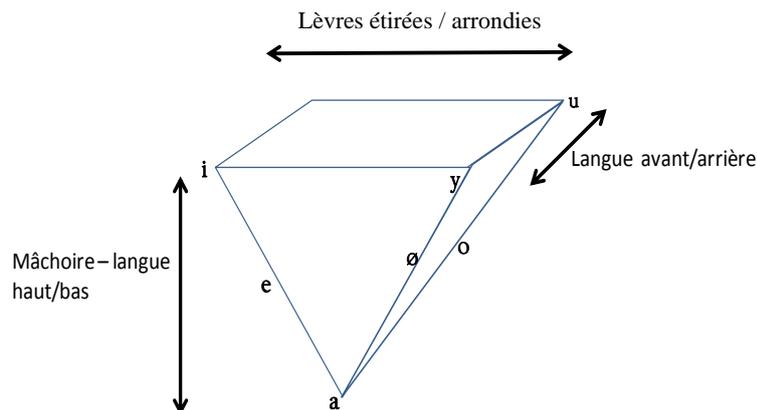


Figure 3-7: Représentation de l'espace vocalique en trois dimensions en fonction des articulations

L'ensemble de ces mouvements articulatoires ont pour conséquence la modification de la forme du conduit vocal et, de ce fait, de sa fonction de transfert. Ces modifications se traduisent (pour les voyelles notamment) par des pics d'énergie au niveau spectral correspondant aux fréquences de résonances du conduit vocal. Ces pics d'énergie sont généralement appelés formants.

La position de ces maxima spectraux dépend de la position des articulateurs. Ainsi, une bouche plus ouverte augmente la fréquence de la première résonance. La forme plus arrondie ainsi que la position avant ou arrière de la langue déplace la deuxième résonance. L'élévation du larynx quant à lui peut être considérée comme un décalage de toutes les fréquences de résonance vers les hautes fréquences. (cf. Figure 3-8)

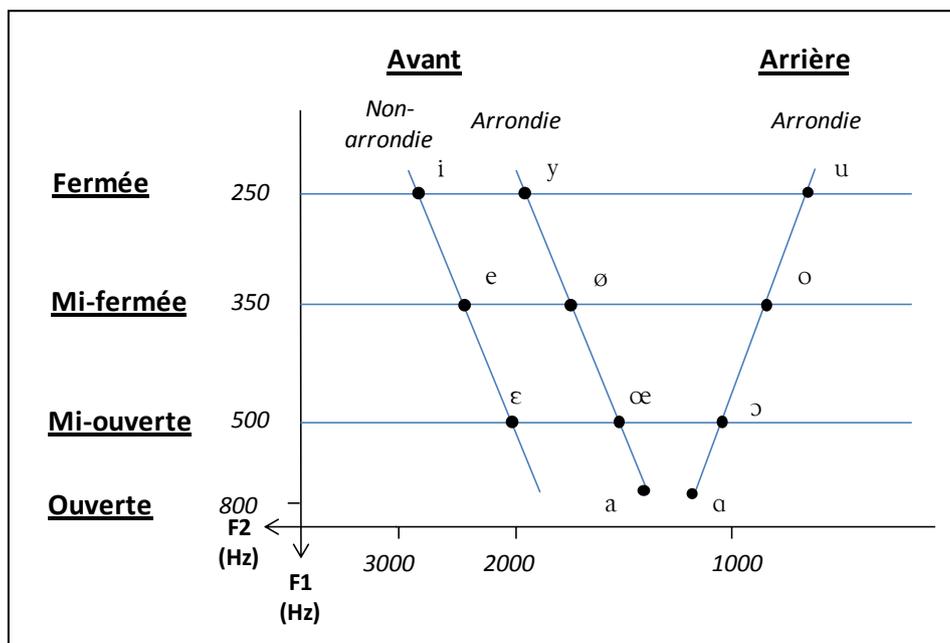


Figure 3-8 : Triangle vocalique française dans le plan F1-F2

La production des consonnes constitue un mécanisme de production plus complexe et dépend de la vibration ou non des cordes vocales (consonne voisée et non-voisée) ainsi que du lieu d'articulation. La production des consonnes fricatives est effectuée par constriction dentale (/s/,/ʃ/,/z/,/ʒ/) ou labio-dentales (/f/,/v/). Les consonnes occlusives (ou plosives) sont réalisées par des strictions bilabiales (/b/,/p/), des strictions entre la langue et le voile du palais (/t/,/d/) ou entre le dos de la langue et l'arrière du palais suivis d'une ouverture brusque (/k/,/g/).

Le mouvement du voile du palais permet de contrôler le couplage entre le conduit vocal et le conduit nasal, laissant passer plus ou moins d'air dans les cavités nasales. Les sons résultants de ce couplage sont les consonnes nasales (/m/,/n/,/ɲ/). L'abaissement du voile du palais permet également de produire les voyelles nasales (/ɛ̃/,/ã/,/õ/,/œ̃/).

La consonne liquide /l/ se réalise en posant la pointe de la langue sur les alvéoles. La consonne /R/ quant à elle se produit principalement par une constriction réalisée entre le dos de la langue et l'arrière du palais. Il existe de nombreuses variantes concernant la production du /R/ essentiellement basées sur des critères régionaux. Par exemple dans certaines régions cette consonne sera roulée.

Enfin la dernière classe de consonnes est appelée semi-consonne (ou semi-voyelle) (/j/, /ɥ/, /w/) Les semi-consonnes sont des fricatives sonores, mais leur articulation se situe au même endroit que certaines voyelles, d'où leur dénomination.

### 3.3 Modélisation acoustique de la parole

---

Bien que les mécanismes de production vocale soient bien connus aujourd'hui, il est encore difficile et lourd, en termes de calcul, de réaliser une modélisation aéromécanique en 3D de l'ensemble de l'appareil phonatoire. La voix se modélise le plus souvent *via* le signal de la parole suivant la théorie source-filtre.

#### 3.3.1 La théorie source-filtre

La théorie de la production de la parole exposée par Fant en 1960 (Fant, 1960) explique ceci :

Une onde est générée par la vibration des cordes vocales (source glottique). Cette onde se propage dans le larynx, le pharynx, les cavités buccales et nasales, dont l'ensemble forme le **conduit vocal**. Le conduit vocal fait office de résonateur. Certaines composantes fréquentielles de l'onde acoustique qui s'y propage sont alors accentuées (formants). L'onde résultante est ensuite rayonnée vers l'extérieur, par les lèvres et les narines (**rayonnement aux lèvres**). Dans le cas des sons dits non-voisés, appelés

également consonnes sourdes (/p/, /t/, /k/, /f/, /s/, /ʃ/), les cordes vocales ne vibrent pas. La source glottique est alors générée par les turbulences de l'air expulsé (**source aperiodique**). La Figure 3-9 résume le principe de la théorie source filtre et associe à chaque composante son équivalent morphologique.

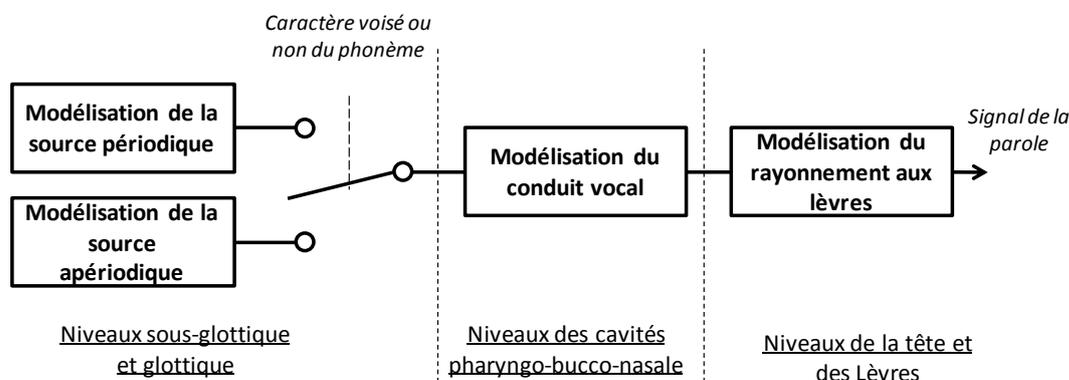


Figure 3-9: Représentation conceptuelle du modèle source-filtre

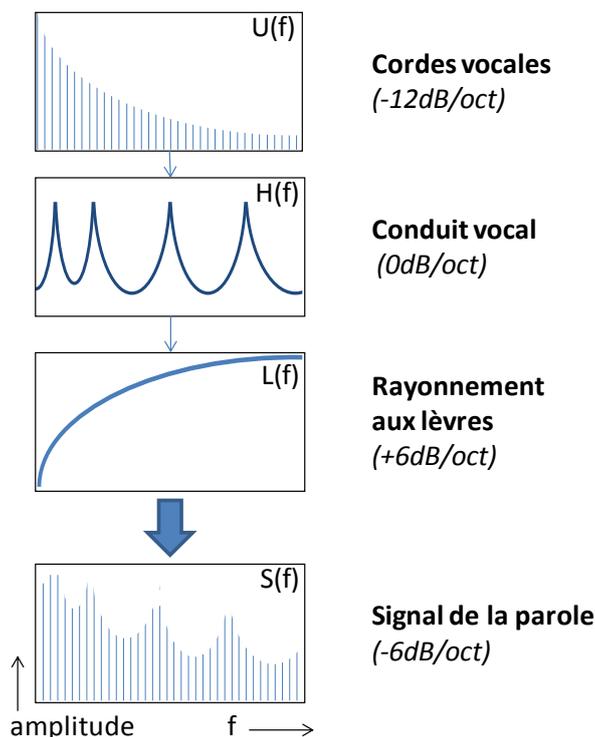


Figure 3-10: Représentation fréquentielle des signaux aux différents niveaux du modèle source-filtre

Le signal de la parole  $s(t)$  sera alors décrit comme la convolution de l'onde de source, notée  $u(t)$  (périodique ou non) par la réponse impulsionnelle du conduit vocal, notée  $h(t)$  et par une fonction de rayonnement, notée  $l(t)$ . (cf. Figure 3-10)

Dans le domaine temporel, la parole peut alors être représentée par l'équation suivante :

$$s(t) = u(t) * h(t) * l(t) \quad (\text{Eq. 3-2})$$

Ou, dans le domaine fréquentielle, comme représenté sur la Figure 3-10 :

$$S(f) = U(f).H(f).L(f) \quad (\text{Eq. 3-3})$$

### 3.3.1.1 Limitation de la représentation source-filtre

Cette théorie se fonde sur une hypothèse d'indépendance entre la source, le filtre et le rayonnement aux lèvres. Cependant, il a été démontré qu'il existe des interactions entre la source et le filtre (Flanagan, 1968). Il existe plusieurs types d'interactions qui sont recensés notamment dans (Childers et Wong, 1994). Ces auteurs recensent notamment une interaction due à l'impédance du conduit vocal (appelée *vocal load*), qui a pour conséquence d'agir directement sur la forme de l'onde de débit glottique et qui a notamment pour effet, de rendre l'ODG asymétrique. Notons que l'impédance du conduit vocal peut être modélisée via le coefficient d'asymétrie du modèle de Liljencrants et Fant (le modèle LF) (Fant et al., 1985). La deuxième source d'interaction forte est directement liée à la position du premier formant (appelée *Ripple*). Celle-ci semble être due à une interaction entre la glotte et le premier formant du conduit vocal, qui engendre une ondulation supplémentaire sur l'onde de débit glottique durant sa phase ouverte. Elle est causée par une relation non-linéaire entre le débit glottique et la pression transglottique (Båvegård et Fant, 1994). Il existe également une autre forte interaction généralement appelée *formant tuning*. En dépit de la théorie source-filtre qui suppose une indépendance entre la F0 et les fréquences des formants, plusieurs auteurs démontrent toutefois que ce n'est pas toujours le cas ; notamment pour les voix chantées (Joliveau et al., 2004a, 2004b; Sundberg, 1977). Le fait qu'une harmonique de la F0 soit centrée sur la première résonance du conduit vocal, favoriserait la vibration des cordes vocales. De plus, cet ajustement permettrait d'obtenir un maximum d'efficacité de la voix en plaçant un des pics d'énergie issue de la glotte à l'aplomb d'une résonance du conduit vocal.

Beaucoup d'études se consacrent aujourd'hui au problème de la modélisation et de la compréhension des phénomènes d'interactions. Voir (Childers et Wong, 1994; Titze, 2008) pour plus de détails et un état de l'art relativement exhaustif. Malgré tous, le modèle source filtre reste encore aujourd'hui le modèle le plus utilisé en traitement et analyse de la parole. En effet précisons que s'il existe effectivement des phénomènes d'interactions, ceux-ci ne sont pas systématiquement présents, ou tout du moins, ont des effets très faibles. Dans certains cas en revanche, comme pour la voix féminine ou pour les voix d'enfants, il semblerait que l'ampleur des interactions soit plus forte (Titze, 2008).

### 3.3.2 Modélisation acoustique du conduit vocal

Le conduit vocal a une forme complexe possédant des cavités résonantes, qui ont pour rôle la modification de l'onde de source. Une première modélisation du conduit vocal peut être entreprise à base d'un modèle multi-tuyaux. Dans ce cas, le conduit vocal est vu comme étant une succession de tubes, de sections et de longueurs différentes (modélisation multi-tuyaux).

Dans l'analyse-synthèse de la parole, la modélisation du conduit vocal est en général effectuée par un filtre numérique (qui dans une certaine mesure peut être vu comme un système multi-tuyaux). Plusieurs solutions sont possibles dans le but de modéliser la réponse en fréquence du conduit vocal. La solution la plus répandue est la modélisation du conduit vocal par un filtre AR (*Auto Regressive*), soit un filtre tout-pôle. Une variante est la modélisation par filtre ARMA (*Auto Regressive Moving-Average*), soit un filtre introduisant également des zéros (des antirésonances) dans la fonction de transfert du conduit vocal. La modélisation ARMA a longtemps été considérée et l'est encore, dans une certaine mesure, comme la solution permettant de modéliser les sons nasaux. En effet, la nasalité introduit une chute d'énergie conséquente dans le spectre de la parole, souvent associée à la présence d'une antirésonance (Titze, 1994). Toutefois, le débat qui existe à ce jour concernant le bien fondé de l'utilisation des filtres ARMA pour les nasalités, sort du cadre de ce travail et c'est pourquoi nous n'entrons pas plus dans les détails (pour plus de détails cf. (Feng, 1986)).

### 3.3.3 Modélisation de la source périodique : la vibration des cordes vocales

D'après la théorie source-filtre, la source périodique est un signal harmonique ayant une décroissance spectrale de -12 dB/octave. Il existe plusieurs méthodes permettant de modéliser le signal correspondant à la vibration des cordes vocales.

La première et la plus simple consiste en une approximation grossière du spectre de la source glottique. Celle-ci consiste à modéliser le signal issu de la vibration des cordes vocales par un **peigne de Dirac**. Le signal  $u(t)$  est alors une suite d'impulsions de Dirac espacées de  $T_0=1/F_0$ . Notons que le spectre de ce signal est un spectre harmonique de période  $F_0$ , qui ne possède pas une décroissance de -12 dB/octave. Cette modélisation possède un inconvénient majeur. Le signal glottique est purement harmonique et génère des voix de synthèses métalliques.

D'autres méthodes offrant de meilleures qualités de synthèse consistent à utiliser des modèles de source glottique, ou plus précisément **de l'onde de débit glottique** (pour un état de l'art cf. (Cummings, 1995)) D'une façon générale, ces modèles se classent en trois catégories :

- (1) Les modèles paramétriques sans interactions : il s'agit de modèles destinés à décrire la forme du signal de l'onde de débit glottique. Ces modèles sont dits sans interactions car ils ne

tiennent pas compte de l'interaction source filtre (cf. §3.3.1.1). Le modèle le plus répandu en analyse-synthèse de la parole est sans doute le modèle LF (Fant et al., 1985). D'autres modèles fréquemment utilisés existent également, comme le modèle KLGLOTT88 basé sur le modèle de Rosenberg (modèle R) (Rosenberg, 1971) et utilisé dans le synthétiseur de Klatt (Klatt, 1980). On trouve également des variantes comme le modèle R++ proposé par (Veldhuis, 1998). Ce modèle est basé sur le modèle de Rosenberg d'où son nom Rosenberg++. Ce modèle propose une alternative au modèle courant LF avec un coût de calcul plus faible.

- (2) Les modèles mécaniques paramétriques avec interactions : il s'agit de modèles mécaniques permettant de simuler, par une structure simple, les oscillations des cordes vocales (modèle d'oscillateur mécanique masse-ressort-amortissement). Ces modèles offrent l'avantage d'inclure l'interaction entre la source et le filtre car ils tiennent compte de l'impédance du conduit vocal pour le calcul des oscillations des cordes vocales. Différents modèles existent et notamment les modèles dits à 1 masse (Flanagan et Landgraf, 1968), à 2 masses (Ishizaka et Flanagan, 1972), voir à 3 masses (Story et Titze, 1995). Ces modèles sont toutefois très lourds en calcul.
- (3) Les modèles physiologiques : ces modèles utilisent une représentation réelle des cordes vocales via une modélisation à 2 ou 3 dimensions. Les calculs de vibration des cordes vocales sont, par la suite, effectués par de méthodes très lourdes type MEF (Modélisation par éléments finis). Ils ont toutefois l'avantage d'être très réalistes. Le modèle le plus connu et le plus élaboré est peut être celui de Titze et Riede (décrit dans (Titze et Riede, 2010))

Le modèle LF étant le plus répandu, nous ne décrivons que ce dernier.

### 3.3.3.1 Modèle d'onde de débit glottique de Liljencrants-Fant (LF)

Nous décrivons ainsi brièvement le modèle de source LF, proposé par Liljencrants, Fant et Lin (Fant et al., 1985), étant donné que la majorité des études liées aux variations de la source glottique se base sur les paramètres de celui-ci.

Le modèle LF se compose de 5 paramètres de formes (cf. Figure 3-11):

- $\alpha_m$ , le coefficient d'asymétrie
- $O_q$ , le quotient ouvert
- $Q_a$ , le quotient de fermeture
- $E$  : amplitude maximum de la dérivée de l'onde de débit glottique
- $T_0$ , la période.

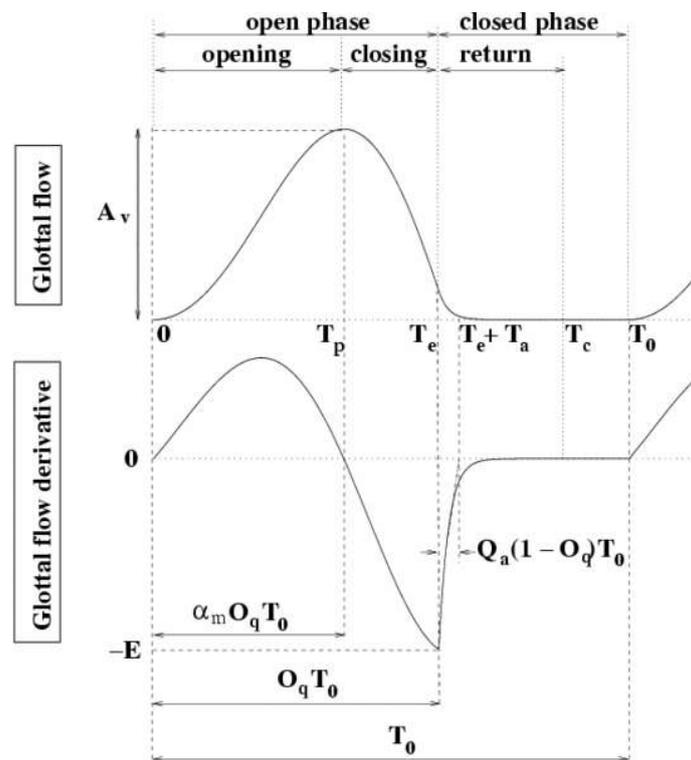


Figure 3-11: Représentation temporelle du modèle LF. En haut le signal temporel, en bas sa dérivée

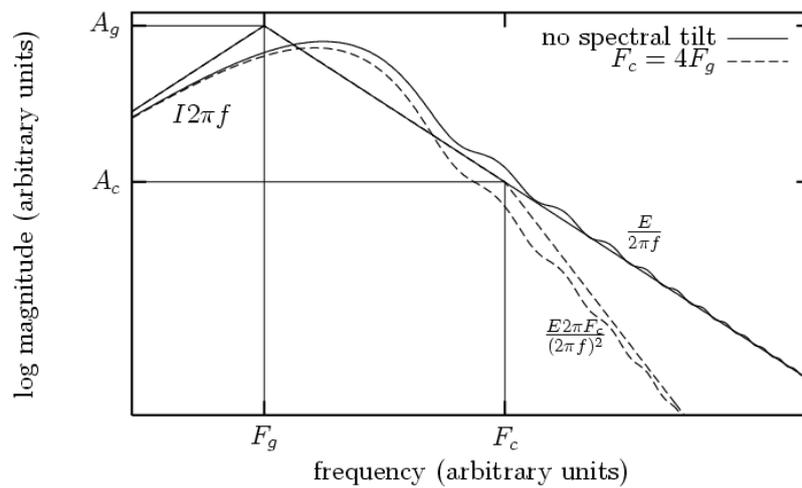


Figure 3-12: Spectre de la dérivée du modèle LF.<sup>1</sup>

<sup>1</sup> Source : [http://rs2007.limsi.fr/index.php/The\\_spectrum\\_of\\_glottal\\_flow\\_models](http://rs2007.limsi.fr/index.php/The_spectrum_of_glottal_flow_models)

L'expression mathématique de la source glottique s'exprime ainsi :

$$u(t) = \begin{cases} E_0 e^{at} \sin(\omega_g t) & \text{pour } 0 < t < T_e \\ -\frac{E}{\varepsilon T_a} [e^{-\varepsilon(t-T_e)} - e^{-\varepsilon(T_c-T_e)}] & \text{pour } T_e < t < T_c \\ 0 & \text{pour } T_c < t < T_0 \end{cases} \quad (\text{Eq. 3-4})$$

$\omega_g$  est la pulsation glottique défini par  $\pi/T_p$ .  $E_0$  est l'amplitude positive et  $E$  l'amplitude négative de la dérivée.  $a$  et  $\varepsilon$  sont des paramètres d'ajustement. Le paramètre  $a$  se détermine en considérant la continuité entre la première et la deuxième partie du modèle LF à  $t=T_e$ , et le paramètre  $\varepsilon$  est obtenu en respectant  $\int_0^{T_0} u(t)dt = 0$ . En règle générale  $T_c=T_0=1/F_0$ .

Les différents paramètres du modèle LF, ont une influence directe sur le spectre de la source (cf. Figure 3-12). D'une façon générale, si  $E$  est fixe, le formant glottique est influencé par  $Oq$ , l'amplitude du spectre par  $\alpha m$ , la pente spectrale est influencée par  $Qa$  et notamment au-delà de  $F_c$  ou une pente de -6 dB/oct vient s'ajouter (Doval et al., 2006).

### 3.3.4 Modélisation de la source apériodique : turbulence causée par l'air

Il existe différentes sources de turbulence dans la parole. Le bruit de turbulence peut être dû à la constriction au niveau buccal (comme pour les fricatives), au niveau du conduit vocal ou encore au niveau glottique. On estime la source de bruit buccal comme ayant un spectre plat de 0 à 4 kHz, puis un spectre qui décroît avec la fréquence (Mariani, 2002). Les bruits de constriction glottique (ou bruits d'aspiration), quant à eux, sont considérés comme des bruits large bande ayant un maximum situé aux alentours de 2 kHz (Mariani, 2002). Les bruits de constriction du conduit vocal (ou bruit de « frications ») sont également des bruits à large bande avec un maximum aux alentours de 4-9 kHz (Mariani, 2002).

Ainsi, la source de bruit peut être considérée comme un bruit blanc qui sera filtré par un passe-bas ou un passe-bande à large bande. La modélisation se réalise toutefois en considérant un bruit blanc. En effet, le filtrage induit par la place de la constriction, sera le plus souvent modélisé via le même procédé que pour les voyelles (filtre AR) modélisées au sein du filtre du conduit vocal. On trouve toutefois dans la littérature, des méthodes d'analyse-synthèse utilisant une modélisation du bruit plus poussée. Parmi celle-ci nous pouvons citer les synthétiseurs CELP (*Codebook excitation, linear prediction*), utilisant des modèles types de bruits, ou encore des synthétiseurs qui séparent la composante périodique de la composante apériodique du signal comme pour les synthèses HNM (*Harmonic plus noise model*) par exemple.

### 3.3.5 Modélisation du rayonnement aux lèvres

La rayonnement aux lèvres est en règle générale modélisé par un piston vibrant dans une sphère modélisant la tête (Flanagan, 1972). Pour des raisons de simplification, le rayon du piston (i.e. la bouche) est considéré comme petit devant le rayon de la sphère (i.e. la tête). De ce constat, le rayonnement aux lèvres peut être modélisé comme un piston vibrant dans un plan infini. D'après (Pierce, 1981) l'impédance  $Z_m(\omega)$  d'un piston rigide monté dans un plan infini s'exprime par :

$$Z_m(\omega) = \frac{-j\omega\rho}{2\pi} \int \int \int \int R^{-1} e^{ikR} dx_s dy_s dx dy \quad (\text{Eq. 3-5})$$

où  $R = \sqrt{(x - x_s)^2 + (y - y_s)^2}$  est la distance entre la surface du piston  $(x_s, y_s)$  et la surface considérée  $(x, y)$ .  $\omega$  est la pulsation du signal,  $\rho$  la densité volumique de l'air, et  $k$  le nombre d'onde défini par  $k = \omega/c$ , avec  $c$  la célérité du son. Plusieurs simplifications peuvent être apportées à ce modèle de propagation très général. La première simplification déjà évoquée est la considération de la surface de la bouche comme étant très petite. De plus la voix est enregistrée via un microphone qui ne mesure la pression acoustique qu'en un point précis. Ainsi, les surfaces  $(x_s, y_s)$  et  $(x, y)$  peuvent être considérées comme ponctuelles. Les deux intégrales de surface sont alors supprimées et  $Z_m$  peut être exprimé par :

$$Z_m(\omega) = \frac{j\omega\rho x}{2\pi} e^{-ikx} \quad (\text{Eq. 3-6})$$

où  $x$  représente la distance entre la bouche et le microphone.

Les enregistrements étant effectués à une position fixe, la distance bouche micro est donc fixe. D'autre part, dans des cas comme l'analyse du signal de la parole uniquement (sans appareil de mesure annexe), le facteur  $e^{-ikx}$ , exprimant le retard de l'onde rayonnée (i.e. mesurée par le microphone) par rapport à l'onde issue de la bouche n'est pas utile.

L'équation (Eq. 3-5) devient alors :

$$Z_m(\omega) = C \cdot j\omega \quad (\text{Eq. 3-7})$$

où  $C$  est une constante représentant un facteur d'atténuation de la pression, en fonction de la distance entre le microphone et la bouche.

En négligeant cette atténuation (car constante), on obtient une approximation du modèle du rayonnement aux lèvres numériques  $L(z)$  qui correspond à un différenciateur numérique soit :

$L(z) = 1 - z^{-1}$ . Il est important de ne pas confondre cette fonction de rayonnement avec les filtres de préaccentuation.

### 3.3.5.1 Rayonnement aux lèvres et filtre de préaccentuation

En traitement du signal de la parole, des filtres de préaccentuation sont utilisés dans le but de réaliser une meilleure approximation des filtres du conduit vocal. En effet, le spectre du signal de la parole a théoriquement une pente de -6 dB/octave. Cette pente peut engendrer des erreurs d'estimation des filtres notamment par les méthodes de prédiction linéaire. En effet, les premiers formants ont des énergies plus fortes que les formants en haute fréquence. Ainsi, lors de la prédiction linéaire, les formants en basse fréquence seront modélisés préférentiellement, en dépit de la modélisation en haute fréquence. Cela implique des erreurs potentielles sur l'estimation des formants situés à des fréquences élevées. Afin de réduire ces erreurs potentielles d'estimation du filtre du conduit vocal, le spectre du signal de la parole doit être redressé. Ainsi, l'estimation de ces filtres se fait après une préaccentuation spectrale du signal de la parole. De façon générale, ce filtre de préaccentuation est défini comme un filtre passe haut du premier ordre ( $V(z) = 1 - a_v \cdot z^{-1}$ ). Le facteur  $a_v$  est choisi proche de 1 afin que la réponse en fréquence de ce filtre soit similaire à une pente croissante de +6 dB/octave. Historiquement on utilise un facteur égal à 0,95.

Toutefois, lors d'une synthèse vocale, le son résultant du filtrage de l'onde débit glottique par la fonction de transfert du conduit vocal, est filtré par la modélisation du rayonnement aux lèvres et non pas par le filtre de préaccentuation (cf. Figure 3-13). Il est fréquent d'observer de telles confusions, à savoir que lors d'une synthèse pure, le rayonnement est modélisé par un filtre de désaccentuation. Il s'agit là d'un point théorique important sur le rôle de chacun de ces filtres. Malgré tout, d'un point de vue applicatif, les conséquences ne sont pas dramatiques étant donné la grande similarité du modèle de rayonnement et du filtre de préaccentuation.

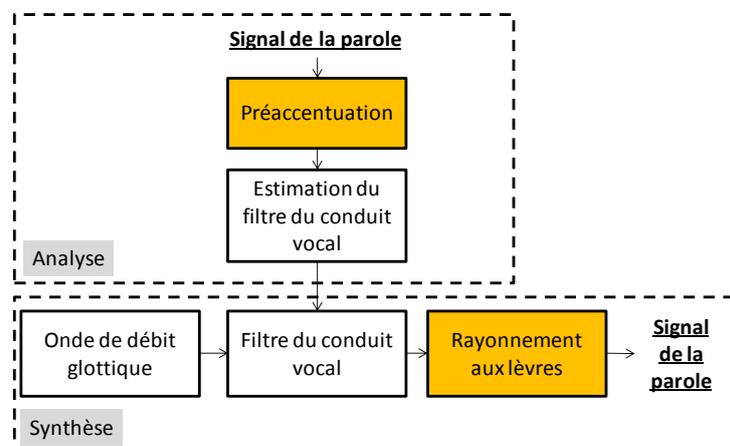


Figure 3-13: Représentation de l'utilisation du filtre de préaccentuation et du modèle du rayonnement aux lèvres en fonction du contexte : analyse ou synthèse

---

## 3.4 La parole : information et variabilité

---

Nous avons vu, dans la première section de ce chapitre, que la production vocale est liée à la physiologie des organes qui constitue l'appareil phonatoire. En considérant le cas de la parole, chaque son produit a, avant tout, un but linguistique : dans le but de reproduire chaque phonème d'un mot que l'on souhaite communiquer. Toutefois il est important de préciser que la production vocale dépend également d'autres facteurs, tels que l'état psychologique de la personne, le but de la production vocale ou encore l'environnement dans lequel est produit la voix.

Ces différents facteurs induisent des variations des paramètres de la parole autour de cette structure phonétique. L'information linguistique de la parole sera toujours présente. Ces variations, trahissent pour la plupart l'état psychologique du locuteur (i.e. l'émotion). En effet, la voix de ce dernier est différente si le locuteur, au moment de la phonation est triste, gai, effrayé, etc... Le contexte dans lequel se déroule la communication influe également sur la production vocale, comme dans le cas d'une communication en environnement bruyant ou plus proche de notre étude, lors d'une communication à distance. L'ensemble de ces variations de la parole n'apporte aucune information supplémentaire d'un point de vue linguistique. En revanche ces variations informent l'auditeur sur la situation ou l'état du locuteur. Par opposition aux paramètres linguistiques, l'ensemble de ces variations de la parole se regroupe sous les termes paralinguistiques.

### 3.4.1 Théorie de l'information

La parole est constituée de deux types d'informations : linguistique et paralinguistique. La théorie de l'information (cf. (Feng, 1996; Liénard, 1977)) offre une approche physique permettant de juger de la part d'informations linguistiques et prosodiques dans le signal. Cette théorie a été développée par les ingénieurs en télécommunication dans les années 50, dans le but d'étudier la quantité d'informations à transporter. Par le biais des caractéristiques d'un canal de transmission, elle permet de quantifier mathématiquement l'information.

On estime que pour transmettre un signal de parole en haute fidélité, 300 000 bits/seconde sont suffisants (bande passante 20 Hz-15000 Hz). Le signal de la parole peut toutefois être transmis en réduisant la bande passante tout en restant intelligible. Par exemple, les lignes téléphoniques possèdent une fréquence d'échantillonnage d'environ 8 kHz et une bande passante d'environ 3 kHz (300 Hz à 3400 Hz). Le signal est codé sur 8 bits. Ainsi, un tel signal se transmet avec un débit de 64 000 bits/seconde. Ceci permet de préserver une bonne intelligibilité de la voix mais dégrade légèrement

l'identité du locuteur. Il s'agit là du débit minimum nécessaire pour transmettre directement le signal de la parole.

Toutefois, le signal de la parole est très redondant. L'élimination de ces redondances correspond à une modélisation du signal de la parole ou plus précisément, au codage de la parole, par le biais de la modélisation. Ce codage peut être effectué en utilisant un minimum de 10 paramètres (soit en général 8 coefficients LP (*Linear Prediction*), 1 valeur de  $F_0$ , une valeur d'intensité). Ces 10 paramètres sont codés sur 100 niveaux (env. 7 bits) à une cadence de 100 fois/seconde (toute les 10 ms). Ainsi, la quantité d'informations dans un tel cas, s'exprime par  $C=100*10\log_2(100) \approx 7000$  bits/seconde. Dans une telle transmission, le signal de la parole est donc, au préalable codé via une technique de type *Vocoder* (*voice coder/decoder*), et transmis sur le canal. Au niveau du récepteur, les paramètres sont utilisés dans un synthétiseur pour reconstruire le signal de la parole.

Considérons à présent le cas extrême ou nous désirons transmettre le signal de la parole, avec un débit le plus faible possible. Dans un tel cas, il est nécessaire de garder l'information la plus importante du signal de la parole : l'information linguistique. Elle correspond à l'information qui doit permettre d'identifier un mot. Ainsi on peut quantifier la quantité d'information minimum pour transmettre un message linguistique. Pour simplifier le calcul de la quantité d'informations linguistiques, on suppose une cadence de production maximale d'environ 5 syllabes/seconde. En français, environ 320 syllabes de base (i.e. les plus fréquentes) de type consonne-voyelle, sont nécessaires pour noter phonétiquement la parole. Ainsi, la quantité d'information pour transmettre un tel message, s'exprime par  $C=5*\log_2(320) \approx 50$  bits/seconde. Ce type de codage permet alors de transmettre de la parole sous une forme phonétique. Le codage s'effectue alors par reconnaissance de la parole suivi d'une transcription en symboles phonétiques. Ceux-ci sont ensuite envoyés sur le canal de transmission pour, au niveau du récepteur, être synthétisés à partir de synthétiseur texte-to-speech.

Ainsi, selon la théorie de l'information, en comparant le débit nécessaire à la transmission linguistique (50 bits/seconde) et celui nécessaire pour une transmission complète de la parole (7000 bits/seconde), la parole peut être considérée comme comprenant environ 1% d'information linguistique. Le reste étant principalement des redondances d'informations ainsi que les informations paralinguistiques. La part d'informations paralinguistiques est donc très élevée par rapport à l'information linguistique présente dans le signal de la parole. La grande variabilité de la parole provient de ce rapport entre informations linguistiques et paralinguistiques.

Penchons nous à présent sur les différents facteurs influençant les paramètres paralinguistiques dans le cas d'une communication parlée. Il existe évidemment d'autres paramètres que ceux mentionnés par la suite. Toutefois, dans le cadre de notre étude, nous ne développons que ceux liés directement à la communication parlée.

### 3.4.2 Conditions adverses : cas de la communication parlée

Liénard (1977), résume la communication parlée comme étant régie par des contextes et des rétroactions (cf. Figure 3-14). Dans une communication parlée l'individu A qui communique avec B désire transmettre de l'information verbale mais également de l'information extra verbale comme les gestes, les mimiques ou les postures. L'individu B doit alors décoder les informations fournies par A en fonction de l'environnement spatio-temporel qui leur est commun, du contexte du message ainsi que du contexte individuel. Lors de la communication, il existe une rétroaction permanente de B vers A exprimée de façon verbale ou extra-verbale qui peut influencer la façon dont A transmet son message.

Ainsi, la production de l'individu A est modifiée en fonction de l'environnement dans lequel il s'exprime, du contexte de la situation ainsi que du contexte du message. D'autre part, sa production vocale est également influencée par les retours effectués par son interlocuteur, soit par des informations verbales soit par des extra-verbales. Pour une information linguistique donnée, l'ensemble de ces variations constitue une partie des paramètres paralinguistiques.

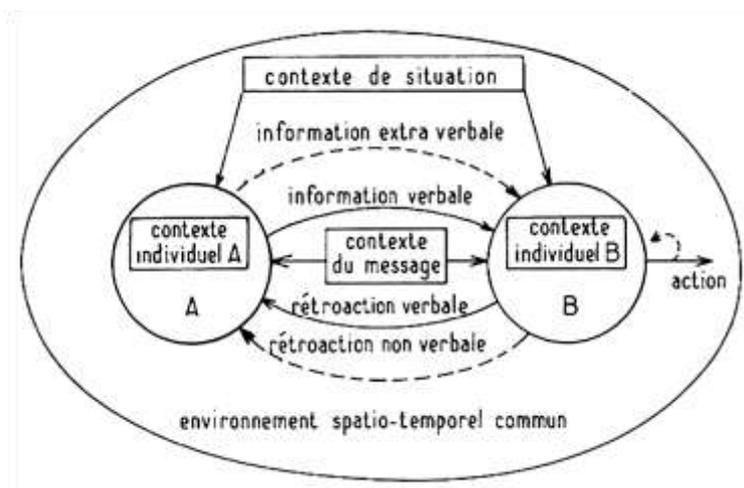


Figure 3-14: Schématisation de la communication parlée (d'après (Liénard, 1977))

### 3.4.3 Comportement et adaptation

Il existe plusieurs facteurs qui influent sur la production vocale d'un locuteur dans une situation de communication parlée. Dans un premier temps, nous évoquons l'influence du locuteur lui-même.

**La voix d'un locuteur** est construite à partir d'unités phonétiques qui forment le message qu'il souhaite transmettre (i.e. information linguistique). Toutefois, la voix véhicule également d'autres informations qui peuvent refléter l'état psychologique du locuteur, (est-il triste, joyeux, en colère, etc. ?), ou plus simplement, des informations sur le sens de la phrase (S'agit-il d'une question, d'une affirmation, d'un ordre, etc. ?).

L'environnement dans lequel se situe la communication joue également un rôle important. **L'environnement** est essentiel dans le contrôle de la production vocale. En effet, le but premier de la production vocale est la communication d'informations à autrui. Toutefois, un élément indissociable à la communication est l'intelligibilité. En effet, il est nécessaire d'être intelligible pour être compris et ainsi, être sûr que son message soit entendu puis correctement interprété. Ainsi, si la communication s'effectue dans un environnement bruyant, pour se faire comprendre il est nécessaire d'augmenter l'intensité de sa voix pour émerger du bruit de fond. De la même manière, si l'environnement est très réverbérant, un locuteur a tendance à ralentir son débit de production afin que chaque unité phonétique soit compréhensible, et qu'elle ne soit pas noyée dans les réverbérations des unités précédentes (Pelegrín-García et al., 2011). C'est le cas des curés qui parlent lentement lors de leurs sermons dans les églises. Dans certains cas on observe également que le locuteur adopte une stratégie d'hyper-articulation, dans le but de parler le plus distinctement possible. De la même manière, en fonction du contexte de la situation, la distance qui sépare les interlocuteurs par exemple, influe sur la production de la voix (Traunmüller et Eriksson, 2000).

**L'auditeur** n'est pas sans reste dans la production vocale, car il peut également influencer la façon de parler du locuteur. En effet, si notre interlocuteur n'est pas attentif, notre stratégie de communication est différente. Nous produisons alors une voix plus forte et autoritaire afin de capter son attention. En effet, la stratégie de communication employée par le locuteur est fonction de son interlocuteur. Des voix hyper-articulées et d'un débit ralenti sont utilisées pour parler à des personnes malentendantes, étrangères ou à des enfants en bas âge (Lindblom et al., 1992).

**Le contexte du message** quant à lui, influence également la production. Les mots prévisibles du message n'attirent pas forcément l'attention de l'auditeur. Le locuteur fait alors moins d'efforts sur ce type de mots que sur les mots qui ne sont pas prévisibles (Jurafsky et al., 2001)

Enfin, **le retour auditif** de sa propre voix, influence également notre production vocale. Nous avons déjà mentionné le cas d'un rapport signal sur bruit trop faible, engendrant une augmentation de la production vocale (effet Lombard). Il a également été observé d'autres mécanismes de régulation de la voix portant sur d'autres paramètres que sur l'intensité. En effet, on observe le phénomène souvent décrit comme la réaction inverse de l'effet Lombard : l'effet *sidetone* (Lane, 1970). Il s'agit de la diminution de l'intensité de la voix lorsque le retour auditif est amplifié. Ce type de régulation est également observé sur la fréquence fondamentale (*pitch-shift reflex*) (Burnett, 1998), sur le timbre de la voix (*filtered auditory feedback*) (Garber et Moller, 1979; Garber et al., 1981), ou encore un retour auditif différé (*delayed auditory feedback*) (Lee, 1950; Stuart et al., 2002). Dans tous les cas, le locuteur ayant un retour auditif différent de sa production vocale a un réflexe tendant à compenser les effets de sa voix. Ainsi, s'il entend sa voix plus aiguë qu'elle ne l'est, il diminue la fréquence fondamentale. Ou encore, si les hautes fréquences de sa propre voix sont filtrées il tendra à augmenter

l'énergie des fréquences élevées pour compenser le retour auditif. Il existe cependant plusieurs types de retour auditif (par voix aérienne directe et réfléchi, par voie osseuse), ainsi que de retours kinesthésiques par la sensation que les vibrations procurent. Notons que le retour auditif de sa voix par voie aérienne est modifié en fonction de sa propre morphologie, et par l'environnement. En effet, dans un milieu fortement réverbérant, notre retour auditif est influencé par les paramètres de l'environnement. Dans un tel environnement, la durée des segments voisés sera allongée, la fréquence fondamentale plus élevée et l'énergie plus grande (Pelegrín-García et al., 2011) dans le but d'être intelligible soi-même ou par prédiction de l'intelligibilité.

Il existe de grandes théories expliquant ces phénomènes. La première est de l'ordre de la régulation psychologique réflexe. Cette théorie tend à expliquer **l'effet Lombard** par le retour auditif que le locuteur a de sa propre voix. Ainsi, la régulation a pour but de maintenir un rapport signal sur bruit constant ou tout du moins acceptable dans le but d'entendre sa propre voix. Il s'agit là d'un comportement physiologique réflexe dans le sens où cette régulation est difficile à inhiber car les mécanismes de régulation sont rapides. Une deuxième explication tend à expliquer cette régulation par des mécanismes cognitifs et communicationnels (Lindblom, 1990). Dans le cas précédent, l'objectif était de s'entendre soi-même, tandis que dans ce second cas, l'objectif est de se faire entendre. Le but premier de cette production est alors la recherche de l'intelligibilité auprès de son interlocuteur. Cette deuxième théorie engendre inévitablement des mécanismes cognitifs complexes permettant de prédire l'intelligibilité de sa propre voix en fonction des différents éléments environnementaux et de son interlocuteur. Nous ne rentrons toutefois pas dans les détails concernant le débat qu'il peut exister entre ces théories, et nous renvoyons le lecteur plus intéressé à (Garnier, 2007) pour une discussion sur le sujet.

### 3.5 Paramètres paralinguistiques

---

L'ensemble de ces informations qui ne correspondent pas à une information linguistique est appelé information paralinguistique. Les variations des paramètres paralinguistiques peuvent être analysées à travers deux notions : la **prosodie paralinguistique**<sup>1</sup> et la **qualité de voix**. Nous détaillons ici ces termes en commençant par la prosodie ainsi que la qualité de voix.

---

<sup>1</sup> D'une manière générale, la **prosodie** est l'inflexion, le ton, la tonalité, l'intonation, l'accent, la modulation que nous donnons à notre langage oral en fonction de nos émotions et de l'impact que nous désirons avoir sur nos interlocuteurs. Source : <http://www.vulgaris-medical.com>

### 3.5.1 La prosodie<sup>1</sup>

Pour définir la prosodie nous nous référons à la définition proposée par Di Cristo (Di Cristo, 2000)

« La prosodie (ou la prosodologie) est une branche de la linguistique consacrée à la description (aspect phonétique) et à la représentation formelle (aspect phonologique) des éléments de l'expression orale tels que les accents, les tons, l'intonation et la quantité, dont la manifestation concrète, dans la production de la parole, est associée aux variations de la fréquence fondamentale (F0), de la durée et de l'intensité (paramètres prosodiques physiques), ces variations étant perçues par l'auditeur comme des changements de hauteur (ou de mélodie), de longueur et de sonie (paramètres prosodiques subjectifs). Les signaux prosodiques véhiculés par ces paramètres sont polysémiques et transmettent à la fois des informations para-linguistiques et des informations linguistiques déterminantes pour la compréhension des énoncés et leur interprétation programmatique dans le flux du discours. » (Di Cristo, 2000) p.15

La prosodie correspond donc aux variations de F0, d'intensité et de durée au cours d'un énoncé. Ainsi la prosodie joue plusieurs fonctions dans le langage parlé. Elle peut avoir un rôle linguistique, pragmatique (intentions, attitudes), émotionnel et idiolectal (âge, sexe, identité). D'un point de vue linguistique, celle-ci joue un rôle de structuration de l'énoncé, de modalisation (expression de la déclaration, de l'interrogation,...), ainsi que de focalisation (mise en perspective d'une partie de l'énoncé), (cf. (Di Cristo, 2000) pour une discussion autour des rôles fonctionnels de la prosodie). Nous nous intéressons particulièrement dans cette étude à l'aspect paralinguistique de la prosodie. Les variations de tempo (débit de parole), de registre, d'amplitude (force de voix) permettent de distinguer les émotions fortes (joie, colère), des émotions modérées (ennui, satisfaction) (Mozziconacci, 1998) ainsi que probablement la voix criée.

Malgré tout, le terme prosodie n'englobe pas la totalité des variations qui sont observées sur un signal de la parole. Dans ce sens, la qualité de voix vient compléter ce manque.

### 3.5.2 La qualité de voix

On définit régulièrement la qualité de voix par une définition par défaut : toutes les informations qui ne sont ni linguistiques ni prosodiques. Or, de plus en plus d'auteurs s'accordent à dire que la qualité vocale constitue une dimension prosodique à part entière (Campbell et Mokhtari, 2003; D'Alessandro, 2006; Pfitzinger, 2006).

Selon (Laver, 1980), la qualité de voix inclue également la configuration des lèvres, des différentes parties de la langue du pharynx et de sa hauteur. Néanmoins, celle-ci se cantonne en général à la classification du comportement glottique (voir (Audibert, 2008) pour une discussion sur le sujet). Dans certains cas, la qualité de voix peut également englober des caractéristiques spectrales (le timbre). Toutefois, nous nous concentrons ici uniquement sur le comportement glottique. La qualité de voix se réfère donc principalement aux différents modes de vibrations des cordes vocales qui la modifie (Laver, 1980).

Laver a réalisé une classification des différents aspects de la qualité de voix en fonction de 3 paramètres de tension musculaire au niveau du larynx (la tension d'adduction, la compression médiale et la tension longitudinale).

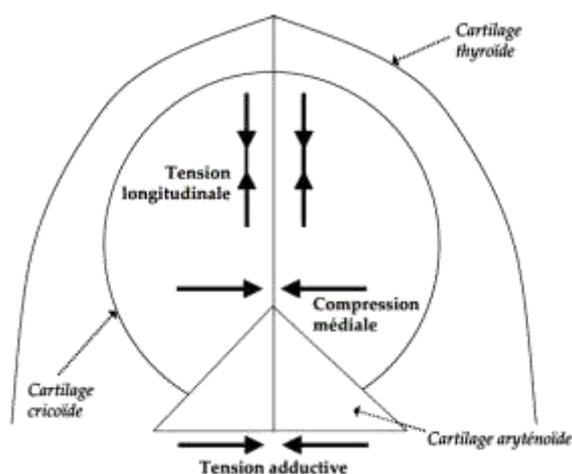


Figure 3-15: Schématisation des actions musculaires au niveau des cordes vocales décrivant les modes de vibrations des cordes vocales (selon (Nî Chasaide et Gobl, 1999) et issue de (Audibert, 2008))

Suivant ces trois descripteurs, Laver décrit, entre autre, les qualités de voix suivantes (Laver, 1980):

- La voix modale (*modale voice*) : tension neutre des muscles durant la phonation. Il s'agit de la voix de référence.
- La voix soufflée (*breathy voice*) : tension médiale et d'adduction faible. La phase de fermeture de la glotte est incomplète créant ainsi des turbulences et une sensation de voix soufflée. Le cas extrême de la voix soufflée est le cas de la voix chuchotée (*whispered voice*). La glotte est ouverte et les tensions faibles. La voix n'est alors créée que par des bruits de turbulences.
- La voix pressée (*pressed voice*) : la tension médiale et d'adduction sont très élevées. La fermeture de la glotte étant fermement maintenue par la tension des muscles, la vibration se fait difficilement et avec une F0 plus faible qu'en voix modale. Dans le cas de F0 très faible (env. 40 Hz), il est possible d'entendre chaque impulsion glottique. Ce type de phonation est appelé voix craquée (*creaky voice*)

Toutefois, la catégorisation proposée par Laver a montré ses limites pour la description de certains comportements vocaux tels que l'effort vocal. Ainsi, D'Alessandro (2007), propose la classification de la qualité de voix selon les 4 dimensions prosodiques :

- La dimension des registres vocaux: liée aux mécanismes de vibrations des cordes vocales. Le mécanisme 0 correspondant à la qualité de voix *vocal fry* et le mécanisme 2 à la voix de fausset (*falsetto voice*)
- La dimension tendue/relâchée : directement liée à la position des cartilages aryténoïdes et ainsi à la force avec laquelle les cordes vocales sont collées l'une à l'autre.
- La dimension d'apériodicité : liée au bruit additionnel continu, ou bruit lié aux variations de la période (*jitter*) ou de l'amplitude (*shimmer*) des vibrations des cordes vocales.
- La dimension d'effort : en rapport avec la sonie de la parole. Cette dimension traduit alors selon D'Alessandro une caractéristique importante des accentuations de la parole. L'augmentation de l'effort vocal pourrait être selon (D'Alessandro, 2006) dû à une tension et un amincissement des cordes vocales, par l'action des muscles extrinsèques, sur les cartilages cryco-thyroïdes et par la contraction du *vocalis*. Il ajoute que la pression subglottique doit alors être plus importante et que certains ajustements de la source glottique ainsi que du conduit vocal, peuvent également jouer un rôle pour les variations de l'effort vocal. Les aspects physiologiques liés à l'augmentation de l'effort vocal sont discutés plus en détails dans le chapitre suivant.

La dimension de l'effort qui est considéré comme une des dimensions de la qualité de voix définit l'effort vocal, ou plus précisément la voix criée, comme étant une voix forte produite avec beaucoup de forces musculaires et ayant les caractéristiques suivantes : pression subglottique élevée, tension forte des cordes vocales, débit d'air moyen, et une amplitude de voisement élevée (D'Alessandro, 2006). Toutefois, le terme effort vocal est, dans la littérature, abondamment utilisé mais reflète régulièrement des aspects et des significations divergentes. Nous proposons ainsi, dans le chapitre suivant, de décrire plus en détail ce terme.

# Effort vocal et effort vocal

---

Le terme *effort vocal*, bien que largement utilisé, souffre d'un manque de définition claire. En effet, en fonction du domaine dans lequel ce terme est employé son interprétation n'est pas forcément la même. Ainsi, d'un point de vue médical (ORL) ce dernier définit l'effort devant être fourni par un patient atteint d'une pathologie vocale pour parler. Le terme effort vocal peut également définir le forçage vocal effectué dans un environnement bruyant ou plus généralement quand le retour auditif est perturbé (i.e. effet Lombard) ; comme le cas de surdité ou de traumatisme sonore. Le fait de parler pendant une très longue période de la journée, comme pour les enseignants, est également considéré comme de l'effort vocal, étant donnée les différentes pathologies que la phonation prolongée peut engendrer. La voix d'une personne énervée constitue également un effort vocal. Au regard de ces différentes interprétations nous avons voulu préciser ce que nous entendons, dans cette étude, par le terme effort vocal. Dans cet objectif, nous débutons ce chapitre en listant les causes et circonstances faisant intervenir l'effort vocal. Par la suite, nous mentionnons la définition qui nous concerne. Enfin, nous évoquons la problématique liée à la mesure de l'effort vocal fourni par un locuteur, et la manière dont nous contrôlons la production de celle-ci.

## 4.1 Causes et circonstances

---

En guise d'illustration donnons les différents résultats obtenus par (Garnier, 2007) lors d'une enquête visant à identifier les différentes situations qui engendrent un forçage vocal. Le terme forçage est un terme encore plus large que l'effort vocal. C'est pourquoi, elle recense que 1 % des personnes interrogées mentionnent une situation de production vocale lors d'une prise de parole en posture inconfortable, que 11% mentionnent une situation de trouble émotionnel (trac, timidité) tandis que 88% mentionnent l'effort vocal comme situation de forçage vocal.

Les causes de cet effort vocal sont alors dissociées en différentes situations. A savoir que 8% mentionnent que l'effort est liée à une mauvaise acoustique de salle, 12% à une hauteur de voix inhabituelle (cas du chant), 19 % par un usage prolongé de la voix et 53% pour un usage de voix plus forte. Encore plus approfondi, l'usage de la voix forte est induit pour 3% des 53% précédents, par une

situation d'urgence (crie d'alerte) 3% par l'éloignement de l'interlocuteur, 13% par des émotions (colère, joie, excitation) et 63% par la présence de bruit ambiant. Sur cette enquête, seulement 1,4% des personnes considèrent la distance comme une situation de forçage vocal.

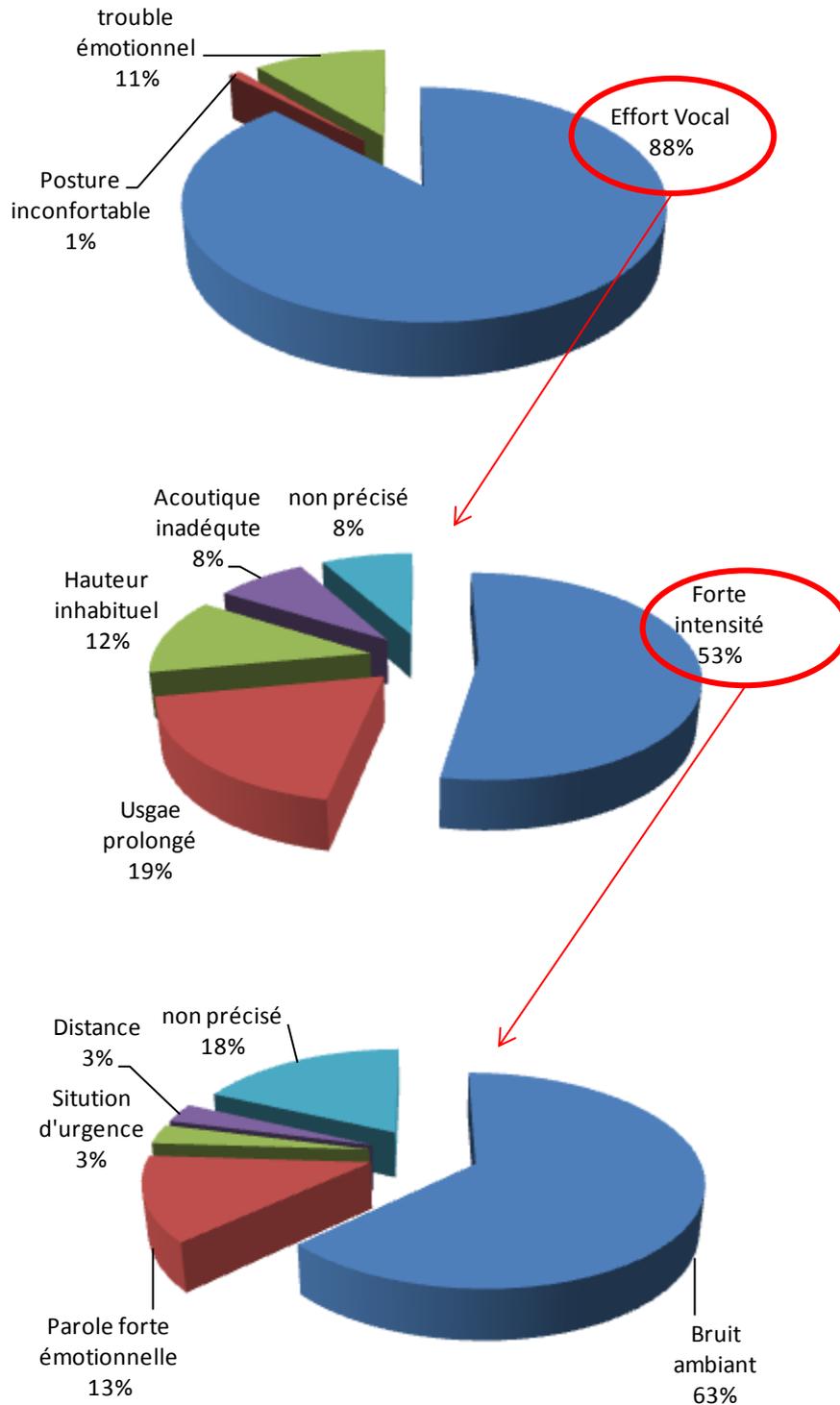


Figure 4-1: Occurrences des différentes situations de forçage vocal (en haut), de l'effort vocal (au milieu) et de l'usage de voix à forte intensité (en bas) (d'après (Garnier, 2007))

L'effort vocal au sens large peut se manifester pour plusieurs raisons et dans différentes circonstances que nous pouvons classer en trois catégories.

1. D'une part l'effort vocal peut être réalisé dans un but de maintenir la communication avec un interlocuteur dans des conditions adverses. Dans cette situation précise, l'effort est réalisé dans un but de conserver l'intelligibilité de la voix face aux conditions adverses (distance, bruit, acoustique de la salle, etc.). Il s'agit donc d'un effort lié à la communication. De plus ces éléments interagissent également. En effet, dans le but de se faire comprendre dans un environnement bruyant, il est nécessaire d'augmenter le niveau de production de sa voix. La source de bruit peut être très variée comme de la musique, d'autres personnes qui conversent, des machines dans une usine, le bruit du vent, etc... On remarque d'ailleurs, en fonction du type de bruit dans lequel les locuteurs sont plongés, des stratégies de production de voix différentes (cf. Garnier, 2007)). Par exemple, une situation de communication à distance va être différente en fonction du bruit et de l'acoustique de l'environnement (Pelegri-García et al., 2011).
2. Une autre situation est liée à la pathologie ou encore à certaines circonstances qui demandent un effort vocal afin de parler à quelqu'un, mais sans conditions adverses. C'est le cas lorsqu'on essaie de parler tout en portant une charge lourde, par exemple, ou encore lorsqu'on souffre d'une pathologie qui demande alors beaucoup d'effort dans le but de produire un son. L'objectif est là encore d'être intelligible.
3. Une troisième catégorie est la production d'effort vocal pour cause d'émotivité. Un cri de joie, un cri de colère, un cri d'effroi ne sont pas directement associés à la communication verbale mais reflètent simplement l'état d'esprit d'une personne. Ce type de production vocale n'est pas forcément destiné à une personne. Il est incontrôlé et peut provoquer des traumatismes (réversibles) lors de sa production.

Après ce bref tour d'horizon concernant les causes et circonstances de l'apparition de l'effort vocal, penchons nous à présent sur la définition qui nous concerne.

## **4.2 Effort vocal : définition nous concernant**

---

Dans notre études nous parlons d'effort vocal dans le sens où il a été défini par (Traunmüller et Eriksson, 2000) :

*“Vocal effort is the quantity that ordinary speakers vary when they adapt their speech to the demands of an increased or decreased communication distance.”*

Ainsi dans notre situation, l’effort vocal traduit une variation de production vocale en fonction de la distance de communication par rapport à la production vocale effectuée pour une distance conversationnelle.

Remarquons également que la production de voix chuchotée est aux yeux de cette définition, assimilée à un effort vocal. Ce qui n’est pas sans raisons. En effet, il est nécessaire de faire un effort afin que les cordes vocales ne vibrent pas et que l’air expulsé par les poumons engendre un son ; ce qui implique un effort de constriction. Ainsi, pour une augmentation de distance de communication il semble abusif de parler « d’augmentation » d’effort vocal.

Le terme *voix projetée* est parfois utilisé pour définir une voix criée pour la communication à distance (Giovanni et al., 2007). Or, le terme « voix projetée » (ou *voix implicatrice*) selon (Lehuche et Allali, 2010), n’implique pas forcément une voix criée. En effet, il s’agit d’un comportement vocal et corporel reflétant la volonté de se faire comprendre ainsi que d’obtenir l’attention de son interlocuteur (d’où le terme voix implicatrice). Ce comportement se traduit par le regard fixé sur l’interlocuteur, par l’utilisation d’un souffle abdominal et le redressement du corps (Lehuche et Allali, 2010).

On pourrait alors, toujours dans le cas de communication à distance, définir deux types d’effort vocal :

1. **l’effort vocal de retenu**, dans le but de produire des voix de plus en plus chuchotées jusqu’à des voix non audibles (NAM, Non audible Murmure)
2. **l’effort vocal de projection**, dans le but de produire une voix qui puisse être intelligible à des distances de plus en plus grandes.

En gardant la même philosophie de dénomination, pour l’effet Lombard on pourrait alors associer le terme **effort vocal d’émergence**, ou dans le domaine médical **effort vocal de production/phonation**. Il ne s’agit là que de certaines propositions visant à éclaircir le terme « effort vocal » en fonction du but de sa production. **Toutefois dans nos travaux ce terme est clairement défini par la définition précédente et c’est pourquoi, nous conserverons le terme effort vocal pour parler de l’effort vocal de projection.**

De plus, il existe plusieurs termes liés à l’effort vocal. Nous proposons ici, une échelle indiquant les différents termes utilisés dans la littérature retraçant les différents stades de l’effort vocal (cf. Figure 4-2). Le passage d’un niveau à un autre est régi par une augmentation de distance.

L’ensemble de ces niveaux d’effort s’accompagne de nombreuses modifications (timbre, durée, formant, etc...) toutefois, nous les classons en fonction des plus révélateurs à savoir l’intensité et la F0.

Dans un premier temps nous considérons la voix modale (ou conversationnelle ou parlée) comme référence de notre échelle. Cette voix correspond alors à un effort de production « nul ». Une légère augmentation de l'intensité de la voix se traduit dans la littérature par le terme *voix forte* (*loud voice*). Il s'agit principalement d'une augmentation de l'intensité de la voix engendrant également une augmentation de la F0. Le niveau supérieur correspond au terme *voix criée* (*shouted voice*) impliquant une intensité et une F0 encore plus élevée. Le terme qui indique un effort encore plus grand est *voix hurlée* (*screamed voice*). Nous estimons également qu'il existe un niveau supérieur, dont nous n'avons pas trouvé de correspondance dans la littérature, qui correspondrait à l'égosillement. Ce type de voix n'est que très rarement utilisée car elle est douloureuse et peut causer de graves lésions au niveau des cordes vocales. Ces deux derniers types de voix s'accompagnent d'intensité de production forte ainsi que de vibrations des cordes vocales, qui ne semblent plus parfaitement périodiques, et dont la vibration est dans une certaine mesure chaotique (Nieto, 2008).

A l'inverse quand l'effort vocal diminue, les termes utilisés sont la *voix douce* (*soft voice*) et la *voix soufflée* (*breathy voice*). Cette dernière a la particularité de présenter dans les zones normalement voisées des absences épisodiques de F0. L'absence systématique de vibration correspond au niveau inférieur et est appelé *chuchotement* ou *murmure*. Le dernier niveau d'effort, le plus faible, correspond au terme NAM (*non audible murmure*) ou *voix silencieuse* (*silent voice*).

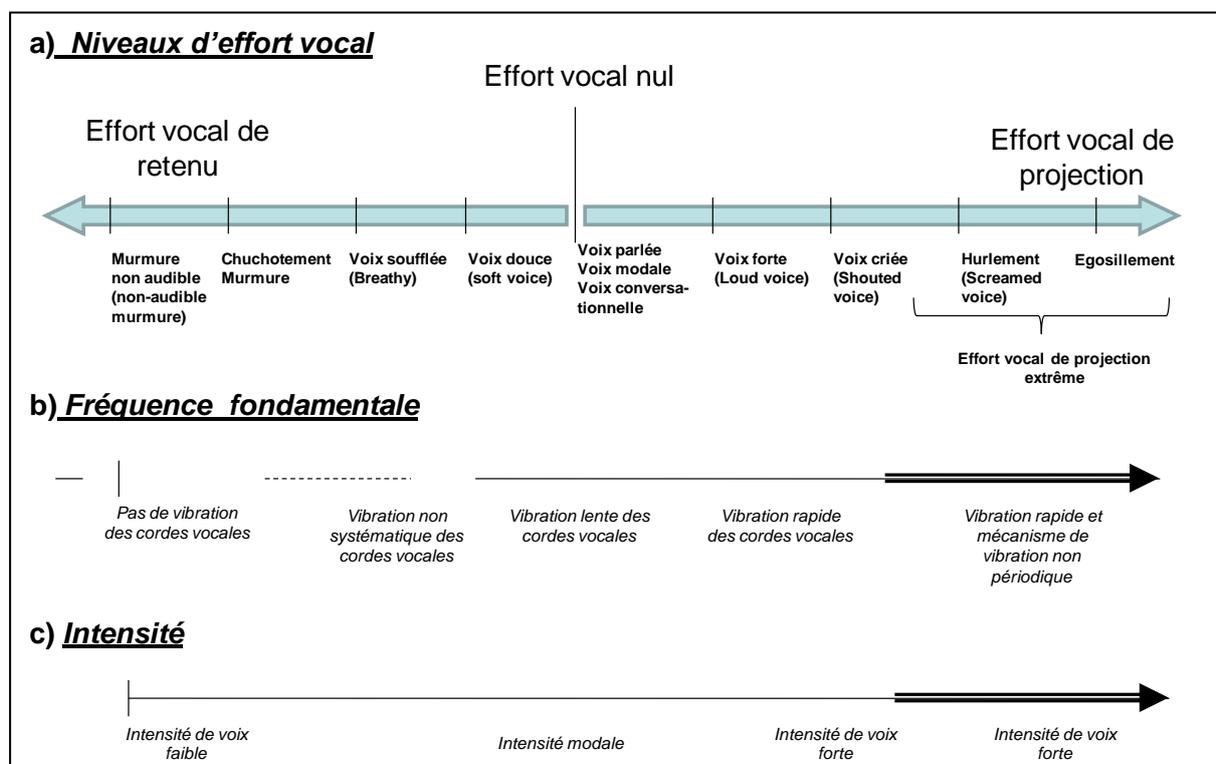


Figure 4-2: Classement des différents termes relatif à l'effort vocal et caractérisé par la F0 et l'intensité

Nous classons ici les différents niveaux d'effort par le biais de l'intensité et la F0. Cependant, ces deux indicateurs sont des conséquences de l'effort vocal qui ne permettent pas, à eux seuls, de quantifier la quantité d'effort fournie. Existe-t-il alors une méthode de mesure permettant de le faire ?

### 4.3 La mesure de l'effort vocal

---

L'effort vocal est fondamentalement une grandeur objective. Toutefois, le manque de compréhension des mécanismes de production et de perception de l'effort vocal, et de ce fait l'absence de mesures objectives de l'effort, rendent cette notion plus ou moins subjective.

En effet, il n'existe pas à ce jour de méthode permettant de mesurer directement le taux d'effort fourni par un locuteur. C'est pourquoi, dans la majorité des cas, l'effort vocal est mesuré par l'intermédiaire d'une grandeur qui influe sur la production de l'effort vocal. Ainsi, dans le cas de l'effort vocal d'émergence (effet Lombard) l'augmentation de l'effort vocal se traduit par l'augmentation du niveau sonore du bruit ambiant. Or, il est largement admis que le type de bruit ambiant influe sur la production vocale et donc sur l'effort. D'autre part, certains locuteurs ont besoin de fournir moins d'effort que d'autres pour émerger du bruit ambiant. D'une manière plus générale, l'effort vocal peut également se représenter en fonction du rapport signal à bruit.

Dans le cas où l'effort vocal intervient après une longue période de parole (cas des enseignants) l'effort vocal se mesure via le temps de parole. Dans des cas pathologique, l'effort vocal peut être représenté via un degré d'évolution de la pathologie par exemple (degré qui reste également qualitatif). Toutefois, dans la majorité des cas, l'effort vocal se quantifie par le biais du niveau acoustique de production vocale, plus précisément par la variation d'intensité par rapport à une phonation modale, ou encore par la variation de la fréquence fondamentale. Une dernière manière d'évaluer le taux d'effort vocal fourni est de réaliser des tests subjectifs afin de quantifier l'effort. Cette méthode reste toutefois hasardeuse du fait que l'échelle de l'effort vocal qui devrait être utilisée est qualitative (pas d'effort, un peu d'effort, moyennement d'effort, etc...). Il est également possible d'utiliser une échelle relative de l'effort (plus ou moins d'effort) mais la problématique reste la même.

Pour identifier les corrélats liés à l'augmentation de l'effort vocal la quantification de ce dernier est primordiale. Cependant, il n'existe à ce jour aucun appareil ou méthode de mesure que l'on pourrait appeler « effort-mètre ». On pourrait imaginer, un appareil permettant de relever certains signaux physiologiques, comme par exemple, la tension des muscles du larynx, la pression subglottique mêlés à des mesures acoustiques telles que l'intensité et la F0. Sur la base de ces relevés ainsi que par une

modélisation précise, une valeur correspondant à l'effort vocal fourni par le locuteur pourrait être calculée. Dans un tel cas, notre étude aurait été facilitée.

En effet, dans ce sens une première étude permettant de quantifier l'effort vocal fourni en fonction de la distance aurait été menée. Dans un second temps, des relevés de paramètres de la voix en fonction du taux d'effort nous auraient permis d'utiliser cette relation dans un synthétiseur de voix permettant ainsi de générer tel ou tel taux d'effort qui à son tour engendrerait une perception de la distance. Malheureusement, « l'effort-mètre » n'existe pas à ce jour et ainsi l'approche que nous venons de décrire est à ce jour utopique. De ce fait, pour mener à bien notre étude nous tentons d'établir les liens entre les variations de paramètres vocaux et les variations de distance (cf. Figure 4-3). En effet, la distance engendre une augmentation de l'effort vocal qui à son tour engendre des modifications sur les paramètres vocaux. Ainsi une relation entre les variations des paramètres et la distance est une façon de quantifier l'effort fourni.

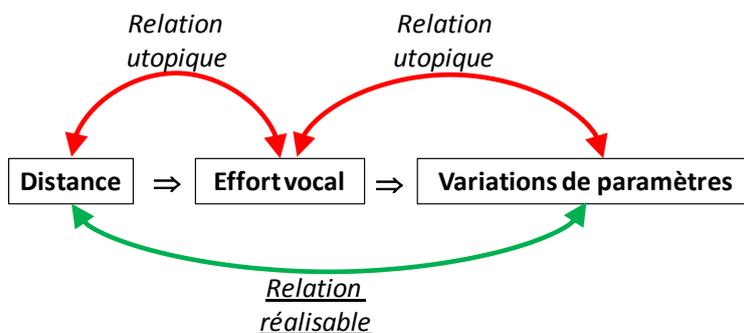


Figure 4-3: Relation entre variations des paramètres et distance de communication

Il faut souligner que, même s'il existait une méthode de mesure fiable de l'effort vocal fourni par un locuteur, la relation entre la distance et l'effort vocal fourni n'est pas triviale. La Figure 4-4 schématise (de manière symbolique) les relations entre effort vocal, intensité locale et distance de communication. Dans un premier temps, pour une distance de communication donnée l'effort vocal devant être fourni n'est pas toujours identique et peut être dépendant de l'environnement dans lequel les interlocuteurs sont plongés. En effet, de jour il est nécessaire de fournir plus d'effort que de nuit pour être intelligible à une distance donnée. Cette relation est directement liée au phénomène météorologique de l'environnement ainsi qu'à la typologie des lieux (champ libre, ville, etc ...) mais également de la nature du sol (béton, herbe, etc ...). Ces différents éléments influent considérablement sur l'effort vocal nécessaire et suffisant à la communication. Dans un second temps la relation entre l'effort vocal et l'intensité (locale) de la voix peut également varier en fonction de la stratégie plus ou moins efficace adoptée par le locuteur. Enfin, toutes ces relations sont également dépendantes du locuteur en lui-même et plus particulièrement des caractéristiques de sa voix (voix qui « porte » ou non) ! Sur cette figure il n'est pas représenté l'évolution de la fréquence fondamentale en fonction de l'effort vocal ou

encore de la distance. Il est fort probable que celle-ci soit à l'image des exemples déjà mentionnés à savoir très variable selon les cas.

Au regard de ces relations multiples entre les différents paramètres, il est impensable de prétendre donner une information de distance absolue par le biais de l'effort vocal. C'est d'ailleurs ce que l'on observe dans la littérature concernant la PAD qui varie beaucoup d'une étude à l'autre ou même d'un auditeur/locuteur à l'autre. C'est pourquoi, dans cette étude, nous n'envisageons pas de transmettre une information de distance absolue mais plutôt une information de la plage de distance (i.e. proche, moyen, loin).

Dans notre étude, qui consiste à étudier la voix dans le cas de distances de communication différentes, on peut s'interroger sur la nécessité d'évoquer le terme « effort vocal » puisque nous ne sommes pas en mesure de le mesurer ou de le quantifier. D'ailleurs pour cette raison, la relation « distance – effort vocal » et la relation « effort vocal – variation de paramètres » ne sont pas accessibles (voir figure ci-dessous), seule la relation « distance – variations de paramètres » pouvant être étudiée directement. La réponse est que nous ne cherchons pas à mesurer l'effort vocal mais nous l'utilisons pour *comprendre* les variations des paramètres de la parole causées par la distance de communication. En effet, sans la notion d'« effort vocal », nous ne pouvons que constater les variations des paramètres sans comprendre pourquoi elles sont ainsi.

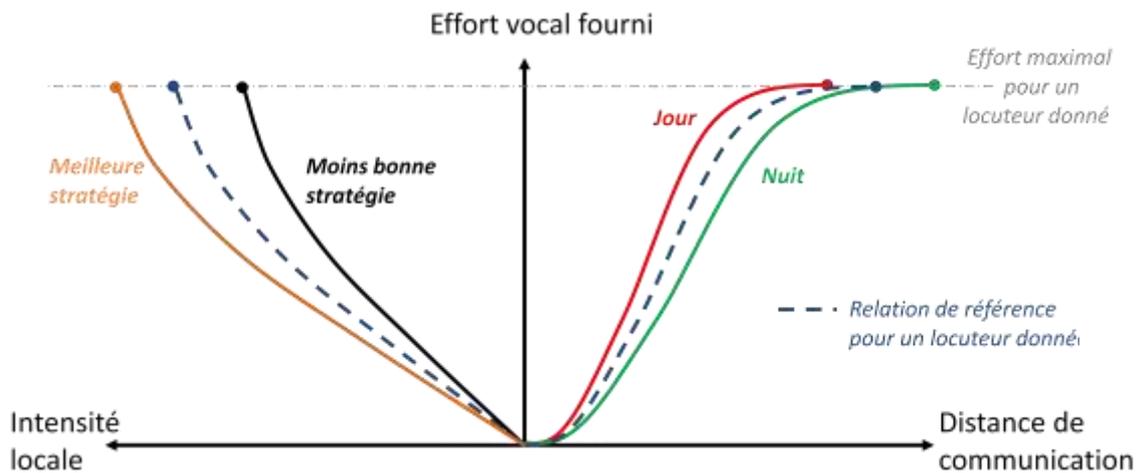


Figure 4-4: Schématisation des relations multiples entre intensité locale, distance de communication et effort vocal

# Caractéristiques des voix en communication à distance : état de l'art

---

*« Parler de la parole pour mieux parler »*  
— Jérôme Liss

Dans la communication à distance le locuteur ajuste sa voix dans le but d'assurer une bonne situation de communication. Ainsi, pour une grande distance de communication le locuteur crie afin de pallier la diminution d'intensité due à la dispersion géométrique (sphérique) de l'onde sonore. Dans le cas d'une courte distance, le locuteur emploie plutôt une voix chuchotée dans un objectif d'assurer un niveau sonore adéquat (pas trop intense) au niveau des oreilles de l'auditeur.

Ces modifications de la voix faisant intervenir un ajustement précis de l'effort vocal, qui est une fonction de la distance de communication, permettent à l'auditeur de déterminer la distance du locuteur. Ainsi, pour notre objectif de transformer une voix parlée en une voix criée ou chuchotée permettant de refléter la distance d'un locuteur, la connaissance des évolutions des caractéristiques de la voix est primordiale.

Ces différentes voix utilisées en communication à distance, présentent des caractéristiques très différentes comparées à la voix modale et engendrent d'importantes variations des paramètres de la voix. Dans le cas de la voix chuchotée, l'absence de vibrations des cordes vocales, ainsi que la production à faible intensité, s'accompagnent de variations des paramètres vocaux causées par ces dernières. La production de la voix criée modifie la configuration de la glotte et du conduit vocal afin de permettre de produire une voix de forte intensité. Ces ajustements à différents niveaux de l'appareil phonatoire, induisent également de fortes variations des paramètres de la parole. D'autre part, ces voix peuvent également s'accompagner de réorganisations prosodiques (ce qui relève de la stratégie du locuteur) dans le but d'accroître l'efficacité de la communication, ou encore de conserver l'intelligibilité du message.

Ce chapitre présente, un état de l'art sur les connaissances actuelles concernant les voix chuchotées, ainsi que les voix reflétant un fort effort vocal dans le but d'une communication à distance. Nous reprendrons les différentes observations connues à travers la littérature. Nous reprenons également certaines interprétations concernant ces variations, dans le but de déterminer par quelles actions physiologiques elles sont induites. Ce chapitre se compose de trois parties. La première est consacrée à l'état de l'art concernant les variations des paramètres des voix chuchotées. La deuxième se consacre à la voix reflétant un fort effort vocal. Enfin, la troisième et dernière partie, traite de l'intelligibilité de ces voix mais également de la capacité à reconnaître un locuteur à partir de celles-ci. En effet, les modifications de nature de la voix, engendrent des changements drastiques des paramètres, dont les conséquences sur l'intelligibilité et l'identification peuvent être significatives. Soulignons cependant que, dans l'ensemble, la dégradation de l'intelligibilité n'est pas aussi grande que ce que nous pourrions penser.

## **5.1 État de l'art des connaissances actuelles sur la voix chuchotée**

---

La motivation première de toute communication est de se faire comprendre facilement et efficacement par son interlocuteur. Ainsi, pour une communication à courte distance, dans le but de garantir une intensité de la voix adéquate au niveau des oreilles de l'auditeur, le locuteur produit une voix de faible intensité. En effet, un niveau de production modale, produit à quelques centimètres de l'auditeur, provoque une intensité beaucoup trop forte et une gêne à l'écoute. De plus, une communication à courte distance a souvent pour but la discrétion. Ainsi, dans le but d'une communication discrète, la voix chuchotée peut également être utilisée. Ainsi nous estimons que la voix chuchotée qui résulte d'une communication à courte distance est issue de deux contributions : d'une part le caractère confortable de la communication et d'autre part la volonté de discrétion. Pour ces raisons, l'intensité d'une voix chuchotée se situe aux alentours de 40 dB SPL, soit 20 dB en dessous de la voix parlée. Les voyelles chuchotées ont une intensité diminuée de 20-25 dB par rapport aux voyelles vocalisées (Ito et al., 2005).

Dans le but de produire une voix de faible intensité, le locuteur met en œuvre des mécanismes de productions volontaires et/ou involontaires. Une voix chuchotée implique, de la part du locuteur, une stratégie permettant l'inhibition de la vibration des cordes vocales lors de l'expulsion de l'air. Malgré cette inhibition de la vibration, les cordes vocales doivent permettre de générer une turbulence lors du passage de l'air au travers la glotte. En effet, c'est le bruit des turbulences de l'air expulsé qui est à la

base des voix chuchotées. Ces différents ajustements vont engendrer plusieurs modifications que nous détaillons par la suite.

### 5.1.1 Le niveau glottique

L'une des caractéristiques principales d'une voix chuchotée est que le signal de la parole n'est à aucun moment périodique. Lors de la production d'une voix modale, l'air expulsé par les poumons provoque la vibration des cordes vocales qui, périodiquement, s'accolent et se séparent. Lors d'un chuchotement, les cordes vocales ne sont plus en contact. La glotte est ouverte de façon à provoquer une turbulence lors du passage de l'air expulsé par les poumons. Cette turbulence est également générée par l'interaction des flux d'air avec les bandes ventriculaires et l'épiglotte (Morris, 2003)

Il existe toutefois une *constriction de la glotte*, permettant de produire le bruit de turbulence. En situation de respiration, la glotte est largement ouverte et ne génère aucun son. Ainsi, une constriction de la glotte est nécessaire afin de produire la turbulence de la voix chuchotée. En effet, lors d'un chuchotement, l'aire glottique est alors de l'ordre de  $0,6 \text{ cm}^2$  (Sundberg et al., 2010) alors que lors de la respiration, elle est de l'ordre de  $1 \text{ cm}^2$  ((Sawashima, 1977) d'après (Hardcastle et Laver, 1999, p.48)). En phonation cette aire est de l'ordre de  $0,05$  à  $0,1 \text{ cm}^2$  (Hardcastle et Laver, 1999, p.48).

La *forme de l'ouverture glottique* est généralement en forme de 'Y' mais peut également adopter des formes en 'V' ou des ouvertures ovales (Sundberg et al., 2010). De plus, Zeroual et al. (2005) ont montré que lors d'un chuchotement, les bandes ventriculaires se rapprochent et que l'écartement entre les cordes vocales est plus grand au niveau des cartilages. Ils observent également une compression antéropostérieure de l'épilynx (région supra-glottique du larynx).

On constate également, afin d'éviter toute vibration des cordes vocales, que la structure supra-glottique se resserre au niveau des bandes ventriculaires (fausses cordes vocales) et que les cordes vocales sont recouvertes par ces dernières ((Tsunoda et al., 1997) d'après (Matsuda et Kasuya, 1999)). Ainsi, une voix chuchotée est entièrement produite à partir de bruits de turbulence.

### 5.1.2 Le niveau acoustique

L'ensemble des variations observées au niveau laryngé vont modifier la forme du conduit vocal et, de ce fait, sa fonction de transfert. Ainsi, au niveau acoustique, outre le fait que le signal soit entièrement composé de bruit, on observe des modifications des formants.

### 5.1.2.1 Les formants

Jovičić (1998), sur la base de voyelles maintenues en voix chuchotée et en voix modale par des hommes et des femmes, a mesuré les positions et les largeurs de bande des quatre premiers formants. La première observation de Jovičić est que **les fréquences centrales des formants** des femmes, sont plus élevées que tous les formants des hommes. Toutefois, il observe des taux de déplacement relatif qui sont similaires entre les hommes et les femmes. Il relève des déplacements significatifs de la position des formants 1 et 2 vers les hautes fréquences pour des voix chuchotées. Ces déplacements sont de l'ordre de 6-7 demi-tons pour F1 et de 1-2 demi-tons pour F2. Aucune modification significative n'a toutefois été observée pour les formants 3 et 4. Pour la voyelle /u/, Jovičić (1998) mesure un déplacement significatif des formants vers les basses fréquences (-2 demi-ton pour F1, -4 demi-ton pour F3 et -9 demi-ton pour F4).

L'étude réalisée par Sharifzadeh et al. (2010) va également dans ce sens. Ces derniers ont également mesuré des déplacements vers les hautes fréquences, plus importants sur le premier formant que sur le deuxième et aucune variation significative du troisième formant. Toutefois, dans cette étude, la voyelle /u/ subit également un déplacement vers les hautes fréquences des formants, contrairement aux mesures effectuées par Jovičić (1998). Sharifzadeh et al. (2010) montrent également que les formants sont plus élevés pour les femmes mais, contrairement à Jovičić (1998), ils montrent des variations de la position des formants moins fortes pour les femmes, en particulier pour F1.

Cette tendance à l'augmentation des 2 premiers formants a, par ailleurs, été confirmée par plusieurs études (Ito et al., 2005; Kallail et Emanuel, 1984a, 1984b; Matsuda et Kasuya, 1999; Petrushin et al., 2010; Swerdlin et al., 2010). Ito et al. (2005) montrent un déplacement de +30% à +60% pour le premier formant, et de 0% à +20% sur le second formant pour un locuteur (sexe inconnu). Ces proportions sont également observées par (Matsuda et Kasuya, 1999) qui mesurent un déplacement de F1 de +30%, et un déplacement de F2 de +6% pour des hommes. Swerdlin et al. (2010), quant à eux, relèvent des variations de +45% pour F1 et de +9% pour F2 chez des jeunes femmes. Toutefois Kallail et Emanuel (1984a) ne montrent qu'une différence significative du premier formant pour les femmes (Kallail et Emanuel, 1984a), ainsi que pour les hommes (Kallail et Emanuel, 1984b).

Plusieurs auteurs ont démontré que le déplacement des formants en voix chuchotée est lié à l'ouverture glottique ; plus précisément à **l'interaction entre la glotte et le conduit vocal**. En effet, une ouverture plus grande de la glotte engendre une impédance glottique plus faible (Barney et al., 2007; Matsuda et Kasuya, 1999; Swerdlin et al., 2010). Les conditions aux limites du conduit vocal sont ainsi modifiées. A partir de ce constat, et sur la base de la théorie acoustique de production vocale, le déplacement des formants peut être simulé par un système acoustique simple. A l'aide d'une telle modélisation, on constate que l'ouverture de la glotte engendre un déplacement des formants vers

les hautes fréquences et particulièrement pour les deux premiers formants (Morris, 2003). La Figure 5-1 illustre ce phénomène.

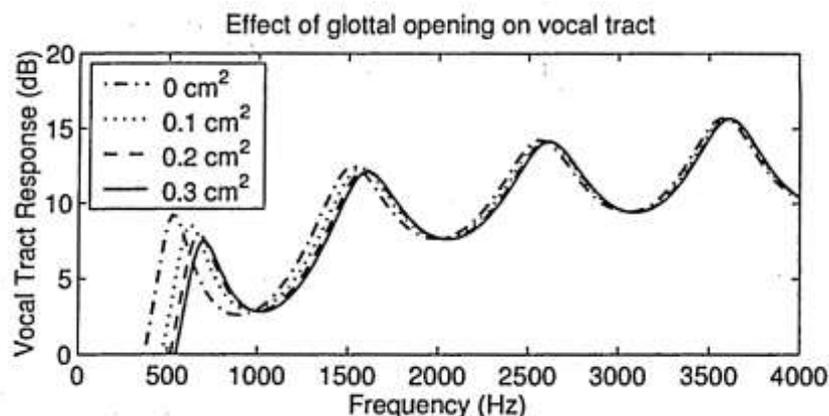


Figure 5-1: Conséquences de la variation de l'ouverture glottique sur le spectre d'une voyelle, calculées à partir d'une simulation du conduit vocal (Morris, 2003)

Concernant la **largeur de bandes** des formants, Jovičić (1998) observe que l'ensemble des largeurs est augmenté, et que cette augmentation est importante pour F1 et faible pour F2. Il observe également que la largeur de bande de l'ensemble des formants tend à avoir des valeurs similaires en voix chuchotée. C'est-à-dire que les largeurs de bande des formants tendent vers une valeur de 120 Hz. Aucun formant ne prédomine, engendrant ainsi un spectre relativement plat. A notre connaissance il s'agit là de la seule étude ayant mesurée la largeur de bandes des formants.

### 5.1.2.2 Spectre

**L'amplitude des formants** est également largement altérée lors de la production d'un chuchotement. L'amplitude des formants le long du spectre décroît plus rapidement en voix parlée qu'en voix chuchotée (Kallail et Emanuel, 1984b). Ainsi, la pente spectrale est altérée et le spectre a tendance à s'aplatir (Zelinka et Sigmund, 2010). En effet, la pente spectrale est de -2.86 dB/octave pour les voix chuchotées alors qu'elle est de -8,29 dB/octave pour les voix parlées (Zhang et Hansen, 2007)

**Le spectre de la voix chuchotée** comparé à celui d'une voix parlée, chute rapidement en-dessous de 500 Hz (Jovičić, 1998) et est généralement plus plat dans la région 200 Hz à 2000 Hz (Jovičić, 1998; Schwartz, 1970). La modification du spectre peut être liée aux modifications apportées à la source. En effet en fonction de la configuration glottique et des constriction effectuées lors de la production de voix chuchotée, le spectre de la source présente de grandes variations (Stevens, 1998). Selon Farner et al. (2008) la pente spectrale de la voix chuchotée, par rapport à la voix criée, s'annule voir s'inverse pour des fréquences inférieures à environ 3 kHz (cf. Figure 5-2).

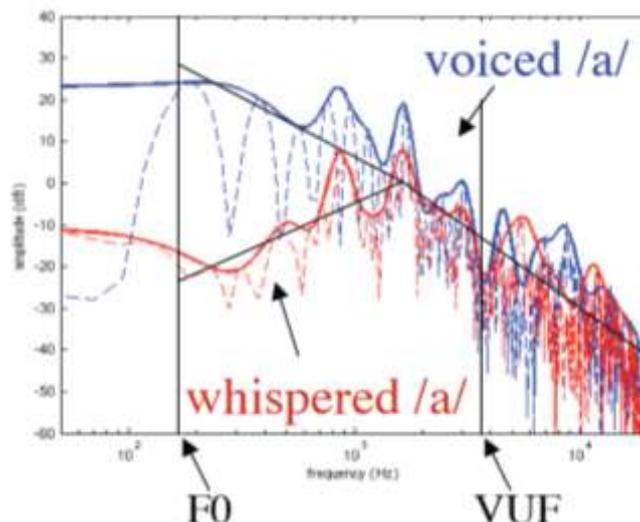


Figure 5-2: Spectre et enveloppe spectrale d'une voyelle /a/ parlée et chuchotée. VUF : voiced/unvoiced frequency (i.e. la fréquence maximale de voisement), (Farner et al., 2009).

### 5.1.3 Le niveau prosodique

Outre les phénomènes acoustiques, la voix chuchotée présente également des variations prosodiques. Mais, en l'absence de vibrations des cordes vocales, les variations prosodiques n'interviennent que sur l'intensité et la durée.

Concernant l'intensité, « Les obstruantes non-voisées ne sont pas affectées mais les sonorantes deviennent non voisées et leur amplitude décroît et ainsi, les fricatives sont plus intenses que les voyelles chuchotées »<sup>1</sup> (O'Shaughnessy 1999 p.67). En effet, il n'y a pas de différence significative de l'intensité des consonnes non voisées en voix chuchotée (max. 3,5 dB). La différence est significative pour des consonnes phonétiquement voisées ainsi que les nasales et les semi-voyelles qui diminuent d'environ 25 dB (Jovičić et Šarić, 2008). Cette grandeur est en accord avec (Ito et al., 2005) qui mesurent une diminution de 20-25 dB sur des voyelles.

En voix chuchotée le **débit de parole** est en général ralenti (Schwartz, 1972). On observe une augmentation de la durée des voyelles d'environ 35% (Heeren et Heuven, 2011) d'environ 10% pour les consonnes (Jovičić et Šarić, 2008) (avec certaines exceptions comme le /m/ (Schwartz, 1972)). De plus, les consonnes initialement voisées sont plus rallongées (15,3 %) que les consonnes non voisées (5,8 %) (Jovičić et Šarić, 2008). Outre le fait que les phonèmes soient rallongés en voix chuchotée, le débit ralenti des voix chuchotées peut également s'expliquer par des pauses plus longues (Andersson et al., 1996; Bonnot et Chevrie-Muller, 1991; Jovičić et Šarić, 2008; Traunmüller et Eriksson, 2000).

<sup>1</sup> Traduction personnelle de : « Unvoiced obstruents are unaffected, but sonorants become unvoiced and decrease substantially in amplitude so that fricatives are louder than whispered vowels »

## 5.2 État de l'art des connaissances actuelles sur la voix criée

---

La voix criée, ou la voix projetée, est produite par un locuteur dans le but de pallier la diminution de l'intensité causée par la distance (-6 dB par doublement de distance). Ainsi, un locuteur force sur sa voix afin d'augmenter l'intensité de celle-ci. Il s'agit là, de la nécessité première pour réaliser une bonne communication. Il existe également une deuxième motivation qui est la conservation de l'intelligibilité du message. En effet, le but de la communication à distance est de maintenir l'intensité au niveau des oreilles de l'auditeur suffisante d'une part, et d'autre part que celui-ci comprenne le message qui lui est destiné. Ainsi, la voix forte selon D'Alessandro (2007) est produite avec une pression subglottique importante, une tension des cordes vocales importante, un débit d'air modéré et une amplitude de voisement élevée. La voix criée est également caractérisée par des réorganisations articulatoires, prosodiques, par des changements de rythme, du contenu spectral, des formants vocaliques mais également de qualité de voix.

L'ensemble de ces modifications, aussi bien dans le but d'augmenter l'intensité de la voix que dans le but de se faire comprendre, engendre des modifications des différents descripteurs acoustiques et prosodiques de la voix. Nous présentons ci-dessous les modifications connues de la littérature concernant la voix criée.

### 5.2.1 Niveau acoustique

Le but premier de la voix criée est l'augmentation de l'intensité vocale. Cette augmentation est principalement générée par l'augmentation de la pression subglottique. Cette modification engendre dans le même temps plusieurs modifications acoustiques.

#### 5.2.1.1 Intensité

**L'intensité moyenne** est sans aucun doute le facteur le plus révélateur de la voix criée. L'ensemble des études sur la voix criée induit par la distance, ou par une consigne d'augmentation de la production vocale, s'accorde à dire que l'intensité moyenne augmente (Alku, 2002; Eriksson et Traunmüller, 1999, 2002; Holmberg et al., 1988; Mooshammer, 2006; Philbeck et Mershon, 2002; Rostolland, 1982a; Watson et Hughes, 2006).

Théoriquement, une augmentation de la distance de communication engendre une diminution de 6 dB par doublement de distance. Plusieurs études ont alors démontré que l'Homme augmente l'intensité vocale dans le but de compenser cette atténuation (cf. CHAPITRE 2). Néanmoins, en fonction des

caractéristiques de l'environnement, la voix du locuteur s'ajuste avec précision et varie de 1,8 à 6,4 dB par doublement de distance (pour des distances de 1 à 8 m) (Pelegrín-García et al., 2011; Zahorik et Kelly, 2007)

Selon Titze (1994), l'intensité vocale peut être contrôlée par trois mécanismes : (1) un ajustement de la pression des poumons (i.e. pression subglottique), (2), un ajustement du larynx et (3) un ajustement du conduit vocal (type *formant tuning*). Par ailleurs, l'augmentation de l'intensité pour des F0 basses est régie par des actions au niveau des cordes vocales (larynx) alors que pour des F0 élevées, l'intensité est régie par le débit d'air (Isshiki, 1964, 1969). Toutefois, d'une façon générale, l'intensité augmente avec la pression subglottique (Lecuit et Demolin, 1998).

« *Le réglage de l'intensité se fait par la combinaison des réglages de la pression sous-glottique et de la géométrie glottique. Il existe en effet une relation quasi-linéaire entre la pression sous-glottique et l'intensité du son mais la fréquence vibratoire a tendance à augmenter dans le même temps. En réalité, l'augmentation des forces d'adduction laryngées est responsable d'une augmentation du temps de contact entre les cordes vocales et s'oppose à cette augmentation de fréquence, ce qui permet la transformation en augmentation de l'intensité vocale. L'augmentation de l'intensité va donc de pair avec la diminution du temps pendant lequel les cordes vocales restent ouvertes* (Giovanni et al., 2002)» (Giovanni et al., 2007).

### 5.2.1.2 Fréquence fondamentale

Le cri nécessite une augmentation de la pression subglottique ainsi que de tendre et de coller entre elles les cordes vocales (Liénard, 1977). Cette réaction a pour effet direct d'augmenter la F0 (Plant et Younger, 2000). Ainsi il n'est pas possible de crier avec une F0 basse (Rostolland, 1982b). En règle générale la F0 augmente de 40 Hz par doublement de distance (selon (Cheyne et al., 2009)). L'augmentation de la F0 est sans doute une conséquence directe de l'augmentation de l'intensité plutôt qu'une stratégie délibérée. Néanmoins, la F0 ne semble pas pour autant être une unique conséquence de l'augmentation de l'intensité, mais une augmentation de la F0 peut également modifier l'intensité en retour (Alku, 2002).

## 5.2.2 Niveau glottique

L'augmentation de la pression subglottique augmente l'amplitude de voisement mais altère également la tension des cordes vocales et réduit la durée de la phase ouverte. Les mêmes phénomènes sont observés pour les voix émotives ou encore les accentuations dans la voix (Cummings et Clements, 1995).

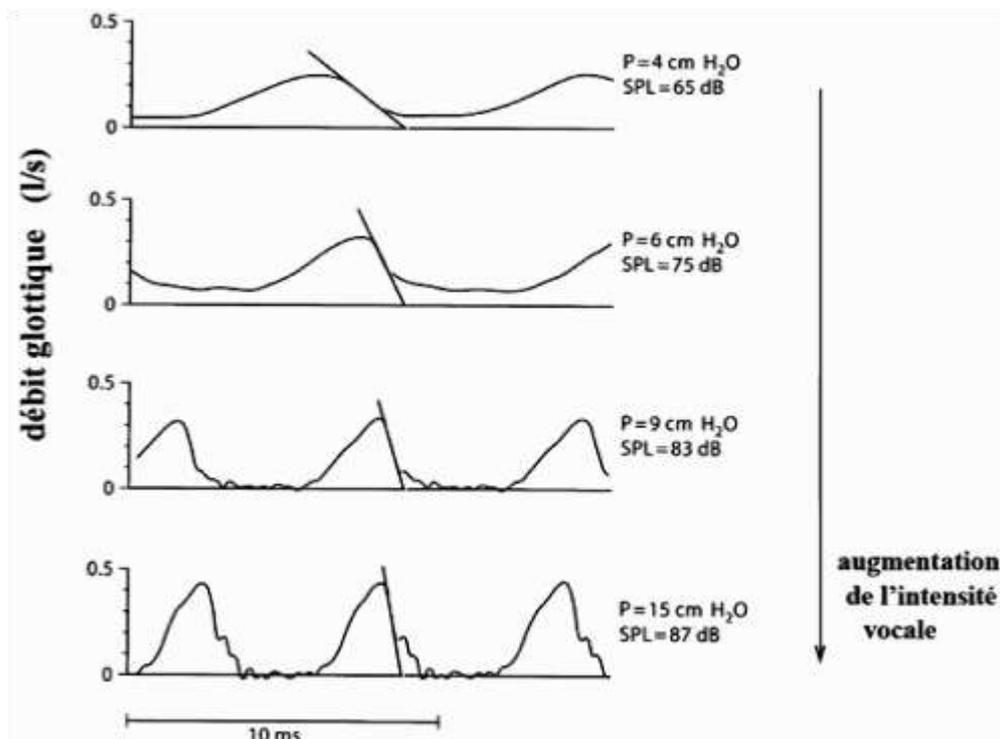


Figure 5-3: Illustration des variations de débit glottique en fonction de l'augmentation de l'intensité vocale. (Gauffin et Sundberg, 1989)

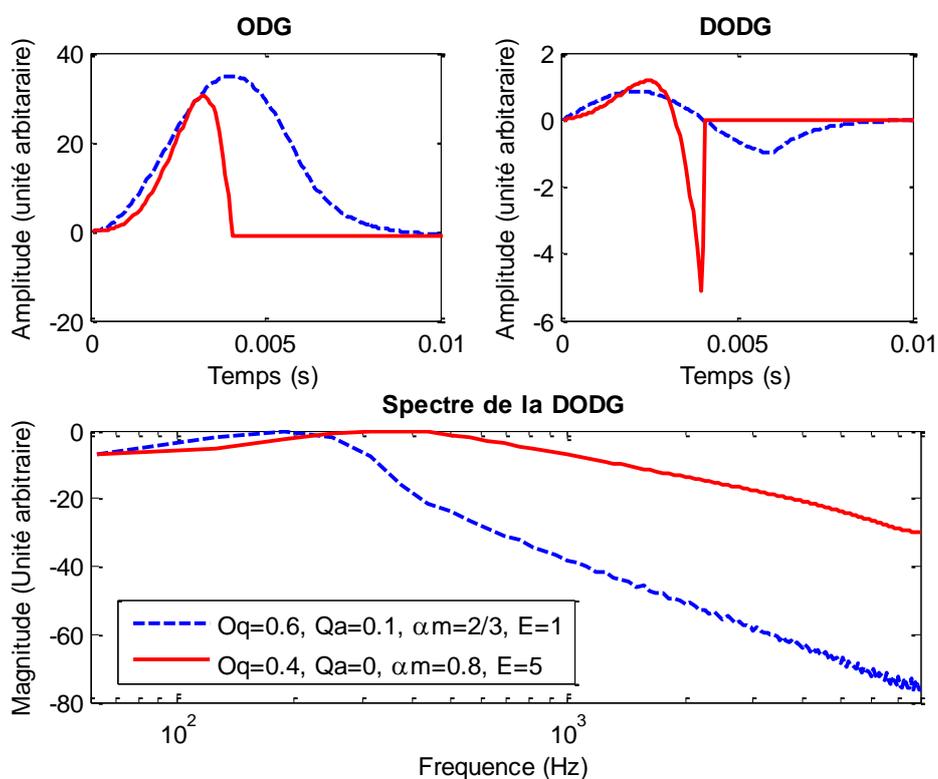


Figure 5-4: Onde de débit glottique (haut gauche), dérivée de l'onde de débit glottique (haut droite) et spectre de deux signaux issu du modèle LF représentant la conséquence de l'augmentation de l'effort vocal. Voix parlée : pointillés, voix criée : trait plein

Les conséquences de l'augmentation de la pression subglottique peuvent être illustrées sur la Figure 5-3. De plus, plusieurs études nous permettent de lister les différentes variations des paramètres du modèle LF en particulier (Henrich, 2001).

- 1- La **vitesse de fermeture**  $E$  (représentée par la tangente sur la Figure 5-3), augmente avec l'intensité (Fant, 1982; Gauffin et Sundberg, 1989; Holmberg et al., 1988; Sundberg et al., 1993).
- 2- L'**amplitude de voisement**  $A_v$ , augmente également avec l'intensité (en voix forte chantée notamment), ce qui s'observe également sur la Figure 5-3 (Sundberg et al., 1993).
- 3- Le **quotient ouvert**  $O_q$ , diminue avec l'intensité jusqu'à atteindre une valeur limite de 0,4 (Nous avons également pu constater dans notre étude une diminution de  $O_q$ , cf. Annexe B). Par la suite, l'augmentation de l'intensité se fait par augmentation de l'amplitude de voisement (Holmberg et al., 1988; Huang et al., 1995; Mooshammer, 2004, 2006, 2010; Sundberg et al., 1993).
- 4- L'**asymétrie**  $\alpha_m$ , est modifiée avec l'intensité et ainsi le **quotient de vitesse**  $S_q = \alpha_m / (1 - \alpha_m)$  (rapport entre la phase ouverte et la phase fermée) est augmenté. L'augmentation de  $\alpha_m$  semble toutefois dépendre du locuteur (Holmberg et al., 1988; Mooshammer, 2004, 2006, 2010).

L'ensemble de ces modifications se traduit par une onde plus abrupte qui contient ainsi plus de hautes fréquences (O'Shaughnessy, 1999 p.67).

L'ensemble de ces modifications de la source glottique va considérablement modifier la forme de l'onde de débit glottique et par conséquent son spectre. Sur la Figure 5-4 sont représentées deux ondes de débit glottique. La première représente les valeurs usuelles de la voix parlée (trait en pointillé) et la deuxième des variations des paramètres en accord avec la littérature. Les valeurs utilisées pour la voix criée sont arbitraires mais évoluent dans le sens de la littérature. Les spectres de ces ondes de débit glottique sont également tracés sur cette figure et normalisés en intensité. On constate ainsi, que les variations de paramètres du modèle LF avec l'effort vocal, engendrent un déplacement du formant glottique vers les hautes fréquences, diminuent la part des basses fréquences et augmentent la part de hautes fréquences. De plus, la pente spectrale, au-delà du formant glottique, est plus faible. Cet exemple illustre parfaitement les différentes variations spectrales observées en voix criée (pente spectrale, plage d'énergie différente sur le spectre).

### 5.2.3 Niveau articulatoire

Lors d'un cri, le **larynx** monte légèrement, ce qui se traduit par un conduit vocal plus court (Acker, 1987; Garnier et al., 2008) et par conséquent par une augmentation de F1, de l'amplitude générale de la voix et altère également la forme du spectre (O'Shaughnessy, 1999). Concernant **les fonctions d'aire**, une étude observe une augmentation de l'aire du conduit vocal à partir de 7 cm au-dessus de la glotte au niveau de l'oropharynx. Pour des conditions extrêmes, l'aire du conduit vocal dans cette zone, peut être jusqu'à doublée (Tom et al., 2001).

Il existe une étude bien connue (Schulman, 1989) qui montre que les **mouvements articulatoires** sont amplifiés en voix criée par rapport à la voix modale. Les mouvements des lèvres en particulier, montrent des différences significatives car elles s'arrondissent plus ou s'étirent plus. Les déplacements des articulateurs sont plus grands mais également plus rapides. La production de voyelles criées entraîne alors une ouverture de la mâchoire plus grande alors qu'aucune variation significative n'est observée pour les consonnes (/s, f, l, n, d, t/ (Geumann, 2001a, 2001b)).

La production de la parole liée à un fort effort vocal s'accompagne de **mouvements posturaux** qui précèdent la production vocale. Ceci semble alors montrer une stratégie globale du comportement vocal visant à améliorer l'efficacité de la communication (Lagier et al., 2009).

Rostolland mentionne qu'il n'est pas possible d'articuler parfaitement et de crier en même temps. Ce qui signifie que les différences d'articulation ont pour but d'accroître l'énergie de la voix au détriment de l'intelligibilité (Rostolland, 1982b)

### 5.2.4 Niveau spectral

Dû aux modifications articulatoires observées dans une voix criée (Geumann, 2001a, 2001b; Schulman, 1989) mais également dû aux modifications liées à la vibration des cordes vocales (notamment la diminution de Oq (Barney et al., 2007)), la position des formants est modifiée (Bond et Moore, 1990; Chládková et al., 2009; Elliott, 2000; Geumann, 2001a; Hansen et Patil, 2007; Harris, 1964; Huber et al., 1999; Junqua, 1993; Liénard et Di Benedetto, 1999; Nanjo et al., 2009; Rostolland, 1982b; Stanton et al., 1988; Traunmüller et Eriksson, 2000).

#### 5.2.4.1 Les formants

Seule une augmentation significative du premier formant est observée. Liénard et Di Benedetto (1999) observent un déplacement du premier formant de 5 Hz/dB, soit un déplacement d'environ 40 Hz dans leurs études. Pour Traunmüller et Eriksson (2000), F1a (formant de la voyelle /a/) varie avec la distance de communication mais pas F3 (ces auteurs n'étudient pas F2, trop dépendant de la voyelle).

Pour une augmentation de la F0 de 100 % F1a varie de 42 % pour les hommes, 71 % pour les femmes, 95 % chez les petits garçons et de 124 % chez les petites filles. Les variations de F3 observées sont toutefois plus corrélées avec l'âge qu'avec la distance de communication. De plus les variations observées sont infimes (pour une augmentation de 100 % de la F0, F3 augmente de 4.1%). Ces auteurs observent ainsi une corrélation positive entre la F0 et la F1 pour l'ensemble des participants (hommes, femmes, garçons, filles).

D'autre part, (Rostolland, 1982b) constate notamment pour les voyelles fermées /i, y, u/ que F1 tend à se rapprocher de la fréquence fondamentale, pour les voyelles semi-fermées /e, ø, o/ la F1 se situe entre F0 et H2 (i.e. la deuxième harmonique de F0), pour les semi-ouverte /ɛ, œ, ɔ/ la F1 se place proche de H2. Toutefois, pour la voyelle ouverte /a/, la variation de F1 n'est pas significative (Rostolland, 1982b). En moyenne, pour les voyelles étudiées, une augmentation de 47 % est observée. L'augmentation différente des formants de chaque voyelle tend à diminuer les écarts phonétiques entre celle-ci et ainsi, favorise la confusion des voyelles en voix criée (Rostolland, 1982b).

Les études mentionnées ci-dessus ne décrivent pas de variations significatives sur F2 et F3. Cependant, certains auteurs relèvent des variations de F2 ou F3. Nanjo et al. (2009) par exemple, constate que F1 et F2 se déplacent vers les hautes fréquences. Plus précisément Bond et Moore (1990) montrent tout d'abord que F1 augmente de 109 Hz en moyenne pour les voyelles /i, u/ entre une voix parlée et une voix criée. En revanche pour les voyelles basses (/a, æ/) F1 tend à diminuer de 46 Hz et surtout F3 diminue de 90 Hz pour toutes les voyelles. Toutefois F2 ne varie que très peu. Pour Chládková et al. (2009), pour une augmentation de F0 d'un facteur 1,37, un déplacement de F1 de 1,125 est observé. Le déplacement de F1 est plus grand pour les femmes que pour les hommes et F2 augmente légèrement.

#### **Remarque :**

On constate que la littérature sur le sujet s'accorde à dire que les positions des formants, et en particulier F1, augmentent avec l'effort vocal. Cependant, les valeurs de déplacement mentionnées par ces auteurs varient beaucoup. Plusieurs explications peuvent alors être données concernant ces différentes valeurs. La première est que les études ne concernent pas les mêmes quantités d'effort vocal. Certains considèrent des voix fortes, d'autres des voix criées ou encore des voix plus intenses. Par ailleurs les protocoles d'enregistrement sont également différents. Dans certains cas il s'agit de communication à distance, dans d'autres cas, il est simplement demandé aux sujets d'augmenter l'effort vocal. Ainsi, des stratégies de communication différentes peuvent également influencer le résultat.

Une seconde explication peut également être donnée. Soulignons que l'estimation de la position des formants sur une voyelle isolée, ou encore sur une voyelle, dans un mot peut également modifier les

valeurs mesurées. En effet, en se basant sur la modification articuloire qui prédit une ouverture plus prononcée et des mouvements plus amples de la bouche lors de la production des voyelles (Schulman, 1989) en voix criée, on peut en déduire que l'évolution du premier formant est différente entre une voix parlée et une voix criée. La plus grande dynamique de l'ouverture de la bouche en voix criée engendre une dynamique de F1 plus grande également. Ainsi, la position à laquelle la mesure du formant est effectuée est également primordiale.

Certains doutes peuvent être soulevés quant à la mesure des formants elle-même. En effet, il est mentionné dans plusieurs études, que la mesure des formants est particulièrement délicate voire même systématiquement biaisée pour des voix qui possèdent des F0 élevées (telles que la voix criée, la voix Lombard, les voix des enfants et des femmes, ou encore la voix chantée) (Atal et Schroeder, 1974; Monsen et Engebretson, 1983; Traunmüller et Eriksson, 1997; Vallabha et Tuller, 2002).

Enfin, la différence entre les études quant aux déplacements ou non de F2 et F3, peut être liée aux types de sujets sélectionnés pour les enregistrements. En effet, Huber et al. (1999) montrent que F1 augmente avec l'intensité de production, mais F2 et F3 ne sont pas modifiés avec l'intensité mais simplement par l'âge et le sexe du locuteur.

Ainsi, toutes ces variantes rendent difficile la comparaison des résultats de chaque étude.

### **Interprétations possibles :**

Comme nous avons pu le voir, plusieurs modifications articuloires, ainsi que les paramètres de la source glottique sont modifiés. Ceci engendre une modification de la position du premier formant en particulier. Plusieurs explications sont reportées (Garnier et al., 2008).

- La diminution de Oq devrait pourtant entraîner une diminution de la fréquence du premier formant, comme le précise Barney et al. (2007) dans une étude réalisée sur un simple tube acoustique. Toutefois, dans certaines circonstances, F1 peut augmenter lorsque Oq diminue (Barney et al., 2007) dans des ordres de grandeur de +40 Hz. De plus, pour une diminution simultanée de Oq et une augmentation de l'aire glottique, F1 augmente de + 160 Hz.
- La montée du larynx peut également être à l'origine de cette modification de la position des formants. L'élévation du larynx entraîne un conduit vocal plus court qui se traduit par une augmentation de la fréquence de tous les formants. Toutefois, Garnier et al. (2008) mentionnent qu'un déplacement de 2 mm du larynx (comme ils l'ont mesuré), engendre uniquement une augmentation de la première résonance de 14 Hz. Ces auteurs n'observent pas de corrélation entre R1 (la première résonance du conduit vocal) et la montée du larynx.

- Le facteur qui semble influencer le plus l'augmentation de F1 lors de l'effort vocal, semble être l'articulation (Schulman, 1989). En effet, Garnier et al. (2008) ne trouvent pas de corrélation significative prouvant l'influence de la montée du larynx ou de la modification de Oq dans le déplacement de R1, mais une forte corrélation avec l'articulation.
- D'un autre point de vue, l'augmentation de F1 peut ne pas être une conséquence directe des différents ajustements faits pour augmenter l'intensité de la voix, mais pourrait être un geste délibéré de la part des locuteurs afin de conserver l'intelligibilité. En effet, la distance F0-F1 est plus révélatrice de la hauteur de la voyelle que F1 ne l'est tout seul (Eriksson et Traunmüller, 2002). Ainsi comme F0 augmente avec l'effort dans le but de conserver la distance F0-F1 lors d'un cri, une augmentation de F1 est nécessaire. Toutefois, cette théorie ne semble valide que pour certaines voyelles (Garnier et al., 2008).
- Enfin une dernière hypothèse se base sur l'efficacité de la parole. Dans le but d'un rendement maximum, la stratégie la plus appropriée est le positionnement du premier formant à l'aplomb d'une harmonique du signal. De cette façon l'intensité de voix résultant est maximisée. Ce phénomène largement observé chez les chanteurs (Joliveau et al., 2004a; Sundberg, 1977; Wolfe et Garnier, 2009) est également observé en voix criée (Garnier et al., 2008). De la même manière que précédemment, l'augmentation de F1 ne semble pas être une conséquence de l'effort vocal mais plutôt une stratégie pour augmenter l'intensité de la voix.
- Rostolland mentionne qu'il n'est pas possible d'articuler et de crier en même temps. Les formants se déplacent vers les zones les plus sensibles de l'audition au détriment de l'intelligibilité (Rostolland, 1982b). En effet, en considérant la sensibilité de l'oreille humaine, le déplacement des formants vers les hautes fréquences tends à les placer dans la zone la plus sensible de l'audition (cf. Figure 5-5) et, par ce fait, a pour conséquence d'augmenter la sonie de la voix (sensation d'intensité perçue)

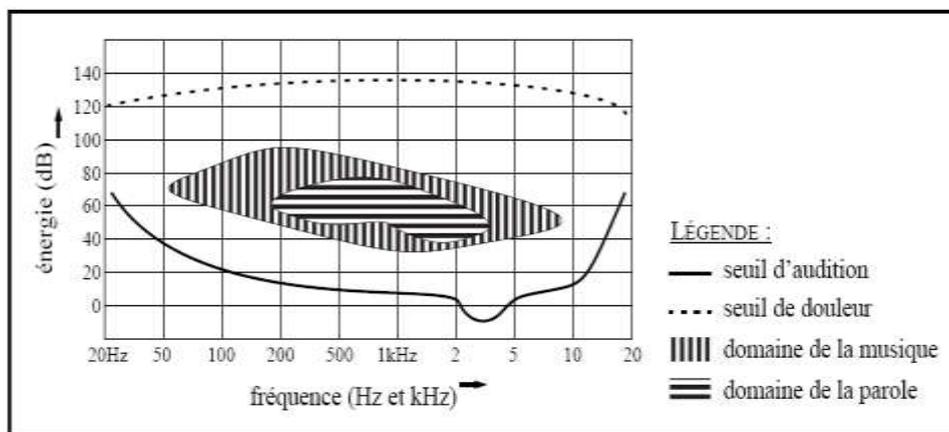


Figure 5-5 : Le champ auditif humain (Zwicker et Feldtkeller, 1981)

#### 5.2.4.2 Pente spectrale

D'une façon générale le spectre de la source glottique possède une pente spectrale de -12 dB/octave. Toutefois quand la voix devient forcée, on constate que les évolutions de la forme d'onde glottique deviennent plus abruptes. On observe notamment une forte vitesse de fermeture et donc une valeur de E plus élevée. Cette fermeture abrupte est liée d'une part à la pression subglottique élevée lors de l'ouverture et d'une forte succion due à l'effet Bernoulli lors de la fermeture des cordes vocales. Ceci engendre alors une pente spectrale de la source glottique différente que celle observée en voix parlée. En effet, en voix criée la pente spectrale possède généralement une pente de -9 dB/octave ((Hansen et Patil, 2007) mentionnée dans (Pickett, 1998) et (Zhang et Hansen, 2011)). Ces observations sont également faites par Zhang et Hansen (2007) mais qui, eux ne constatent pas d'aussi fortes variations. En effet, ces auteurs n'observent qu'une différence de 0,78 dB/octave entre la voix modale et la voix criée (i.e. voix modale : 8,29 dB/octave, voix criée : -7,51 dB/octave). Ces observations sont en accord avec (Liénard et Di Benedetto, 1999) qui montrent une variation plus forte de l'amplitude des formants élevés. Ainsi, quand l'intensité augmente de 10 dB, A1 (amplitude du premier formant), A2 et A3 augmentent respectivement de 11, 12,4, et de 13 dB selon (Liénard et Di Benedetto, 1999).

#### 5.2.4.3 Aspect du spectre

Des auteurs se sont également intéressés à la forme générale du spectre de la voix criée. Par les variations de Oq et du coefficient d'asymétrie, la forme du spectre du signal de la parole se voit également modifié. En effet, les variations sur ces paramètres en voix criée engendrent un déplacement du formant glottique vers les hautes fréquences. Sur le spectre de la parole, ceci se traduit par une chute d'énergie au niveau des 4-5 premières harmoniques. De plus, une perte d'intensité dans la région de 600 Hz à 1 kHz est d'après certains auteurs, prévisible. En effet, une chute dans cette région peut être causée par une absorption de ces sons, par la partie subglottique (d'après (Pickett, 1998) p.62).

En voix criée, l'énergie ne se concentre plus uniquement autour de F0 et F1, mais s'étale jusqu'à des fréquences de 4-5 kHz (Rostolland, 1982b). Il mentionne d'ailleurs que ce phénomène se traduit non pas par une accentuation des fréquences moyennes et hautes, mais par une chute des basses fréquences. On constate ainsi une accentuation de l'énergie dans les parties les plus audibles par l'homme (Rostolland, 1982a).

Bien que ces modifications de la répartition de l'énergie puissent être attribuées au déplacement du larynx, à l'affinement des parois du conduit vocal et à une ouverture buccale plus intense, la piste de la source glottique est privilégiée. En effet, Rostolland (1982b) fait l'analogie avec les instruments de musique. Il mentionne que pour des instruments à vent, comme la clarinette, une augmentation de

l'intensité du jeu, engendre incontestablement une intensification des harmoniques supérieures, alors que les caractéristiques de l'instrument (i.e. du conduit vocal) n'ont pas changées.

Il semble également que la modification de la forme du spectre soit réalisée de manière volontaire, afin d'augmenter l'efficacité de la production vocale. Chez les chanteurs professionnels ténors, une modification de la position de F1 et F2 est observée afin de leur permettre d'accroître l'intensité de la voix. Les chanteurs peuvent également ajuster la position du pharynx, afin d'augmenter l'énergie dans les bandes de fréquences autour de F3 et F5 (Sundberg, 1973). Cette prééminence dans les hautes fréquences est appelée formant du chanteur (*singer's formant*). Ce concept est également observé dans le cas des voix criées non chantées, comme le formant du locuteur (*speaker's formant*) (Bele, 2006; Nawka et al., 1997) ou encore dans le cas des voix d'acteurs (*actor's formant*) (Leino, 1993). De toute évidence, ce formant quel qu'il soit, se présente essentiellement pour des professionnelles de la voix, qui arrivent à ajuster certains articulateurs permettant d'accroître l'énergie dans cette bande de fréquences. L'augmentation de la bande d'énergie de 2-4 kHz (qui correspond au formant du chanteur), peut être expliquée par la mise en vibration des bandes ventriculaires (Bailly et Henrich, 2010).

**Remarque :** Les modifications de l'aspect du spectre et tout particulièrement celles de la pente spectrale peuvent éventuellement être une stratégie visant à compenser l'atténuation des hautes fréquences engendrée par la propagation de l'onde (cf. § 2.1.2) et ce afin d'assurer une meilleure intelligibilité du message.

### 5.2.5 Prosodie

Outre l'aspect acoustique des variations, on retrouve également dans la littérature des études qui relèvent des variations prosodiques entre la voix modale et la voix criée.

#### 5.2.5.1 Durée

En voix criée, les énoncés ont tendance à être rallongés. Il s'agit d'une manière générale d'une augmentation de la voyelle et une diminution des consonnes (Bonnot et Chevrie-Muller, 1991; Geumann, 2001b; Rostolland, 1982a; Traunmüller et Eriksson, 2000).

Chez Traunmüller et Eriksson (2000) les voyelles augmentent de 16 % pour les adultes et seulement de 7 % chez les enfants, alors que les consonnes augmentent de 11 % chez les adultes, mais diminuent de -14% chez les enfants. Dans certains cas, aucune variation de la durée des consonnes n'est observée mais un allongement de la durée des voyelles reste présent (Hansen et Patil, 2007).

On trouve également chez Rostolland (1982a) une augmentation de la durée des voyelles mais dans des proportions plus élevée. En effet, cet auteur relève une augmentation de 33% pour des voyelles non finales et une augmentation de 67% pour des voyelles en position finales. Contrairement, à l'étude précédente une diminution de 20% de la durée des consonnes en voix criée est observée.

Fónagy et Fónagy (1966) montrent que la durée des consonnes diminue en voix criée, sauf pour les nasales et les liquides. Ces auteurs mentionnent d'ailleurs, que la modification de la durée est en relation directe avec la portée en distance de chaque consonne. En effet, chaque consonne n'a pas la même portée en distance du fait de sa nature. Ainsi, pour les consonnes occlusives, fricatives et liquides, Fónagy et Fónagy (1966) constatent que plus la consonne porte loin, moins elle subit les effets dus à la distance, et donc plus elle est compréhensible. De ce fait la communication de ces consonnes ne demande pas beaucoup d'effort de la part du locuteur pour la faire comprendre à un auditeur placé à distance.

#### *5.2.5.2 Intensité*

L'intensité présente un aspect dynamique non négligeable. Il existe, en effet, une modification significative du rapport d'intensité entre consonnes et voyelles. Fairbanks (1957), mesure pour des mots de type /sVs/ (exemple /sas/) des voix produites avec peu d'effort, que le rapport d'intensité consonne/voyelle est en moyenne de -7,2 dB, alors que pour des voix produites avec un niveau d'effort important, le rapport est de -13,8 dB. Soit une augmentation de l'intensité des voyelles d'environ 7 dB de plus que celle des consonnes, pour une augmentation de la voyelle de 18 dB. L'étude de Traunmüller et Eriksson (2000) va également dans ce sens. L'intensité des zones voisées augmente de 4,6 dB par doublement de distance, et l'intensité des /s/ n'augmente que de 2,2 dB SPL par doublement de distance (Traunmüller et Eriksson, 2000).

#### *5.2.5.3 Fréquence fondamentale*

La dynamique de F0 change également avec l'effort vocal (Shriberg et al., 1996; Watson et Hughes, 2006). Rostolland (1982b) observe une plus grande excursion du contour de F0 en voix criée. Cet auteur observe notamment des contours de F0 convexes sur les voyelles, des creux de F0 sur les consonnes ainsi qu'une chute brutale en fin d'énoncé. Il constate également dans le cas où la dernière voyelle est accentuée, que la F0 de celle-ci montre une forme de plateau. Il évoque ici l'hypothèse de la saturation de F0. Il mentionne que cet effet apparaît car les locuteurs ne sont pas capables d'augmenter plus la F0. De ce fait, un phénomène de saturation apparaît, alors que sur les autres voyelles du mot, la F0 arbore une forme de bosse. D'autres études dédiées à la classification de voix criée, constatent que pour un énoncé donné, la distribution des valeurs de F0 est plus grande en voix criée qu'en voix parlée (Nanjo et al., 2009; Zhang et Hansen, 2007). Toutefois, une étude retraçant les

différents travaux sur ce sujet montre que l'écart-type normalisé (écart-type divisé par la valeurs moyenne de  $F_0$  :  $\sigma F_0/F_0\text{mean}$ ) entre une voix parlée et une voix criée, n'est significativement plus élevée que pour des tâches de lecture et non pas dans la parole spontanée (Jessen et al., 2005).

La production d'une voix criée influe également sur la vibration des cordes vocales : les bandes ventriculaires peuvent entrer en vibration (périodique ou non) et influencer l'oscillation des cordes vocales (voir inhiber) la vibration celles-ci (Bailly et Henrich, 2010; Bailly, 2009).

### 5.2.6 Qualité de voix

On constate également, pour des productions à fort effort vocal, que le *jitter* (fluctuation involontaire de la  $F_0$ ) et le *shimmer* (fluctuation involontaire de l'amplitude) diminue en voix spontanée (Huang et al., 1995). D'autre part, les variations du quotient de contact (CQP : contact quotient perturbation) diminue avec l'effort vocal, ce qui signifie que d'un cycle de vibration à l'autre, les valeurs du quotient de contact (et donc du quotient ouvert) sont plus constants en voix criée.(Huang et al., 1995). Notons également que la part de bruit dans le signal glottique semble diminuer en voix criée (D'Alessandro et Doval, 1998; Richard et D'Alessandro, 1996).

## 5.3 Identification du locuteur et intelligibilité

---

Nous avons constaté à travers l'état de l'art présenté dans ce chapitre, que certains des paramètres de la voix varient fortement. On peut alors s'interroger, au regard de ces fortes modifications, sur l'intelligibilité de telles voix. Nous présentons ci-dessous un bref état de l'art à ce sujet. Par ailleurs, l'identification du locuteur est également un point important dans la communication verbale. Nous mentionnons également ce point dans cette section.

### 5.3.1 Intelligibilité de la voix criée

Rostolland a réalisé une suite d'études visant à quantifier l'aspect acoustique (Rostolland, 1982a) et phonétique de la voix criée (Rostolland, 1982b). Il achève ce travail par une étude sur l'intelligibilité de ces voix (Rostolland, 1985). A partir de mots bi-syllabiques, il mesure en moyenne un taux d'intelligibilité des voix parlées de (98%, 100% et 100 % pour des niveaux sonores de présentation de 55 dB SPL, 80 dB SPL et 100 dB SPL) alors que pour les voix criées les intelligibilités sont de 88%, 88% et 95 %. Cette étude mesure également l'influence du locuteur (français ou étranger), ainsi que l'influence du bruit de fond sur l'intelligibilité de ces voix à différents niveaux de présentation. Ici, nous ne rentrons pas plus dans le détail de cette étude. Toutefois, d'une manière générale,

l'intelligibilité des voix criées décroît pour des locuteurs étrangers et les voix criées sont en générales plus intelligibles pour un niveau de bruit donné que les voix parlées.

Pickett (1956), quant à lui, mesure l'intelligibilité de mots suivant différentes forces vocales (voix très basses, jusqu'à des voix très fortement criées) en faisant varier le rapport signal à bruit (+6 dB, 0 dB et -6 dB). Il constate que l'intelligibilité des voix modérément basses, jusqu'à des voix très fortes, ne varient pas plus de 5 %, mais que les voix très basses ou fortement criées quant à elles subissent une chute drastique d'intelligibilité, qui est linéaire avec l'intensité de la voix. Il constate que quand un locuteur parle de plus en plus fort d'une voix basse à une voix forte (55 dB à 1 m jusqu'à 78 dB à 1 m) pour un même RSB (apport signal sur bruit), l'intelligibilité ne varie pas plus de 5% (soit 90 % d'intelligibilité). A l'inverse, entre une voix forte et une voix criée, cet auteur constate une perte d'intelligibilité de plus de 30 % par rapport à la voix criée maximale. Pour une voix fortement criée, l'intelligibilité pour un RSB de +6 dB est d'environ 70 %. Alors que pour les voix très faibles elle est d'environ 40%.

Ces différences peuvent s'expliquer de différentes manières. D'une part, l'intelligibilité d'un énoncé est plus forte pour un rapport d'intensité consonne/voyelle élevé (Kennedy, 1998). Nous avons constaté que dans la littérature il est mentionné que ce rapport diminue. C'est-à-dire que l'intensité des voyelles augmente plus que celle des consonnes. Ainsi, d'après (Kennedy, 1998) pour une voix criée ce phénomène engendre une perte d'intelligibilité. De plus, le déplacement des formants réduit les écarts entre les formants des différentes voyelles, ce qui favorise la confusion et donc réduit l'intelligibilité (Rostolland, 1982b).

### 5.3.2 Intelligibilité de la voix chuchotée

Pour les voix chuchotées, on relève des taux d'intelligibilité de 65 % (82% en voix parlée) pour des voyelles isolées (Kallail et Emanuel, 1985). Tartter (1991) mesure une intelligibilité des voyelles placées en position centrale d'une structure monosyllabique (/hVd/, exemple /had/), un taux de 82% (comparé à 92% en voix parlée). Chez Eklund et Traunmüller (1996) pour des voyelles isolées, il mesure un taux de 88% (95 % en, voix parlée). Ainsi, d'une manière générale, les voyelles sont correctement identifiées. Les consonnes, en revanche, sont identifiées à 64 % (Tartter, 1989). Toutefois, la majorité des erreurs observées (58 %) sont attribuées à la mauvaise décision entre consonnes voisées et non voisées.

Ainsi, en voix chuchotée, la baisse d'intelligibilité semble principalement due à l'absence de F0 qui rend difficile la différenciation des consonnes voisées et non voisées.

### 5.3.3 Reconnaissance du locuteur

Brungart, Scott, et Simpson (2001), test l'habilité des auditeurs, à identifier la voix d'un locuteur sous différents modes de phonation et pour des voix d'hommes et de femmes. Pour ce faire il réalise 2 types de tests. Dans la première version, il réalise une série d'entraînements puis mène le test d'identification pour chaque mode de phonation (conversationnel, chuchoté, et crié). Quand l'entraînement est réalisé sur le même mode de phonation que le test les résultats de l'identification sont de 90 % (Hommes) et 86 % (Femmes) pour les voix parlées, de 87 % (Hommes) et de 95 % (Femmes) pour les voix criées et de 69 % (Hommes) et de 69% (Femmes) pour les voix chuchotées. Cependant, quand les tests d'identification des locuteurs sont effectués sur des stimuli en voix conversationnelle uniquement, le taux est de 70 % (Hommes) et 43 % (Femmes) en voix criée, et 51 % (Hommes) et 37 % (Femmes) en voix chuchotée.

Dans une étude qui vise à juger de la capacité des auditeurs à identifier la voix d'une personne familière en voix criée et masquée, Blatchford et Foulkes (2006) mesure un taux de reconnaissance de 52 % pour une phrase de 2 mots (« *get him !* ») et de 81 % pour une phrase longue (« *face down on the ground and hands behind your back now !* »).

Une étude traite de l'identification du sexe du locuteur. En voix chuchotée sur la base des voyelles /i/ et /a/ Schwartz (1968) n'observe aucune erreur sur 80 classifications pour le /i/ et seulement 4 pour le /a/. Pour les voyelles /i, ε, æ, a, o, u / observe un taux de reconnaissance des sexes de 96 % pour des voyelles parlées et un taux de 76% pour des voyelles chuchotées.

## 5.4 Conclusion du chapitre 5

---

Nous avons, dans ce chapitre, présenté un état de l'art concernant les modifications des paramètres de la voix chuchotée et de la voix criée, par rapport à la voix modale. Un récapitulatif des principales variations est présenté dans le Tableau 5-1. Pour la voix chuchotée, les modifications sont bien connues et sont similaires d'une étude à l'autre. Mais le cas de la voix criée est plus complexe. On observe une grande variation des paramètres mentionnés en fonction des études. Nous pensons que ceci est normal, étant donné les différences de protocoles utilisés pour la production de la voix criée dans ces études, ainsi que les différents niveaux possible de la voix criée. D'ailleurs en voix chuchotée, l'absence de vibrations des cordes vocales écarte déjà une grande source de variabilité.

Au chapitre suivant, nous nous concentrons sur la caractérisation de la voix criée qui constitue le cœur de notre étude.

Tableau 5-1: Récapitulatif des principales variations observées entre les voix modales et les voix criées et chuchotées

Paramètres	Normal Vs Criée	Normal Vs Chuchotée
Valeur de F0	▲	∅
Dynamique de F0 ( $\delta F0$ )	▲	∅
Valeur de I	▲	▽
Dynamique de I ( $\delta I$ )	▲	▽
Valeur du quotient ouvert (Oq)	▽	∅
Vitesse de fermeture (E)	▲	∅
Coefficient d'asymétrie ( $\alpha m$ )	▲	∅
Pente spectrale	▽	▽
Durée totale	▲	▲
Durée des voyelles	▲	▽
Durée des consonnes	▽	▽

Légende : ▲ : augmentation, ▽ : diminution, ∅ : inexistant



## **PARTIE II :**

# **CARACTÉRISATION DES VOIX CRIÉES**



# Élaboration des bases de données

---

*“Imagination is more important than knowledge”*  
— Albert Einstein

Dans le but de transformer une voix modale en une voix criée reflétant la distance, il est indispensable d'établir des règles de transformation liées à la distance de communication. Or l'établissement de ces règles requiert une profonde connaissance sur les paramètres pertinents reflétant la distance de communication. Faute d'avoir suffisamment de descriptions précises sur ces paramètres dans la littérature, nous sommes amenés à effectuer des analyses approfondies afin de caractériser les voix criées. Plusieurs corpus ont alors été enregistrés dans ce but, plus précisément pour refléter une communication verbale plausible à distance entre 2 personnes. Afin de s'assurer que les locuteurs produisent une voix avec suffisamment d'effort vocal, des grandes distances de communication verbales ont été choisies : jusqu'à 100 m. Bien évidemment pour atteindre de telles distances, des enregistrements en extérieur ont été indispensables. Cependant, devant la complexité de réalisation de tels enregistrements, plusieurs corpus ont été nécessaires. En effet, les enregistrements en extérieur engendrent beaucoup de difficultés : bruit ambiant trop important, lieux d'enregistrement trop éloignés, difficultés d'organisation pour le transport du matériel. Ainsi, chaque corpus enregistré avait un but précis et a été mis en œuvre avec précautions dans cet objectif. Malgré tout, les analyses ou encore des réflexions ultérieures nous amènent à penser que les corpus que nous détenons peuvent encore être perfectionnés.

Le premier corpus présenté ici a été enregistré par 3 locuteurs jusqu'à des distances de 100 m en extérieur. Le nombre de locuteurs étant insuffisant pour être représentatif, un deuxième corpus comprenant plus de locuteurs a ensuite été réalisé. Ce corpus tient également compte des difficultés et problèmes rencontrés sur l'enregistrement du premier corpus et a été réalisé de nuit en pleine campagne avec 6 locuteurs. Malgré nos efforts, nous avons estimé qu'un protocole de simulation de communication à distance est indispensable pour réaliser des enregistrements de bonnes qualités et reproductibles. Un troisième corpus a été réalisé dans cet objectif. Nous avons mis en place une simulation en chambre d'audiométrie permettant de simuler des distances de communication de 60 m

et 120 m. Ce protocole s'est révélé très efficace pour réaliser un enregistrement avec un grand nombre de locuteurs dans de bonnes conditions. Le dernier corpus, qui reprend le protocole d'enregistrement précédent, a quant à lui été réalisé dans l'objectif d'enrichir notre base de données d'un point de vue phonétique. Ce corpus, est composé de structures phonétiques de base de type (CV, CVC, VCV et CVCV), plutôt que des phrases courtes utilisées jusqu'à présent.

Nous tenons à préciser, que l'ensemble des enregistrements ont été réalisés à partir d'une tâche de lecture de listes de phrases, de voyelles, de nombres ou encore de logatomes. Il est reconnu que la production vocale, dans le cas de la lecture, est différente de la parole spontanée (Lieberman et al., 1985). En effet, la parole lue est en générale plus lente et plus articulée. Étant donné que notre étude se concentre sur l'extraction des faits pertinents pour la perception de l'effort vocal et de la distance d'un locuteur, nous estimons que ces variantes n'influent pas ou peu sur la perception de l'effort vocal directement. Il restera toujours possible, à partir d'une voix criée lue ou spontanée de percevoir l'effort vocal. Par ailleurs, l'effet de liste, qui dans certaines études peut être un inconvénient, est dans notre étude plutôt un avantage. En effet, l'effet de liste a pour conséquence d'homogénéiser la production vocale, ce qui pour une tâche de modélisation est un plus.

## **6.1 Protocole d'enregistrement en champ libre : DB1**

---

Le premier corpus de voix enregistré est nommé « corpus DB1 », l'objectif étant d'enregistrer des voix dans le cadre d'une communication verbale à distance de plus en plus grande.

### **6.1.1 Protocole d'enregistrement**

Afin de réaliser des analyses sur les caractéristiques des voix pour des distances croissantes, nous avons choisi d'utiliser dans un premier temps des phrases courtes (listées dans le Tableau 6-1). Ces phrases ont été choisies car elles correspondent à une structure de message plausible lors de communication militaire, c'est-à-dire à des phrases courtes évoquant soit une question, soit un ordre. Nous avons également choisi 5 voyelles ([a], [e], [i], [o], [y], [u]) représentatives du triangle vocalique. L'enregistrement de voyelles isolées permet de réaliser des analyses sur la position des formants de manière plus aisée que sur des phrases. En effet, en fonction de la position de la voyelle dans une phrase, plusieurs inconvénients, comme les phénomènes de coarticulation, peuvent intervenir et biaiser les mesures.

Cinq distances de communication ont été utilisées pour réaliser ce corpus: 5 m, 25 m, 50 m, 75 m et 100 m. Le choix de ces distances a été régi par la volonté d'enregistrer des voix pour des écarts de distance constants. Ainsi, 25 m séparent chaque distance de communication (hormis pour le premier écart de distance entre 5 m et 25 m). Le choix de 5 m a été choisi en temps que référence, c'est-à-dire en temps que distance de communication standard. Nous avons volontairement choisi cette distance supérieure à la distance standard utilisée de 1 mètre pour des raisons de cohérence avec le reste du corpus. En effet, si l'environnement dans lequel se déroule les enregistrements est trop calme, à une distance de 1 mètre il est probable que la voix utilisée soit trop détendue voir soufflée par moment. D'autre part, le but de ce corpus n'est pas de juger de la différence entre une voix modale et une voix pour une distance donnée, mais bien d'étudier les variations des paramètres vocaux en fonction de la distance et d'en déduire les éléments pertinents. En ce sens, le choix d'une distance de référence de 5 m n'est pas dérangeant.

*Tableau 6-1 : Liste des 5 phrases utilisées dans les différents corpus*

<b>Nomenclature</b>	<b>Phrase</b>	<b>Transcription phonétique</b>
Ph. 1	Lève le bras.	/lɛv lə br a/
Ph. 2	Tu me vois?	/ty mə vwa/
Ph. 3	Tu m'entends?	/ty mɑ̃tɑ̃/
Ph. 4	Bouge la tête.	/buʒ la tɛt/
Ph. 5	Viens vers moi.	/viɔ̃ vɛr mwa/

Dans le but de forcer le locuteur à produire un niveau d'effort adéquat pour chaque distance de communication, une situation de communication entre un locuteur et un interlocuteur a été mise en place. En effet, dans le cas contraire, le locuteur n'aura aucun retour sur l'efficacité de la voix qu'il utilise, ce qui laisse un degré de liberté sur l'effort vocal trop important. Ainsi, le locuteur peut crier plus fort qu'il n'est nécessaire pour une distance donnée ou encore, ne pas faire l'effort d'articulation nécessaire à la bonne compréhension du message. De plus, Garnier (2007) montre dans le cas de la parole Lombard, que la variabilité de certains paramètres (intensité, F0, durée, mouvements articulatoires, ...) est différente entre une situation de production vocale isolée (sans interlocuteur) et une situation d'interaction avec un interlocuteur. Ces différences montrent que dans le cas de la parole Lombard, la production vocale est motivée par la recherche de l'intelligibilité. La présence d'un interlocuteur dans notre cas semble également primordiale.

Pour nos enregistrements, le locuteur était situé à une position fixe. L'auditeur quant à lui s'est placé successivement aux 5 distances de communication. Le locuteur prononce alors les 5 phrases et les 5 voyelles de manière à se faire comprendre par son interlocuteur. Afin d'assurer un retour de

compréhension, ce dernier exécutait les actions prononcées par le locuteur (pour Ph.1, Ph.4 et Ph.5), ou répondait aux questions (Ph.2 et Ph.3), dans le cas où l'effort vocal produit par le locuteur lui permettait de comprendre la phrase. Pour les voyelles, l'interlocuteur avait pour consigne de répéter ces dernières afin qu'une tierce personne puisse juger de la bonne compréhension. Dans le cas contraire, le locuteur devait répéter le message (phrases ou voyelles) jusqu'à ce qu'il soit compris par l'auditeur.

Les enregistrements ont été effectués pour 3 locuteurs masculins différents. En première approche, aucune locutrice n'a été choisie. En effet, ayant conscience de la difficulté de l'analyse des voix féminines (même en voix modale) et conscience de la difficulté a priori plus grande de transformation de voix féminines, nous nous sommes contentés, pour ce premier corpus, uniquement de voix masculines. Le nombre de 3 personnes, n'est pas le fruit d'une réflexion ou d'un choix mais plutôt d'une restriction liée à des problèmes d'organisation indépendamment de toute volonté.

Les enregistrements ont été effectués en champ libre sur une route en macadam bordée par des champs dont les alentours étaient dégagés (cf. Figure 6-1). Le choix de cet emplacement a été fait dans le but de diminuer les effets liés à la réverbération de l'onde sonore qui peuvent être gênants du point de vue des analyses ultérieures. La voix des locuteurs est enregistrée simultanément par deux microphones : l'un situé à environ 1 mètre de la bouche du locuteur et l'autre près de son interlocuteur (i.e. à chaque distance de communication). Cette configuration permet à la fois d'enregistrer la source (la voix du locuteur), et ce qu'entend l'auditeur. Les enregistrements ont été effectués par deux microphones ½ pouce B&K (type 4189), un conditionneur B&K (type 5935L) et un enregistreur numérique portable (M-Audio, Microtrack II) à une fréquence d'échantillonnage de 24 kHz. Ces enregistrements ont été, par la suite, ré-échantillonnés à 16 kHz.



Figure 6-1 : Photos du protocole d'enregistrement

### 6.1.2 Remarques

Suite à l'enregistrement de ce corpus, nous avons pu observer plusieurs problèmes. Le premier de ces problèmes concerne le bruit ambiant. DB1 est pollué par des bruits de fond assez conséquents, qui sont susceptibles de perturber les mesures qui suivront. De plus, le bruit dû au passage d'avions, de voitures, au souffle du vent ou encore à de cris d'animaux étant par nature très changeant, la production vocale des locuteurs pourrait également être influencée par ces phénomènes extérieurs. On pourrait alors observer des efforts non constants tout le long de l'exercice et en rapport avec le bruit de fond, ce qui peut alors s'assimiler à de l'effet Lombard. Un deuxième problème rencontré lors de l'enregistrement de DB1 a été le contrôle de l'effort vocal nécessaire à la compréhension. En effet, dans DB1, les phrases sont connues par avance et donc les auditeurs sont susceptibles de comprendre les phrases prononcées, alors que l'ensemble de la phrase peut ne pas être parfaitement intelligible. Ceci peut engendrer des efforts non constants sur l'ensemble du corpus ou de l'énoncé.

## 6.2 Protocole d'enregistrement en champ libre version améliorée : DB2

---

Un deuxième corpus de voix (DB2) a été enregistré afin de palier certains inconvénients rencontrés lors des enregistrements du premier corpus (DB1). En effet, DB1 a révélé certains problèmes liés à l'enregistrement de communication à distance en environnement extérieur.

L'inconvénient majeur du premier corpus est le bruit de fond. La solution envisagée pour palier ce problème a été, d'une part de choisir un environnement plus calme, et d'autre part de rapprocher le microphone de la bouche des locuteurs afin d'augmenter le rapport signal sur bruit. Pour ce faire, une monture de fixation, dédiée au microphone B&K permettant d'inclure une boule anti-vent, a été conçue (cf. Figure 6-2). De plus, les enregistrements de DB2 ont été réalisés de nuit, dans une prairie au milieu des champs, loin des villes et des villages. De cette manière, le niveau du bruit de fond dans les enregistrements a pu être diminué. De plus, des enregistrements réalisés dans une prairie permettent également de diminuer les effets liés à la réverbération au sol. En effet, sur un sol composé d'herbes, il y a moins de réflexion que sur un sol en macadam ce qui a pour effet de réduire la quantité d'énergie totale arrivant jusqu'à l'auditeur.

Le deuxième problème lié à DB1 est le problème de la connaissance *a priori* des énoncés par l'auditeur. Une première solution serait de changer les phrases d'un locuteur ou d'un auditeur à l'autre. Or cette solution n'est pas envisageable étant donné que nous souhaitons réaliser une étude

systematique sur l'ensemble des phrases des corpus. Nous devons donc conserver les 5 phrases précédemment établies. Afin de pallier ce problème, nous avons choisi d'enregistrer les phrases dans un ordre aléatoire et d'interposer entre les phrases des séries de nombres aléatoires compris entre 0 et 100. L'avantage des nombres est qu'ils sont peu redondants en termes d'informations linguistiques. Chaque phonème doit être parfaitement entendu afin de comprendre le nombre qui a été prononcé. Par le biais des nombres, nous espérons forcer les locuteurs à conserver une stratégie et un niveau d'effort constant tout au long des enregistrements.

### 6.2.1 Protocole d'enregistrement

Ce corpus contient les mêmes phrases que le corpus précédent. Les distances d'enregistrement ont été choisies telles que chaque distance soit le double de la précédente. En effet, les distances utilisées lors de DB1 ont montré que les écarts entre 50 m, 75 m et 100 m étaient faibles ; d'après la loi du carré inverse les pertes d'intensité sont respectivement (par rapport à 1 mètre) de -33,9 dB, - 37,4 dB et - 39,9 dB. Pour maintenir une diminution de l'intensité de 6 dB entre chaque distance les 5 distances de communication choisies sont : 2 m (référence), 12,5 m, 25 m, 50 m et 100 m.

De la même manière que précédemment, nous avons mis en place une situation de communication entre un locuteur et un auditeur. Pour l'enregistrement de ce deuxième corpus, chaque locuteur disposait d'une liste comprenant des nombres aléatoires, ainsi que 5 phrases. Les nombres et les phrases sont classés de manière aléatoire pour chacune des listes utilisées. Les nombres sont également différents d'une liste à l'autre. Ces listes débutaient par 5 nombres puis une phrase suivie de 2 nombres, et ainsi de suite pour les 5 phrases. Dans le but d'obtenir un niveau d'effort vocal adéquat à chaque distance, il a été demandé au locuteur de prononcer le premier nombre de la liste moins fort que nécessaire. Par la suite, le locuteur devait augmenter progressivement l'effort vocal jusqu'à ce que le nombre soit compris par l'auditeur. De ce fait, le locuteur régule son effort vocal à un niveau adéquat à la distance de communication ; cette procédure évite que le locuteur utilise un niveau d'effort démesuré par rapport à l'éloignement de l'auditeur.

Afin d'assurer la bonne compréhension des messages par l'auditeur, un système de communication textuel via deux ordinateurs a été mis en place. Nous avons choisi cette solution du retour de compréhension par écrit afin de ne pas fatiguer la voix de certains auditeurs qui, par la suite, devenaient à leur tour locuteurs. De plus, le retour par informatique plutôt que le retour verbal, permet de pallier une éventuelle influence de la prononciation de l'auditeur sur celle du locuteur. Lors de la prononciation de nombres par le locuteur, l'auditeur placé à distance avait pour consigne de taper sur un clavier d'ordinateur le nombre qu'il avait entendu. Via une liaison RS323, un deuxième ordinateur placé devant le locuteur, affichait alors le chiffre entendu et tapé par l'auditeur (cf. Figure 6-3). Dans le cas où le nombre aurait été mal compris, les locuteurs avaient pour consigne de répéter ce nombre

jusqu'à ce qu'il soit compris. Lors de la prononciation des phrases, comme pour les enregistrements de DB1, l'auditeur devait répondre aux questions via l'ordinateur ou exécuter l'action qui lui était demandée.

Ces enregistrements ont été réalisés par 6 locuteurs uniquement masculins pour les raisons déjà évoquées en section 0. Les enregistrements des signaux sonores ont été effectués à trois endroits : à environ 15 cm de la bouche du locuteur, à environ 1 mètre du locuteur ainsi qu'au niveau de l'auditeur (cf. Figure 6-4). Ainsi, nous disposons dans ce corpus de la voix du locuteur enregistré proche de la bouche afin de faciliter les analyses. Les 2 microphones (l'un placé à 1 mètre et l'autre placé à la distance de communication) permettent d'avoir deux points fixes d'enregistrement permettant de quantifier éventuellement les effets liés à la propagation de l'onde sonore dans l'air. Le locuteur étant debout, et pouvant ainsi légèrement bouger, il est plus judicieux de placer un microphone fixe proche du locuteur. De plus, un enregistrement électroglottographique (EGG) a été fait sur chaque locuteur afin d'enregistrer avec précision la fréquence de vibrations des cordes vocales. Ce type d'enregistrement est également primordial pour d'éventuelles analyses par filtrage inverse.

Pour ces enregistrements, nous avons utilisé 3 microphones ½ pouce B&K (type 4189), 2 conditionneurs B&K (type 5935L) et un enregistreur numérique TEAC Lx-10, dont la fréquence d'échantillonnage a été fixée à 24 kHz. Les enregistrements électroglottographiques ont été effectués avec l'appareil *Glottal Enterprises EG2*. Ces enregistrements ont été, par la suite, ré-échantillonnés à 16 kHz.



Figure 6-2 : Monture de fixation réalisée pour des microphones B&K



Figure 6-3 : Photos représentant le pupitre du locuteur (à gauche) et le pupitre de l'auditeur (à droite)



Figure 6-4 : Photos représentant le protocole d'enregistrement DB2. A gauche, la photo d'un locuteur muni du micro de bouche et de l'EGG. Au milieu, une vue d'ensemble du locuteur et de l'auditeur. A droite, la photo de l'auditeur équipé de son pupitre muni d'un ordinateur et d'un microphone.

### 6.2.2 Remarques

Malgré les avantages de ce protocole d'enregistrement, le principal problème de ce corpus concerne l'enregistrement nocturne. De nuit, à l'inverse de la journée, la température de l'air à proximité du sol est plus froide que la température de l'air en altitude ; ceci engendre que la vitesse du son est plus élevée en altitude (Embleton, 1996). Ce phénomène engendre une inversion du gradient de température qui aura pour conséquence de canaliser les ondes sonores. En effet, comme le montre la Figure 6-5, de jour, la majeure partie des ondes sonores seront propagées vers le ciel et seule une faible partie de la puissance de la source arrivera au niveau du récepteur. A l'inverse, de nuit, ces mêmes ondes seront réorientées vers le sol par le gradient de température qui à pour conséquence

directe, une intensité plus forte au niveau du récepteur que celle obtenue de jour. Ce phénomène peut alors engendrer des niveaux de production vocale moins intenses de nuit que de jour pour la même distance de communication. Cette production plus basse s'explique d'une part, par un bruit ambiant plus faible, et d'autre part, par la plus grande quantité d'énergie arrivant aux oreilles de l'auditeur, due à la propagation de l'onde. Ainsi le locuteur a besoin de fournir moins d'effort pour pallier la distance et pour émerger du bruit.

La restriction majeure pour ce type d'enregistrements (en champ libre) est qu'ils ne sont pas reproductibles car comme nous l'avons largement évoqué, les stratégies utilisées par les locuteurs (niveau de production, etc.) sont très dépendantes de l'environnement. Ainsi, d'un jour à l'autre, les phénomènes de propagations acoustiques peuvent changer drastiquement. Les conditions météorologiques peuvent engendrer, soit une meilleure, soit une moins bonne propagation des ondes sonores. Cette variation dans les caractéristiques de propagation de l'environnement influencera à son tour le niveau sonore des bruits de fond car de nuit, les bruits lointains se propagent mieux. Le bruit de fond peut également changer radicalement d'un jour à l'autre.

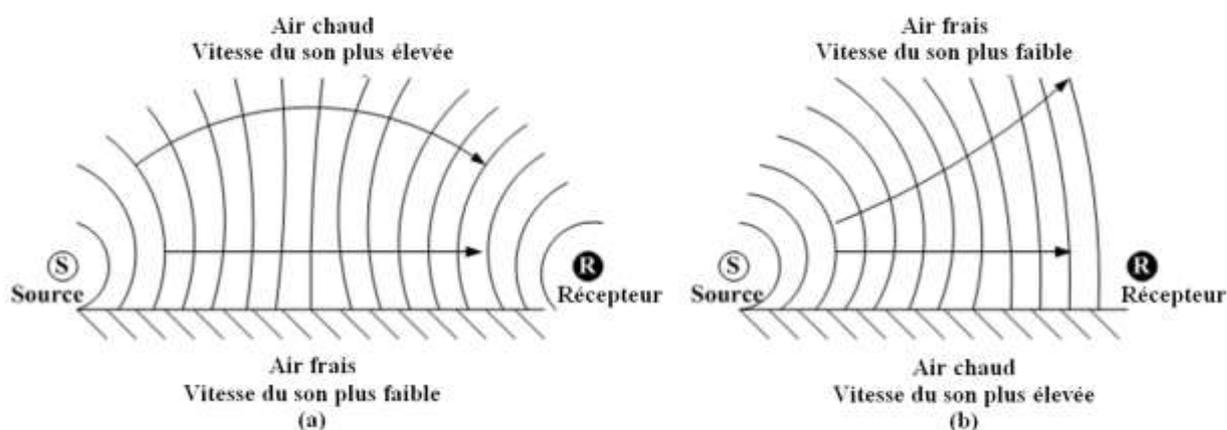


Figure 6-5 : Effet du gradient de température sur la propagation des ondes sonores. a) de nuit, b) de jour

Par ailleurs, nous avons pu juger lors de l'enregistrement de DB2, de la complexité de mise en place d'un tel protocole d'enregistrement : déplacer tout le matériel d'enregistrement, microphone, conditionneur, enregistreurs, les ordinateurs, les tables, de la lumière, des câbles de plus de 100 m, et bien évidemment une source d'énergie pour alimenter tout le matériel. De plus, il n'est pas évident de persuader des personnes à se déplacer en pleine nuit pour se rendre dans un champ !

---

## 6.3 Protocole de simulation de communication orale à distance : DB3

---

Face aux inconvénients et restrictions rencontrés lors de l'enregistrement de DB1 et DB2, nous avons jugé primordial de développer un protocole aisément reproductible, permettant d'assurer des enregistrements de bonne qualité et permettant d'enregistrer le nombre de sujets que nous souhaitons. La meilleure solution permettant de répondre à toutes ces attentes est sans aucun doute un protocole d'enregistrement en laboratoire. Le problème qui se pose alors est de savoir comment enregistrer des situations de communication à 100 m de distance dans un laboratoire ! En effet, aucun laboratoire (à notre connaissance) ne dispose de chambre insonorisé (semi-anéchoïque) d'une telle ampleur. Nous avons donc opté pour une simulation de communication de distance en laboratoire.

Lors d'une communication orale à distance, l'objectif premier d'un locuteur est d'augmenter son niveau de production vocale, afin de palier la chute d'intensité due à la distance. Dans notre simulation, cette chute d'intensité est simulée par l'isolation acoustique d'une chambre insonorisée. Ainsi, un locuteur placé à l'intérieur d'une chambre insonorisée, devra ajuster son effort vocal, en fonction de l'atténuation acoustique, pour se faire comprendre d'un auditeur placé à l'extérieur. En contrôlant soigneusement l'isolation acoustique, il est possible de simuler plusieurs distances de communication. Dans notre cas, il n'a pas été possible de modifier l'isolation des chambres. C'est pourquoi nous avons utilisé deux chambres ayant chacune une isolation différente. L'atténuation de chaque chambre a été mesurée en plaçant un haut-parleur à l'intérieur des chambres, un microphone placé à l'endroit où se situeront les locuteurs, et un microphone placé à l'extérieur à la position de l'auditeur. En utilisant ensuite un analyseur de spectre, nous avons enregistré la fonction de transfert entre les 2 microphones à l'aide d'un sinus glissant de 100 Hz à 10 kHz. Nous avons ainsi obtenu, pour la première chambre, une atténuation moyenne de 36 dB. Cela correspond à une distance théorique de communication d'environ 60 m. Une atténuation de 41 dB a été mesurée pour la seconde chambre, ce qui correspond à une distance théorique d'environ 120 m.

### 6.3.1 Protocole d'enregistrement

Dans un souci de cohérence, nous réutilisons pour ces enregistrements les 5 phrases du Tableau 6-1. Nous avons également choisi d'enregistrer des voyelles isolées. N'étant pas dans ce protocole restreint par le temps alloué aux enregistrements, un ensemble plus complet de voyelles a été retenu. Le corpus contient ainsi les voyelles : [a], [ə], [i], [o], [y], [u], [e], [ɛ], [œ].

Les enregistrements se sont donc déroulés dans 2 chambres d'audiométrie permettant de simuler 2 distances de communication : 60 m et 120 m. Les voix modales ont également été enregistrées aussi bien pour la distance de 60 m que pour la distance de 120 m. Ceci permet d'avoir une référence de voix modale enregistrée peu de temps avant l'exercice de simulation de communication à distance. Cette précaution a été prise car les enregistrements se sont déroulés sur plusieurs jours.

Ces enregistrements sont organisés de manière à recréer virtuellement une nécessité d'augmentation de l'effort vocal dans le but de se faire comprendre. Pour cela, nous plaçons, comme représenté sur la Figure 6-6, un locuteur dans une salle d'audiométrie. Le locuteur possède une liste similaire à celle utilisée dans le protocole DB2 (alternativement des nombres, des voyelles et 5 phrases (cf. Tableau 6-1)) qu'il doit communiquer à l'auditeur, placé à l'extérieur de la salle.

Lorsque le locuteur a prononcé l'un des nombres ou l'une des voyelles, l'auditeur tape à l'ordinateur ce qu'il a compris. Le locuteur, qui voit également l'écran, a de cette manière un retour de son auditeur concernant la compréhension du message transmis. Ainsi, si l'effort vocal fourni par le locuteur n'est pas assez grand pour que l'auditeur comprenne le message, il peut s'en apercevoir via l'écran et prononcer à nouveau le chiffre avec un effort vocal plus approprié. Concernant les phrases, la consigne donnée à l'auditeur était d'exécuter les actions demandées ou de répondre aux questions posées. Ce retour de compréhension permet au locuteur d'adapter son effort vocal à la situation. Pour l'enregistrement en voix modale, le locuteur lisait également la liste qui lui était fournie à un auditeur placé à 1 mètre de lui dans la chambre.

Étant donné l'utilisation d'un protocole de simulation en laboratoire, il est bien plus aisé d'enregistrer un nombre représentatif de locuteurs. Ce corpus de voix contient ainsi 12 locuteurs (9 hommes et 3 femmes) qui ont chacun participé aux 2 simulations de distance (60 m et 120 m). Lors de ces enregistrements, nous avons choisi d'enregistrer les voix de sujets féminins, afin de comparer les effets de l'effort vocal en fonction du sexe du locuteur. Nous disposons ainsi des voix modales et criées des 12 sujets pour le protocole de simulation à 60 m. Nous disposons également des voix modales et des voix criées pour la simulation à 120 m.

Pour réaliser ces enregistrements, un microphone a été placé à environ 1 mètre de la bouche du locuteur. Un électroglottographe a également été placé sur le cou de chaque locuteur. Ces enregistrements ont été effectués via un microphone ½ pouce B&K (type 4189), 1 conditionneur B&K (type 5935L) et un enregistreur numérique TEAC Lx-10, dont la fréquence d'échantillonnage a été fixée à 24 kHz. Les enregistrements électroglottographiques ont été effectués avec l'appareil *Glottal Enterprises EG2*. Ces enregistrements ont été par la suite ré-échantillonnés à 16 kHz.

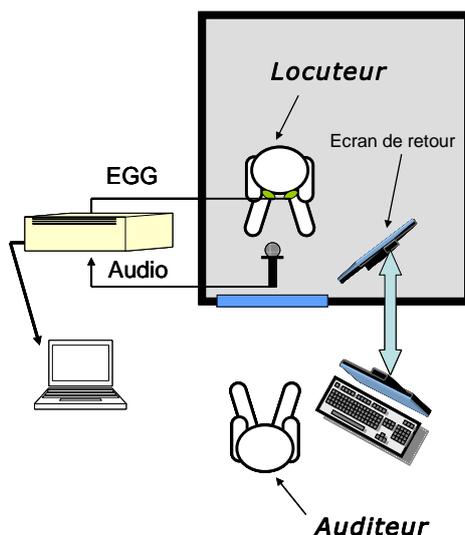


Figure 6-6 : schématisation du protocole d'enregistrement de DB3

### 6.3.2 Remarques et observations

Bien que ce protocole d'enregistrement offre des avantages certains, quelques remarques sont nécessaires. Le problème majeur de cette simulation est la proximité du locuteur et de l'auditeur. En effet, la nécessité d'augmenter l'effort vocal pour le locuteur n'est pas régie par l'augmentation de la distance de communication mais par un obstacle plus ou moins perméable à l'onde acoustique. Nous faisons ici l'hypothèse forte que la réaction du locuteur dans ces deux cas distincts sera la même. Toutefois il est envisageable que ce ne soit pas le cas. Nous verrons pourtant, en comparant les corpus précédents à celui-ci, que les réactions sont très similaires.

Un deuxième aspect non négligeable est le fait que l'auditeur voit le locuteur à travers une vitre. Les enregistrements peuvent alors être biaisés dans le cas où l'auditeur lirait sur les lèvres du locuteur. Ce problème pourrait être résolu en masquant la bouche du locuteur. Toutefois, une attention particulière devra être prise quant au choix du matériau et à sa position. Un matériau trop réfléchissant et trop proche viendrait perturber les mesures et diminuer l'efficacité de la voix du locuteur. En effet, la voix ne serait plus projetée vers l'avant de la chambre (vers l'auditeur), mais réfléchi vers l'arrière. Dans notre cas nous avons opté pour une solution plus simple. Nous avons choisi un seul auditeur pour tous les enregistrements. Celui-ci détachait régulièrement son regard du locuteur. Ceci permet donc d'assurer que le niveau d'effort vocal produit par le locuteur est adéquat à la situation et que la perception auditive est suffisante à la bonne compréhension.

## 6.4 Corpus dédié à l'analyse micro-prosodique : DB4

---

Les différents corpus enregistrés jusqu'à présent (DB1, DB2, DB3) contiennent les 5 phrases du Tableau 6-1, ainsi que plusieurs voyelles. Nous avons, en première approche dans cette étude, voulu étudier et transformer des phrases courtes. Comme nous le verrons pas la suite, il nous est rapidement apparu qu'il était trop complexe et ambitieux de vouloir transformer une phrase parlée entière en une phrase criée. Il nous a alors semblé plus réaliste de faire le choix de transformer uniquement des mots. Cependant, hormis les nombres présents dans 2 d'entre eux (DB2 et DB3) qui ne sont pas les mêmes d'un locuteur à l'autre ou d'un corpus à l'autre, les corpus ne contiennent pas de structure phonétiques simples. C'est pourquoi, un quatrième et dernier corpus a été enregistré dans le but d'une meilleur compréhension et modélisation de l'effort vocal dans la communication verbale à distance.

### 6.4.1 Protocole d'enregistrement

Ce corpus a pour but de réaliser une étude précise de la prosodie et de la micro-prosodie des voix projetées. Ainsi, nous avons choisi d'enregistrer 4 structures phonétiques de base du français, à savoir : CV (i.e. consonne-voyelle), CVC (i.e. consonne-voyelle-consonne), VCV (i.e. voyelle-consonne-voyelle) et CVCV (i.e. consonne-voyelle-consonne-voyelle).

Ces structures de base sont construites à partir des 17 consonnes du français (hors semi-consonnes) :

- 2 liquides : [l], [ʀ]
- 3 nasales : [m], [n], [ɲ]
- 3 fricatives voisées : [v], [z], [ʒ]
- 3 fricatives non voisées : [f], [s], [ʃ]
- 3 occlusives voisées : [b], [d], [g]
- 3 occlusives non voisées : [p], [t], [k]

et, des 3 voyelles [a], [i], [u]. Ces trois voyelles ont été choisies car elles correspondent chacune à des caractéristiques précises de l'ensemble vocalique. Ces trois voyelles sont les trois sommets du triangle vocalique représentant les voyelles orales (également appelées voyelles cardinales). La voyelle [a] est une voyelle ouverte, la voyelle [i] est une voyelle fermée avant non-arrondie et la voyelle [u] est également une voyelle fermée arrière arrondie.

A partir de cette sélection d'unité phonétique, nous avons composé le corpus de logatomes de la manière suivante. Nous avons construit les logatomes en combinant chacune des 17 consonnes avec les 3 voyelles. Ainsi, pour les logatomes CV, on crée 51 logatomes : 17 avec la voyelle [a], 17 pour [i] et 17 pour [u]. Pour les structures faisant intervenir plusieurs consonnes ou voyelles, nous avons fait le choix d'utiliser les mêmes consonnes ou les mêmes voyelles au sein d'un logatome. Par exemple, pour la combinaison d'un [a], et d'un [m], nous avons créés, /ma/, /mam/, /ama/, /mama/. L'ensemble des logatomes utilisés est listé en Annexe A. Ainsi, pour chaque structure (CV, CVC, VCV, CVCV), 51 logatomes ont été créés, ce qui constitue une base de données de 204 logatomes.

Pour l'enregistrement de ces logatomes, nous avons utilisé le même protocole que pour l'enregistrement de DB3 (simulation de distance). Les essais simulant une distance de 120 m étaient problématiques, dans le sens où il était trop difficile de comprendre l'ensemble des mots prononcés. Ceci avait pour conséquence directe, la fatigue des locuteurs et la nécessité de lire sur les lèvres du locuteur pour mener à bien les enregistrements. C'est pourquoi, nous n'utilisons que la chambre équivalente à une distance de 60 m. Pour chaque logatome prononcé, l'auditeur placé à l'extérieur écrivait sur l'ordinateur les mots qu'il avait entendus assurant ainsi un retour de compréhension.

Les voix modales quant à elles, ont été enregistrées avec un auditeur placé à 15 cm du locuteur dans la chambre insonorisée (cf. Figure 6-7). Ce corpus contient 3 locuteurs masculins. Cette base de données comprend alors 204 logatomes prononcés par 3 locuteurs utilisant 2 niveaux d'effort vocal ; soit un ensemble de 1224 logatomes. Les enregistrements ont été effectués via un microphone ½ pouce B&K (type 4189), placé sur la monture de fixation permettant de tenir le microphone proche de la bouche (cf. Figure 6-7), 1 conditionneur B&K (type 5935L) et un enregistreur numérique portable MicroTrack II dont la fréquence d'échantillonnage était de 48 kHz. Ces enregistrements ont été par, la suite, ré-échantillonnés à 16 kHz. Un électroglottographe a également été placé sur le cou de chaque locuteur.

#### 6.4.2 Remarques et observations

Lors de l'enregistrement de ce corpus nous avons fréquemment observé une tendance à employer une intonation interrogative en voix parlée (notamment pour 2 des locuteurs). Cette attitude est interprétée comme une réaction du locuteur lui permettant de poser la question : « *As-tu bien compris ce que je te dis?* ». Malgré tout, les enregistrements restent exploitables du fait que ce biais a été identifié. De plus nous verrons qu'en voix criée, cette tendance disparaît au profit d'une uniformisation de la structure prosodique pour les trois locuteurs.

Une solution envisageable pour des enregistrements plus naturels serait l'utilisation d'une situation d'échange réaliste. Pour étudier des mots précis utilisés lors d'une communication (ou d'un échange)

plus réaliste, Garnier (2007) propose l'utilisation d'un jeu. Ce jeu est basé sur l'échange d'informations. Deux sujets disposent de certains éléments permettant de résoudre une énigme que l'autre n'a pas. Ainsi, pour résoudre cette énigme ils doivent échanger des informations. Ces informations contiennent des mots clés, qui eux seuls seront considérés et analysés. Toutefois, en fonction de la position du mot dans la phrase, ou encore du contexte dans lequel le mot sera prononcé (par exemple dans le cas où l'on répète car le mot a mal été compris), les informations acoustiques et prosodiques seront différentes.

Ainsi, il n'existe pas de solution parfaite pour ce type d'enregistrement, et chaque protocole a de ces défauts et limites. Toutefois, l'essentiel est d'en avoir conscience afin de ne pas faire de conclusions hâtives.

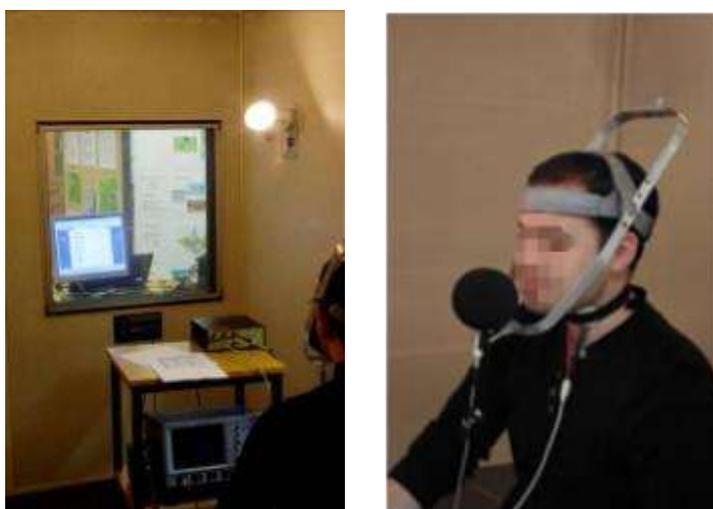


Figure 6-7 : Photos représentant le protocole d'enregistrement de DB4

## 6.5 Conclusion du chapitre 6

---

Pour mener à bien notre étude, quatre corpus ont été enregistrés, chacun possédant un rôle précis. Le premier corpus (DB1) a été réalisé afin de mener à bien une étude perceptive, quant à la pertinence des paramètres prosodiques par rapport aux paramètres spectraux dans la perception de l'effort vocal (cf. CHAPITRE 7). Ce corpus possède plusieurs désavantages et notamment le nombre limité de locuteurs. C'est pourquoi, le deuxième corpus (DB2) a été enregistré afin de disposer de plus de locuteurs ainsi que de limiter le bruit ambiant. Ce corpus a été effectué dans le but de réaliser des analyses globales de la prosodie (cf. CHAPITRE 8). Néanmoins, les enregistrements réalisés en champ libre, sont difficiles à mettre en œuvre, et de plus, ne sont pas facilement reproductibles. Ainsi, nous avons souhaité développer un protocole d'enregistrements simulant une communication à

distance au sein du laboratoire. Ainsi, le corpus DB3 a été conçu, afin de permettre de valider le protocole de simulation, en comparant certaines grandeurs (notamment les moyennes et les dynamiques de l'intensité et de F0) aux enregistrements effectués en extérieur (cf. CHAPITRE 8). Ce protocole étant validé au regard des grandeurs observées, nous avons alors pu enregistrer un corpus destiné à l'analyse prosodique détaillée lors de la production de voix criée (cf. CHAPITRE 9).

# Perception de l'effort vocal : Prosodie ou paramètres spectraux ?

---

*"More effort results in greater intensity and spectral complexity"*  
— John M. Chowning

Parmi le nombre conséquent de modifications observées par les nombreux travaux cités jusqu'à présent, il nous a semblé important avant d'entreprendre des transformations de voix, d'essayer de déterminer quelles sont les paramètres les plus pertinents pour la perception de la distance et donc pour la perception de l'effort vocal. En effet, l'Homme a la capacité de percevoir l'effort fourni par un locuteur et ainsi d'en déduire la distance à laquelle celui-ci se situe (cf. CHAPITRE 2). Toutefois, la question qui reste encore aujourd'hui en suspens est la suivante :

*Qu'est ce qui permet à l'Homme de différencier une voix parlée d'une voix criée ou encore d'une voix chuchotée ?*

En d'autres termes quelles sont les caractéristiques du signal de la parole que nous prenons en compte pour juger de l'effort vocal fourni par un locuteur.

Au cours de nos travaux il nous est arrivé de nombreuses fois de poser cette question à différentes personnes. Cette question laisse la majorité des gens muets. Bien sûr, certains d'entre eux développent des arguments qui leur sont propres, reflétant leurs sentiments mais rarement étayés par des preuves scientifiques. C'est pourquoi nous nous penchons sur cette question, en élaborant un état de l'art des éléments pertinents, permettant la perception de l'effort vocal et donc de la distance d'un locuteur.

Lorsqu'on interroge les gens, les esprits les plus cartésiens nous annoncent alors qu'une voix criée est plus forte qu'une voix parlée. En effet, mais une voix forte n'est pas forcément une voix criée ; augmenter le niveau sonore du signal ne permet pas de ressentir plus d'efforts dans la voix. D'autres nous annoncent alors que, bien qu'une voix criée soit plus forte, la principale caractéristique est le caractère plus aigu. En effet ces personnes ont tout à fait raison car une voix criée s'accompagne d'une augmentation de F0. Or, de la même manière que précédemment, la réciproque n'est pas forcément

vrai ; une voix aiguë n'est pas forcément une voix criée. Ainsi, une voix criée est plus intense et plus aiguë qu'une voix parlée mais l'inverse n'est pas toujours vérifié. En effet, une simple transformation de voix parlée en une voix plus intense et plus aiguë nous permet rapidement de nous rendre compte que celle-ci ne constitue pas une voix reflétant plus d'efforts vocaux. De la même manière une voix d'enfant ou une voix féminine, qui sont plus aiguës qu'une voix d'homme, ne paraissent pas criées.

Nous avons déjà pu constater que l'intensité ne constitue pas, en soi, un élément déterminant pour la perception de l'effort vocal et donc de la distance d'un locuteur. Néanmoins, l'intensité peut être primordiale pour la perception de la distance d'une source sonore quelconque, mais n'est pas pertinente dans le cas de la parole (Eriksson et Traunmüller, 2002; Fux et Zimpfer, 2009; Philbeck et Mershon, 2002; Traunmüller, 1997). D'autre part, une étude, dont nous avons déjà exposé les détails (cf. §2.1.2.3), nous permet de dire que les phénomènes liés à la propagation acoustique de l'onde sonore de la parole, et notamment la chute des hautes fréquences due à la propagation dans l'air, semblent pour des distances convenant à une communication orale, ne pas être pertinents.

Afin d'éclaircir ce point, nous réalisons dans un premier temps, un état de l'art des connaissances sur le sujet. Il n'existe, à notre connaissance, que très peu d'études directement liées à la pertinence des paramètres qui sont modifiés avec l'effort vocal. C'est pourquoi cet état de l'art traite aussi bien des études directement consacrées à la question, mais également aux études qui se destinent à la transformation de la parole pour augmenter ou diminuer l'effort vocal ressenti.

## **7.1 État de l'art sur les paramètres pertinents de l'effort vocal**

---

Dans cette section nous effectuons un état de l'art concernant les éléments pertinents pour la perception de l'effort vocal, aussi bien pour la perception de la voix criée que de la voix chuchotée. Le cas des voix chuchotées, étant *a priori* moins complexe que celui des voix criées, nous ne faisons qu'un bref état de l'art concernant les différentes méthodes proposées afin de transformer une voix parlée en voix chuchotée. Pour les voix criées, en revanche, nous mentionnons des études directement dédiées à l'évaluation de la pertinence de certains paramètres, ainsi qu'un état de l'art des différentes études destinées à la transformation de voix parlée en voix criée.

### 7.1.1 Éléments pertinents pour la perception des voix chuchotées

La majorité des solutions proposées dans le but de transformer une voix modale en voix chuchotée n'applique qu'une seule modification, le dévoisement du signal, et se base sur la qualité des algorithmes d'analyse synthèse dans le but d'obtenir des qualités et un naturel de voix. En effet, il est clair que l'élément majeur, nous permettant de reconnaître une voix chuchotée, est l'absence de vibrations des cordes vocales. On note par exemple la modification en utilisant le modèle ARX-LF visant à séparer de manière plus efficace le filtre du conduit vocal de la source (Agiomyrghiannakis et Rosec, 2009), des méthodes basées sur le vocodeur de phase comme dans (Zölzer et al. 2002, p. 290) ou dans des version plus évoluées (Verfaille et Arfib, 2001). On trouve également des modifications en voix chuchotée à l'aide du synthétiseur de KLATT88 (Burkhardt, 2009) ou encore par des méthodes de décomposition périodique-apériodique du signal permettant de conserver la partie stochastique du signal intacte (D'Alessandro et Doval, 1998).

Dans une étude différente des précédentes, (Nordstrom, 2008; Nordstrom et al., 2006, 2008) qui vise à réduire l'effort vocal d'une voix pressée en une voix *breathy* mentionne que « l'enveloppe spectrale de la source vocale, constitue l'un des facteurs les plus importants pour la perception de l'effort vocal »<sup>1</sup>. Ceci va également dans le sens de Sun (2000). Ces auteurs modifient l'effort vocal d'une voix en appliquant un algorithme APLP (*Adaptive Pre-emphasis Linear Prediction*), dans le but de modéliser la forme générale du spectre (en utilisant des prédicteurs LP d'ordre 1 ou 3) et de la modifier. En effet, ils montrent qu'une voix qui contient de l'effort vocal (appelée *high-effort voice* en référence à une voix tendue) possède un spectre relativement plat jusqu'à 4-5 kHz, alors que pour une voix *breathy* celle-ci décroît linéairement.

De la même manière, pour la diminution de l'effort perçu et dans le même esprit que Nordstrom, en plus du dévoisement du signal, Farner et al. (2008) s'attachent à modifier le spectre des parties voisées du signal, en annulant ou en inversant la pente spectrale dans la zone du signal compris en-deçà de 3 kHz.

Ainsi, au regard de ce bref état de l'art, il en ressort que la modification principale consiste à dévoiser le signal de la parole. Toutefois, certaines modifications en termes d'enveloppe spectrale, peuvent également s'avérer nécessaires pour obtenir une bonne qualité de transformation, afin de rendre la voix synthétisée plus naturelle.

D'autres phénomènes qui ne sont pas directement liés à la production de la parole nous paraissent également être importants. En effet, la proximité d'un locuteur, lorsqu'il chuchote, va permettre à l'auditeur de percevoir les bruits de bouche. Il est alors possible d'entendre par exemple le bruit

---

<sup>1</sup> Traduction personnel de : "The spectral envelope of the voice source provides one of the most important cues for the perception of vocal effort"

effectué lors de l'ouverture de la bouche, le bruit fait par le contact de la langue avec d'autres parties de la bouche ou encore le bruit fait par l'ouverture de la glotte. D'un point de vue perceptif, la logique veut que si nous sommes capables d'entendre ces types de bruits, cela signifie que le locuteur se trouve à proximité de nous. Ainsi, une sensation de proximité d'un locuteur pourrait être apportée en amplifiant les bruits de bouche. D'autre part, la proximité d'un locuteur modifie de façon significative les HRTFs (*Head Related transfer functions*) et notamment, si la source sonore ne se situe pas en face du locuteur (Brungart et Rabinowitz, 1999; Brungart, 1999; Brungart et al., 1999). Ainsi, dans le cas de l'utilisation de voix chuchotée naturelles ou transformées, à travers un casque et une représentation 3D, les HRTFs doivent être prises en compte.

### 7.1.2 Éléments pertinents pour la perception des voix criées

Penchons-nous à présent sur le cas des voix criées. Nous faisons tout d'abord un état de l'art des travaux concernant la pertinence de certains paramètres pour la perception de l'effort vocal ou de la distance. Étant donné qu'il n'existe que très peu de travaux à ce sujet, nous listons également les différents travaux qui proposent des méthodes afin de réaliser des transformations de voix parlée en voix criée.

#### 7.1.2.1 Les éléments perceptivement pertinents

Rares sont les études qui traitent directement de la pertinence des paramètres pour la perception de l'effort vocal et/ou de la distance du locuteur. L'une d'entre elles est l'étude réalisée par Brungart et al. (2002). Dans cette étude, Brungart et al. visent à déterminer le **rôle de la F0** dans la perception du changement de l'effort vocal fourni et ainsi de la distance perçue du locuteur. Pour ce faire, ces auteurs ont utilisé un l'algorithme de transformation PSOLA pour intervertir les contours de F0 de différents mots, prononcés à des efforts vocaux variés. A partir des leurs enregistrements (30 mots produits à des niveaux acoustiques de 60, 66, 72, 78, 84, 90 et 96 dB SPL par quatre locuteurs), ils ont choisi de construire deux séries de mots. La première série se compose de l'ensemble des mots dont le contour de F0 a été interverti avec le contour de ces mêmes mots, mais pour le niveau de production supérieur de 6 dB SPL. La deuxième série a été créé de la même manière, mais en prenant les contours de F0 des mots produits à des niveaux inférieurs de 6 dB SPL. Dans une expérience visant à déterminer l'éloignement d'un locuteur parmi 6 distances (0,5, 1, 2, 4, 8 et 16 m), 6 sujets ont entendu, via des hauts parleurs (2 réels et 4 leurres), les 3 séries de mots. A savoir, la série de mots contenant les enregistrements de base, la série dont les mots ont subi une augmentation de la F0 et la série où la F0 a été diminuée. Les résultats obtenus sont sans appel. En effet, pour le cas des séries ayant subi une augmentation de la F0, bien que la distance estimée ne soit pas identique à celle du mot produit à 6 dB SPL de plus, la distance perçue augmente. Le constat inverse a été fait sur la série de mots où la F0 a

été diminuée, qui sont perçues plus proches que les voix naturelles. Ainsi Brungart et al. concluent que la F0 joue un rôle important dans la perception de la distance. Il signale toutefois que, d'après les résultats obtenus, la F0 n'est pas le seul facteur permettant de modifier la distance perçue à partir de la voix ; ceci car la distance perçue pour une voix ayant la F0 d'un mot produit à 6 dB SPL de plus, n'est pas perçue aussi loin que le mot qui est réellement produit à 6 dB de plus. La F0 est donc un facteur très important mais qui n'est pas le seul.

Dans une étude visant à mettre en avant le rôle des paramètres glottiques dans la perception de l'effort vocal Yi et al (2000) évoquent le fait que le rôle de la F0, notamment la montée initiale et la chute finale, est un paramètre très important pour la perception de l'effort vocal. En effet, dans leurs études, ces auteurs s'attèlent à modifier la voyelle parlée /a/ pour atteindre 3 niveaux d'efforts différents. Pour ce faire, ils modifient la F0 en accord avec des enregistrements de cette même voyelle en voix criée. La F0 des voix criées est toutefois modélisée par des courbes décrites par 5 points. D'autre part, des modifications du quotient ouvert (Oq), du quotient de vitesse (Sq) ainsi que de la pente spectrale (TL) de la source sont appliquées via le synthétiseur KLATT88 (Klatt et Klatt, 1990) où la source glottique est remplacée par la modélisation LF (Fant et al., 1985). A la suite d'un test perceptif et d'analyses statistiques, ces auteurs concluent que les évolutions de F0 jouent un rôle primordial dans la perception de l'effort vocal. De plus ils mentionnent que pour la perception de l'effort, les paramètres glottiques ne semblent pas tous avoir la même contribution et que, par ordre d'importance, les paramètres glottiques sont Sq, Oq et enfin TL (pente spectrale). Cette étude est toutefois critiquée par Eriksson et Traunmüller (1999) qui notent que la forte influence statistique de l'évolution de F0 observée dans la perception de l'effort vocal peut être liée uniquement à l'augmentation de la F0 elle-même. En d'autres termes, le fait d'augmenter la F0, sans pour autant modifier le contour de F0, serait suffisant.

D'une façon plus générale, l'étude de Tassa et Liénard (2000) va également dans le sens de Brungart et al. (2002). Tassa et Liénard ont, dans leur étude, décrit une expérience visant à transformer une voyelle parlée en une voyelle criée. Pour ce faire, ils ont choisi de greffer sur la voyelle parlée, les paramètres prosodiques de la voyelle criée. Dans une seconde expérience ils greffent les paramètres spectraux de la voix criée sur la voix parlée dans le but de juger de chaque groupe de facteurs, indépendamment d'un point de vue perceptif. Tassa et Liénard concluent sur le fait que les paramètres prosodiques sont plus pertinents pour la perception de l'effort vocal que les paramètres spectraux (position des formants, pente spectral). Ce constat n'est malheureusement pas confirmé par des tests perceptifs.

Soulignons cependant qu'il existe des auteurs qui considèrent que les paramètres spectraux sont les plus pertinents. Allant dans ce sens nous pouvons citer Traunmüller (1985) qui, sur la base d'une série

de tests visant à modifier la distance tonotopique<sup>1</sup> entre la F0 et F1 (noté Z1-Z0) ou encore entre F1 et F2 (notés Z2-Z1), démontre qu'un déplacement de F0 et de F1, tout en maintenant la distance Z1-Z0 constante, procure une sensation d'augmentation de l'effort vocal (voir également (Traunmüller, 1988, 1997)). Nous avons effectué le même type de mesure. La position des formants a été mesurée par l'identification des maxima des filtres LP sur les voyelles des logatomes CV du corpus DB4. Pour chaque logatome CV nous ne considérons toutefois qu'une seule valeur de formants correspondant à la moyenne sur un intervalle de confiance de 80%. Toutefois, à partir des relevés de la position des formants et de la F0 que nous avons effectués, nous n'observons pas une telle constance (cf. Figure 7-1). Néanmoins, Garnier (2008) observe une constance de la distance tonotopique pour une voix Lombard.

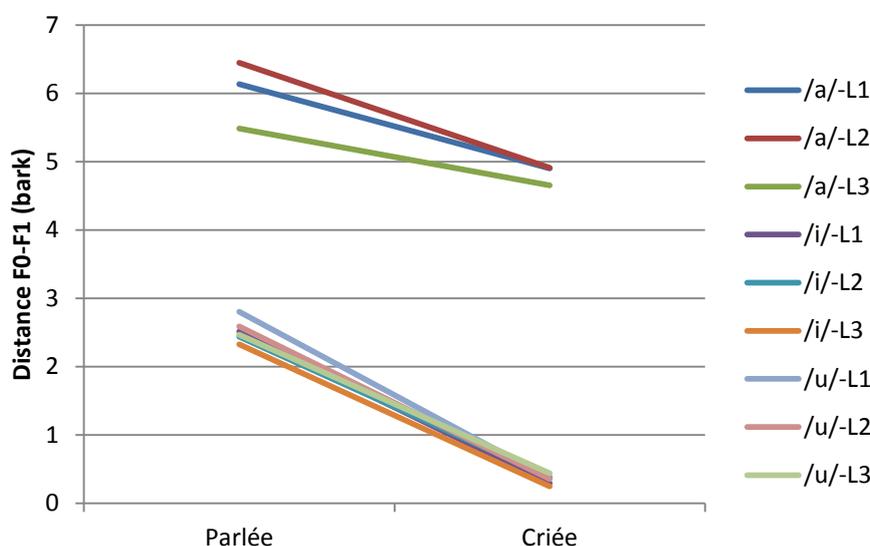


Figure 7-1: Distance tonotopique (F1-F0) en bark, pour les voyelles des logatomes CV des trois locuteurs en fonction de l'effort vocal

Malgré la tendance dans la littérature à considérer la prosodie comme étant l'élément le plus pertinent, il n'existe que peu d'études sur le sujet. Afin de compléter ces propositions, nous explorerons également les études qui ne traitent pas directement du sujet mais qui proposent des méthodes de transformation qui peuvent alors nous apporter des renseignements supplémentaires.

#### 7.1.2.2 Considérations pour la transformation de voix dans l'objectif de percevoir l'effort vocal

Cheyne et al. (2009) ont réalisé des transformations de voix modale en voix plus ou moins criées, dans le but d'apporter une notion de distance grâce à ces transformations. Les transformations sont effectuées sur des mots courts en utilisant un vocodeur MELP (*Mixed-Excitation Linear Prediction*).

<sup>1</sup> Distance tonotopique : écart en tons ou en Bark entre deux fréquences.

Ces auteurs proposent de réaliser les modifications suivantes. Une augmentation de l'intensité de 8 dB par doublement de distance, une F0 augmentée de 40 Hz par doublement de distance, F1 augmentée de 25 Hz par doublement de distance et une modification de la pente spectrale (A3-A1 : amplitude du 3<sup>ème</sup> formant moins l'amplitude du 1<sup>er</sup> formant) diminuée de 2 dB par doublement de distance. Après différents tests perceptifs, considérant la variation d'intensité causée par la propagation atmosphérique, les résultats des tests sont encourageants. En effet, plus les modifications apportées sont importantes, plus la distance perçue est importante. Toutefois, ces auteurs ne testent que des variations de distance allant de 1 m à 32 m. Or, fondamentalement, l'effort vocal pour 32 mètres est faible. Ces auteurs ne concluent malheureusement pas sur l'importance de tels ou tels paramètres, mais se réfèrent principalement aux travaux liés à l'analyse des variations de paramètres par l'effort vocal que nous avons déjà cités jusqu'ici.

D'Alessandro et Doval (1998) quant à eux, préconisent pour transformer une voix modale en une voix contenant plus d'effort vocal, d'augmenter le rapport entre les composantes périodiques et apériodiques de la source (i.e. le résidu de l'analyse LP), de modifier le spectre du résidu (en particulier augmenter la zone entre 1500-2500 Hz) mais également en augmentant la pente globale du résidu de -12 dB/octave à -6 dB/octave et d'augmenter le formant glottique (Richard et D'Alessandro, 1996). Toutefois, nous n'avons pas trouvé dans la littérature des méthodes de transformation qui appuient la marche à suivre décrite par ces auteurs.

La thèse de Nicolas D'Alessandro (2009) propose un continuum permettant de modifier une voix modale en une voix plus ou moins forcée, en modifiant les paramètres de la source par le modèle LF. Il propose ainsi, en fonction du mécanisme de la voix, de faire varier  $O_q$ ,  $\alpha_m$  et TL afin de contrôler l'effort vocal. Néanmoins, ces auteurs n'ont pas confirmé leurs propositions par l'utilisation de tests perceptifs, ce qui ne permet malheureusement pas d'infirmer ou de confirmer la démarche proposée.

Drugman et al. (2010), par le biais du modèle déterministe plus stochastique (DSM) (Drugman et al., 2009), transforme une voix modale, générée par une synthèse de parole basée sur des HMMs (*Hidden Markov Models*), en une voix plus tendue ou plus douce. Pour ce faire, ces auteurs utilisent les résidus propres (voir (Drugman et al., 2009)) comme base déterministe du signal de source. Ils modifient également la fréquence maximale de voisement, en fonction du caractère souhaité de la voix (4600 Hz pour une voix tendue, 3990 Hz pour une voix modale et 2460 Hz pour une voix douce), et modifient la pente spectrale en utilisant les spectres moyens des trois types de qualité de voix qu'ils étudient. Les résultats d'un test perceptif montrent que ces modifications entraînent des efforts vocaux ressentis différents. En effet, la transformation en voix tendue procure bien une sensation voulue. Il s'agit ici de travaux qui utilisent le fruit des analyses d'un corpus afin d'en modifier la qualité vocale de voix de synthèse, dans le but d'apporter un caractère tendu ou doux à ces voix. Il ne s'agit pas directement de la transformation de l'effort vocal dans le sens où cette étude ne cherche pas à obtenir une voix criée

mais une voix tendue. Ces auteurs montrent toutefois, que le caractère tendu se reflète par les paramètres de la source.

On retrouve également des auteurs allant dans ce sens comme par exemple Watterson et al. (1993) qui suggèrent que la perception de l'effort vocal est une perception liée au larynx, ou comme Seshadri et Yegnanarayana (2009) qui nous précisent que la perception de la sonie semble être plus influencée par les caractéristiques de l'excitation de la voix (*strength of excitation* : l'apparence plus ou moins impulsionnelle du résidu de la parole) que par les facteurs prosodiques.

Bou-Ghazale et Hansen (1995, 1996a, 1996b), quant à eux, proposent une méthode permettant de modifier une voix modale, en des voix *Lombard* fortes ou des voix énervées. Pour ce faire, ils utilisent un générateur de source, ainsi que des modèles représentant la position du premier formant et le gain de la parole, construit à partir de modèles HMM. Par la suite, ils utilisent un vocodeur CELP (*Code-Excited Linear Prediction*), ou encore un vocodeur RELP (Bou-Ghazale et Hansen, 1998) avec modification du pitch par ré-échantillonnage afin d'appliquer les différentes modifications à 16 mots du vocabulaire de l'aviation. « *Les évaluations effectuées dans cette étude montrent que le pitch est capable de refléter l'état émotionnel du locuteur, alors que l'information sur les formants n'est pas aussi bien corrélée au stress. Néanmoins, dans le cadre d'une modélisation CELP de la parole, c'est la combinaison de la localisation des formants, du pitch et de l'énergie qui produit l'indicateur le plus fiable du stress émotionnel.* »<sup>1</sup>(Bou-Ghazale et Hansen, 1996b). Cette étude n'est cependant pas liée à la perception de la distance à partir des voix modifiées, mais a pour rôle d'améliorer les systèmes de reconnaissance ainsi que la synthèse de voix sous stress.

On retrouve également des méthodes de modification de l'effort vocal par concaténation d'unités (Schröder et Grice, 2003), ou encore de la conversion de voix incluant des interpolations de paramètres, pour atteindre différents niveaux d'effort (Calzada et Socoró, 2011) dans la synthèse TTS (*text to speech*) (Turk et al., 2005). Chez ces derniers, la conversion est faite entre une base de données contenant des voix avec 3 niveaux d'efforts vocaux (*soft, modal, loud*) et ils interpolent les spectres de ces différents niveaux, afin de réaliser des conversions d'un type de voix à un autre. Ainsi, leur approche tend à dire que le spectre semble être primordial pour la perception de l'effort vocal. Toutefois, ils précisent que la prise en compte et une modification du résidu du signal semble être nécessaire (ce qu'ils n'appliquent pas dans leurs études).

<sup>1</sup> Traduction personnelle de : “*The stressed speech parameter evaluations from this study revealed that pitch is capable of reflecting the emotional state of the speaker, while formant information alone is not as good a correlate of stress. However, the combination of formant location, pitch and gain information proved to be the most reliable indicator of emotional stress under a CELP speech model.*”

Au regard de cet état de l'art, il ne ressort pas d'éléments pertinents admis par tous, mais différentes approches plus ou moins efficaces. Nous avons donc choisi de mener une étude complémentaire afin d'éclaircir certains points.

## 7.2 Le rôle de la prosodie dans la perception auditive de l'effort vocal

Dans le but d'une meilleure compréhension des phénomènes liés à la perception de l'effort vocal (donc à la perception de la DLA), et à la recherche d'éléments pertinents pour indiquer la distance d'un locuteur, nous avons mené une étude permettant d'apporter des éléments quant à l'axe de recherche à privilégier. .

A travers la littérature, on peut trouver un nombre conséquent de paramètres reliés à la perception de l'effort vocal. Par ailleurs, plusieurs études démontrent que la prosodie constitue l'élément le plus pertinent pour la perception de la l'effort vocal. Afin de faire notre propre opinion sur le sujet, nous avons réalisé à notre tour, une étude permettant de déterminer les paramètres à privilégier pour réaliser une transformation de voix parlée en voix criée. Il ne s'agit pas de tester un à un l'ensemble des paramètres vocaux, qui varient entre une voix parlée et une voix criée. Ainsi, afin d'étayer l'hypothèse concernant la prédominance de la prosodie, nous avons fait le choix de distinguer deux groupes de paramètres. Le premier groupe que nous appelons « *paramètres prosodiques* » regroupe les variations de la F0, de l'intensité et de la durée.

*Remarque :*

**Dans une certaine mesure, la valeur moyenne de F0 fait partie des paramètres prosodiques, cette valeur étant directement liée à la variation de la F0. Nous faisons ici la distinction et mettons l'accent sur l'évolution de F0. Le but de cette étude est de vérifier la pertinence des évolutions car il a été montré qu'augmenter uniquement la valeur moyenne de F0 sans tenir compte des évolutions n'est pas suffisant (Fux, Feng, & Zimpfer, 2012).**

Le deuxième groupe englobe l'ensemble des paramètres dits « paramètres spectraux ». Ils décrivent toutes les variations pouvant être capturées par l'enveloppe spectrale, incluant le déplacement des

formants, les largeurs de bande, la pente spectrale et les différences notées sur la forme générale du spectre. Notons également que la qualité de voix n'est pas prise en considération ici.

Cette étude a pour but d'identifier lequel des deux groupes de facteurs est prépondérant dans la perception de l'effort vocal. Pour ce faire, nous avons voulu « greffer » sur une voix modale, dans un premier temps les paramètres spectraux d'une voix criée, et dans un second temps les paramètres prosodiques de celle-ci. Ceci nous permet alors, à l'image de l'étude de Tassa et Liénard (2000), d'estimer isolément la contribution des facteurs prosodiques ou spectraux, dans la perception auditive de l'effort vocal. Il existe tout de même trois différences entre notre étude et celle de Tassa et Liénard (2000). D'une part, notre étude se base sur des phrases parlées et criées, alors que Tassa et Liénard (2000) n'appliquent leurs transformations que sur des voyelles. D'autre part, ces auteurs n'ont malheureusement pas confirmé leurs dires par un test perceptif. La dernière différence, et non des moindres, est la méthode de synthèse utilisée. Tassa et Liénard (2000) utilise la méthode TD-PSOLA permettant de modifier la prosodie de la voix. De cette manière, la prosodie d'une voix est modifiée mais, la qualité de voix liée à la source glottique est préservée. Il nous semble alors que la préservation de la qualité de voix (ou du moins d'une partie de la qualité de voix) peut influencer les résultats. C'est pourquoi, dans notre étude, nous avons utilisé un vocodeur LP (excitation par une suite d'impulsions et bruit blanc). Celui-ci considère comme signal d'excitation, une suite d'impulsions espacées de  $T_0=1/F_0$  ou un bruit blanc en fonction du caractère voisé ou non du signal à synthétiser (cf. Figure 7-2). Nous avons choisi d'utiliser ce synthétiseur afin de considérer les paramètres prosodiques et spectraux, indépendamment de la source glottique (hors biais liée à l'estimation du filtre LP pouvant modéliser certains traits liés à la source glottique). Ainsi, nous ne considérons pas le résidu de l'analyse LP comme source d'excitation, mais nous recréons le vecteur d'excitation de façon indépendante.

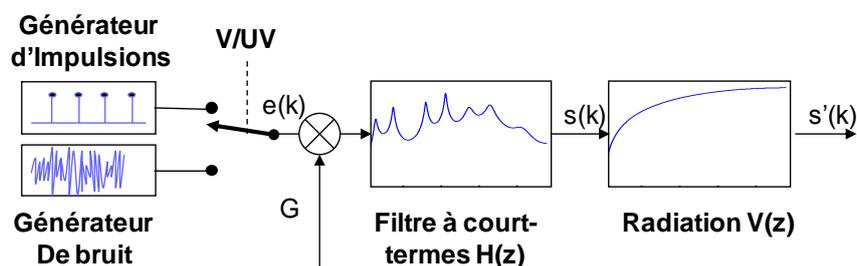


Figure 7-2: Schématisation du principe de synthèse du vocodeur LP

### 7.2.1 La base de voix utilisée

Afin de mener à bien cette étude, nous avons utilisé les enregistrements de voix du corpus DB1 et retenu les deux distances extrêmes ; à savoir, les enregistrements réalisés pour des distances de 5 m et de 100 m. En effet, nous considérons qu'avec une faible distance de communication (5 m), la voix du

locuteur est une voix normale, alors qu'à 100 m, sa voix reflète une augmentation significative des efforts vocaux. Rappelons que la base de voix utilisée pour cette étude comprend 3 locuteurs masculins et 5 phrases dont nous considérons 2 distances de communication (soit un total 15 couples de voix parlée-criée). Pour plus de détail se reporter au CHAPITRE 6.

### 7.2.2 Méthodologie

La première étape consiste à estimer les différents paramètres de chaque voix à intervalle régulier. Ainsi, toutes les 10 ms, une estimation du filtre du conduit vocal, de la fréquence fondamentale et de l'intensité, est effectuée. La reconstruction du signal est effectuée à l'aide d'un vocodeur à prédiction linéaire, suivant la technique de *l'overlap-add* (addition recouvrement), afin de réduire les discontinuités engendrées par la concaténation des trames. Concrètement, cela signifie que la fenêtre d'analyse est plus longue que son pas d'avancement. Cette fenêtre a une longueur de 20 ms pondérée par une fonction de *Hanning*, et est décalée par pas de 10 ms (soit un recouvrement de 50%). Le prédicteur du vocodeur est un filtre autorégressif tout-pôle d'ordre 18. Le coefficient  $a_v$  de la préaccentuation ( $V(z) = 1 - a_v z^{-1}$ ) est égal à 0,98.

Pour améliorer la qualité de la synthèse, les différents paramètres utilisés sont lissés entre deux trames successives, en particulier pour la période de la fondamentale. Ce lissage consiste à faire varier progressivement la période fondamentale d'une trame à l'autre selon une loi linéaire. Malgré sa simplicité, ce vocodeur permet de manipuler aisément et séparément les paramètres spectraux et prosodiques de la parole. Certes, ce vocodeur ne permet pas d'obtenir une très grande qualité de synthèse mais nous estimons que la qualité de la parole synthétisée est suffisante pour notre étude.

### 7.2.3 Dynamic Time Warping (DTW)

Afin de greffer les paramètres d'une voix sur un autre signal, une étape préliminaire est nécessaire. En effet, compte tenu des structures temporelles différentes des deux voix prises en compte (i.e voix parlée et voix criée), il est indispensable de réaliser un alignement temporel. Afin de confondre les filtres LP pour chaque fenêtre d'analyse ainsi que la valeur de F0 et d'intensité d'un signal à l'autre il est primordial de faire correspondre entre les analyses de ces deux signaux les éléments d'une même zone. Par exemple sur une voyelle il existe plusieurs zones correspondant à une zone d'établissement, de maintien et de relâchement. Celles-ci peuvent, et sont en général, différentes d'un enregistrement à l'autre et d'autant plus lorsqu'on considère des types de phonation différente. Ainsi pour faire correspondre les fenêtres d'analyse de chaque signal (et ainsi le vecteur de paramètres propre à la fenêtre) il est nécessaire de réaliser un alignement temporel entre ces deux vecteurs temporelles. Pour ce faire, nous utilisons la technique appelée *Dynamic Time Warping* (DTW) (Sakoe et Chiba, 1978) ou

déformation temporelle dynamique. Dans cette méthode, la similarité entre les différentes trames des deux voix est assurée à l'aide du calcul d'une distance spectrale. Le meilleur alignement consiste à trouver une correspondance entre ces trames minimisant les distances spectrales.

Concrètement cette méthode consiste à segmenter les deux signaux en fenêtres d'analyses ( $a(n)$  : voix parlée et  $b(n)$  : voix criée) et de calculer une distance spectrale locale  $d(i,j)$  pour l'ensemble des combinaisons possibles entre ces fenêtres (cf. Eq. 7-1). Une distance spectrale minimale signifie un maximum de ressemblance spectrale. Cette méthode de calcul de distance a été proposée par Dan Ellis<sup>1</sup>.

$$d(i, j) = \frac{\sum_{k=1}^{N/2} |A_i(k)|^2 \cdot |B_j(k)|^2}{EA_i \cdot EB_j} \quad i = 1, \dots, I; j = 1, \dots, J. \quad (\text{Eq. 7-1})$$

Dans l'équation 7-1,  $I$  et  $J$  sont respectivement le nombre total de fenêtre d'analyse des signaux  $a(n)$  et  $b(n)$ .  $A_i$  et  $B_j$  sont respectivement les transformées de Fourier (FFT) des  $i^{\text{ème}}$  et  $j^{\text{ème}}$  fenêtres d'analyses de  $a(n)$  et  $b(n)$ .  $EB$  et  $EA$  sont les énergies totales de  $A_i$  et  $B_j$  et  $N$  la longueur de la FFT. L'équation 7-1 calcule ainsi une matrice de distances spectrales croisées. Pour chercher le meilleur alignement entre le signal  $a(n)$  et le signal de référence  $b(n)$  il faut alors chercher le meilleur chemin dans cette matrice  $d(i,j)$  qui minimise la somme des distances locales pour aller d'un point initial ( $i=I, j=J$ ) à un point final ( $i=1, j=1$ ). On peut facilement démontrer que la somme des distances minimales locales est égale à la distance minimale globale. Ainsi le chemin (ou distance) optimal  $D$  se calcule de la manière suivante :

$$D(i, j) = d(i, j) + \min[D(n-1, j), D(n, j-1), D(n-1, j-1)] \quad (\text{Eq. 7-2})$$

De cette manière, on contraint le chemin  $D$  à suivre un trajet « monotone » et « plausible », tout en suivant le chemin le plus court. Ainsi, grâce au vecteur  $D$ , il est possible d'associer une fenêtre d'analyses de  $a(n)$  à son équivalent spectral (le même son) sur le signal  $b(n)$ . En pratique, on crée une seconde matrice de même dimension que  $d$  et dans laquelle on mémorise pour chaque entrée ( $i,j$ ) un pointeur vers le prédécesseur optimal. C'est-à-dire vers l'un des éléments du terme  $\min[D(n-1, j), D(n, j-1), D(n-1, j-1)]$ . Sur la Figure 7-3, on peut voir les deux signaux à aligner ( $a(n)$  : voix parlée, et  $b(n)$  : voix criée), la matrice des distances spectrales croisées  $d$ , et représentée par les croix blanches, le chemin optimal  $D$ .

<sup>1</sup> Source : <http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw/simmx.m>

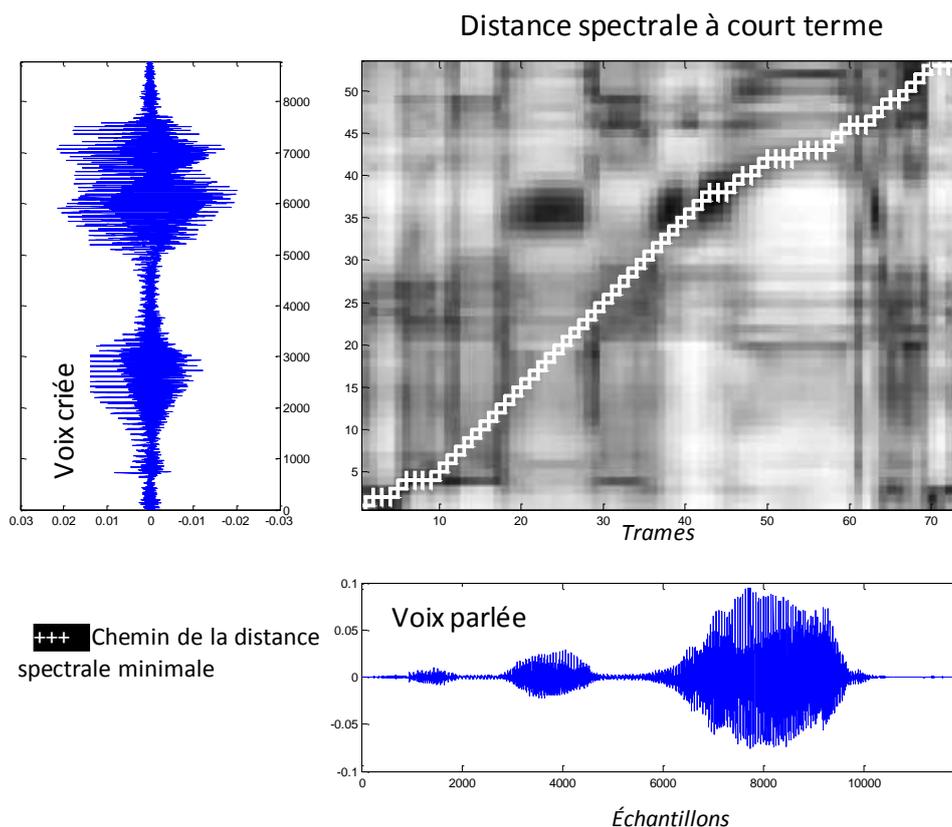


Figure 7-3 : Exemple de l'analyse DTW. Le chemin représenté par des croix blanches représente la courbe d'alignement entre les fenêtres d'analyse des deux signaux considérés.

## 7.2.4 Génération des signaux tests

Disposant à présent des estimations des paramètres nécessaires à la reconstruction de la voix effectuée toutes les 10 ms, d'un synthétiseur ainsi que de l'information concernant l'alignement, la dernière étape consiste à générer les signaux tests.

Pour tester la pertinence des paramètres spectraux d'une voix criée sur une voix parlée, on conserve la prosodie de celle-ci (structure temporelle, évolution de la  $f_0$ , évolution de l'intensité) et on lui « greffe » les paramètres spectraux de la voix criée en utilisant les correspondances établies avec la DTW. Notons que la valeur moyenne de  $F_0$  est également altérée mais que l'intensité moyenne est inchangée. Après la synthèse, on obtient une voix mixte appelée *Smod\_spec* (cf. Figure 7-4).

Pour mettre en évidence le rôle des paramètres prosodiques d'une voix criée, on utilise ces paramètres pour la synthèse, tout en conservant les paramètres spectraux initiaux à la voix normale (toujours via la DTW). La synthèse nous donne une voix mixte appelée *Smod\_pros* (cf. Figure 7-4).

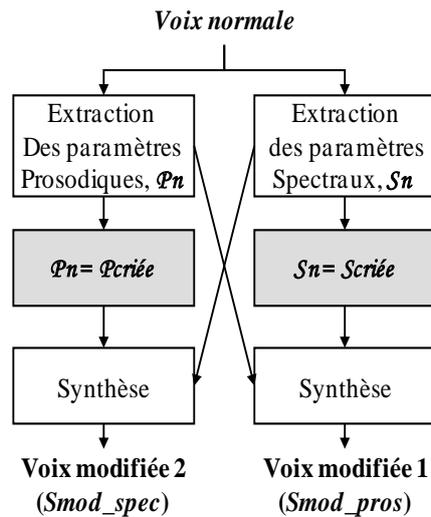


Figure 7-4: Principe de la modification de la voix par « matching »,  $S_n$  et  $S_{criée}$  sont respectivement les paramètres spectraux de la voix parlée et de la voix criée.  $P_n$  et  $P_{criée}$  correspondent aux paramètres prosodiques de la voix parlée et de la voix criée.

La Figure 7-5 donne un exemple des résultats de cette manipulation. Sur chacune des figures sont représentés, le contour de la F0 (en bleu) ainsi que le suivi des 3 premiers formants (en pointillés rouges). Toutes les figures correspondent au même locuteur, ainsi qu'à la même phrase. La figure a) correspond à la phrase parlée pour une distance de 5 m. La figure b) correspond à la voix criée pour une distance de 100 m. La figure c) correspond à la voix parlée où les paramètres spectraux de la voix criée ont été « greffés ». Enfin, la figure d) correspond à la voix parlée où les paramètres prosodiques de la voix criée ont été « greffés ». On observe clairement que les durées et la variation de d) ( $S_{mod\_pros}$ ) correspondent à celles observées sur b) ( $S_{criée}$ ), tout en concevant les variations spectrales observées pour a) ( $S_{parlée}$ ). De la même manière, c) ( $S_{mod\_spec}$ ) conserve les variations prosodiques de a) ( $S_{parlée}$ ) et les variations spectrales de b) ( $S_{criée}$ ).

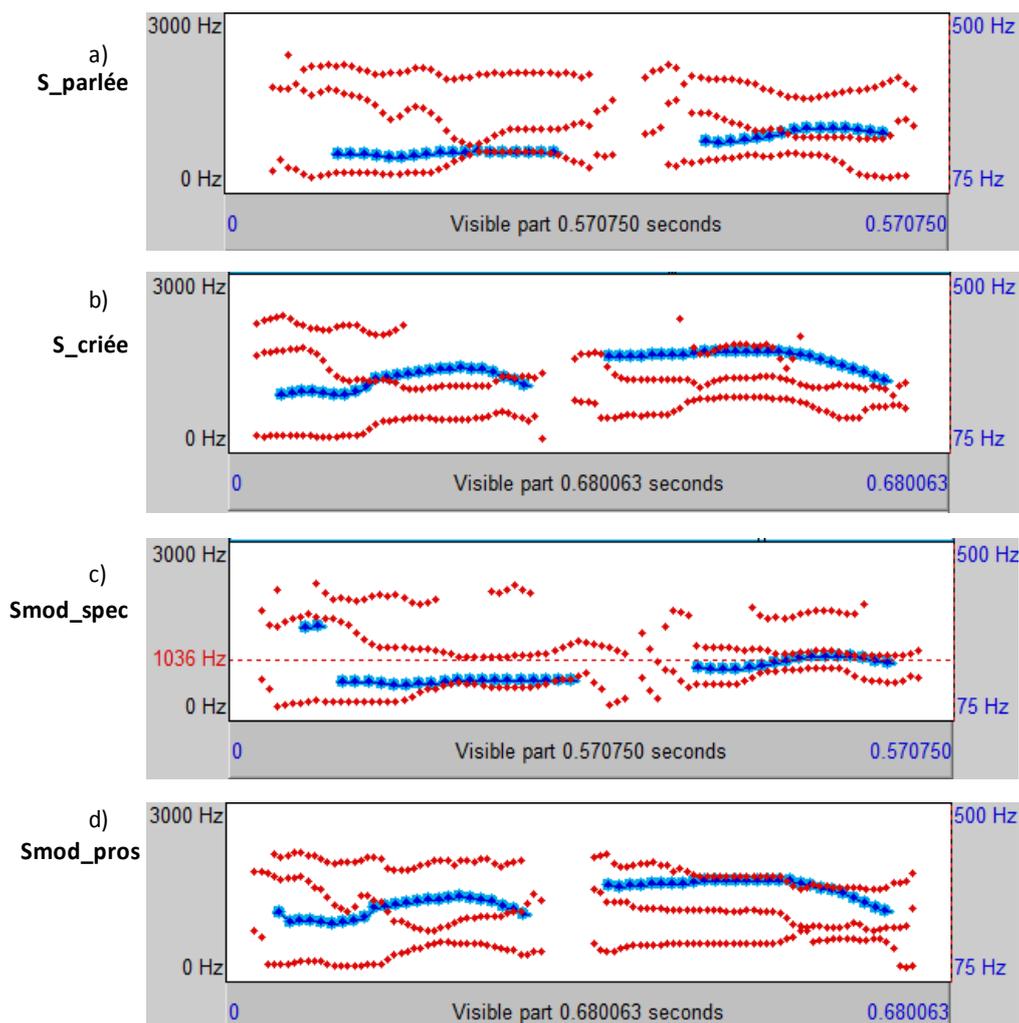


Figure 7-5 : suivi des contours de la phrase « tu m'entends ? » de F0 et des trois premiers formants pour les voix parlées (a) et criées naturelles (b) ainsi que pour la voix parlée modifiée en accord avec le spectre de la voix criée (c) ou la prosodie de la voix criée(d).

### 7.2.5 Test perceptif

Un test perceptif a été réalisé pour étudier l'apport de chaque jeu de paramètres dans la perception de l'effort vocal. Il s'agit de comparaisons en terme de ressemblance. Les voix d'origine sont également re-synthétisées afin d'être insérées dans le test. De cette manière, les voix de références contiennent les mêmes artefacts (liés à la méthode de synthèse) que les voix parlées qui ont été modifiées. De plus nous pensons que ce choix facilite le test pour les sujets. En effet, ces derniers n'ont alors pas besoin de faire un effort d'abstraction concernant le naturel des voix de synthèse afin de réaliser des comparaisons objectives avec les voix naturelles. Notons que l'ensemble des signaux ont été égalisés en intensité pour ce test.

Pour un locuteur donné et une phrase donnée deux voix modifiées ont été créées (*Smod\_pros*, *Smod\_spec*) à partir des enregistrements de la voix parlée et de la voix criée du locuteur pour la même

phrase. Le test consiste alors à comparer chaque voix modifiée aux voix d'origines, à partir desquelles elles ont été créées. Le test se présente de la manière suivante : on fait entendre aux sujets successivement, et une seule fois, la voix parlée, la voix modifiée 1 (i.e. *Smod\_spec*) puis la voix criée. Un deuxième triplet de stimuli est également entendu pour la voix modifiée 2 (i.e. *Smod\_pros*). Ainsi, nous obtenons deux triplets de stimuli pour chaque locuteur et pour chaque phrase. Ceci constitue une base de trente triplets à comparer. Lors de l'écoute, les phrases ont été séparées par une pause de 300 ms. L'ensemble des trente triplets de stimuli ont été comparés par chaque sujet dans un ordre d'apparition aléatoire. Les sujets devaient alors, pour chaque triplet de stimuli, répondre à la question : *à laquelle des deux références ressemble le plus la phrase modifiée (c'est-à-dire la phrase du milieu)?* Ce test perceptif s'est déroulé dans une chambre d'audiométrie afin que les sujets ne puissent être perturbés par du bruit ambiant.

### 7.2.6 Résultats

Le Tableau 7-1 donne les résultats du test subjectif des dix-huit sujets. Chaque sujet a réalisé 30 comparaisons. C'est-à-dire que chaque sujet a comparé l'ensemble des voix modifiées (prosodiquement et spectralement) pour un locuteur donné. Ainsi, chaque série de voix modifiées concernant un locuteur, a été entendue et comparé 6 fois, soit un nombre total de comparaison égal à 540 (270 pour *Smod\_pros* et 270 pour *Smod\_spec*).

La première voix modifiée (*Smod\_pros*) est clairement perçue comme étant très similaire à une voix criée. On suppose que ceci est due au fait qu'un effort vocal significatif est perçu. Cette voix est alors perçue comme une voix produite dans le but de pallier une grande distance (*Voix criée*) dans 86% des cas. La seconde voix synthétisée (*Smod\_spec*) ne donne pas la même perception de l'effort vocal. Elle est plutôt perçue comme une voix produite dans un but de communication à faible distance (*Voix parlée*) dans 93% des cas.

Tableau 7-1: Ressemblance des phrases modifiées par rapport aux phrases de référence.

	Voix parlée	Voix criée
<b>Smod_pros</b>	14%	86%
<b>Smod_spec</b>	93%	7%

Notons également que le facteur *Locuteur* ne semble pas avoir une influence déterminante pour ces résultats. Bien que les valeurs soient différentes pour les trois locuteurs, les ordres de grandeurs ainsi que la tendance globale des résultats sont similaires.

### 7.2.7 Discussion

Ces résultats montrent clairement que les paramètres spectraux à eux seuls, contrairement aux paramètres prosodiques, ne contribuent pas de manière significative à la perception des efforts vocaux. Ainsi, nous pouvons dire que les paramètres prosodiques sont prédominants dans la perception des efforts vocaux. Soulignons que ceci ne nie pas pour autant la contribution globale des paramètres spectraux. En effet, leur apport n'est pas à négliger du fait que l'effort vocal ne se reflète qu'à travers un tout. De toute évidence, les voix modifiées manquent de naturel. La perte de qualité due au synthétiseur, combinée aux modifications (prosodiques ou spectrales), engendrent des voix qui ne semblent pas être produites par l'être humain. Les paramètres prosodiques et spectraux étant étroitement liés, modifier un seul jeu de paramètres, génère des phrases qui ne sont pas naturelles. Les voix obtenues restent néanmoins parfaitement intelligibles et ne sont pas désagréables à entendre. De plus, les sujets qui ont participé à l'expérience ont jugé ne pas avoir été perturbés par l'écoute des voix modifiées. De ce fait, nous pensons que la qualité des voix synthétisées n'a pas biaisé les résultats obtenus.

Rappelons aussi que l'objectif de cette étude n'est pas de quantifier de manière absolue l'apport de chacun de ces groupes de paramètres, mais que nous cherchons à mettre en évidence l'apport relatif à chacun d'eux. Ces résultats nous indiquent clairement qu'une plus grande prise en considération des variations prosodiques par rapport aux variations spectrales est nécessaire pour la perception de la distance par le biais de l'effort vocal.

## 7.3 Conclusions du chapitre 7

---

Ce chapitre a pour but de restreindre le champ des analyses concernant les paramètres liés à l'effort vocal. En effet, le nombre de paramètres variant avec l'effort vocal est considérable. C'est pourquoi réduire le champ de notre étude, par l'identification des paramètres pertinents pour la perception de l'effort vocal, permet de rendre plus efficace notre étude ainsi que de faciliter les différents choix qui seront effectués lors de l'élaboration des méthodes de transformation de voix.

Nous avons ainsi dans ce chapitre, établi un état de l'art concernant les éléments pertinents liés à la perception de l'effort vocal. Dans un premier temps nous avons repris la bibliographie directement dédiée à l'identification des paramètres perceptivement pertinents. La littérature sur ce sujet fait ressortir une tendance sur le fait que la dimension prosodique constitue l'élément le plus pertinent pour la perception de l'effort vocal de projection, bien que certains auteurs proposent la distance tonotopique entre F0 et F1 ou encore les paramètres de source glottique (mais dans de moindre

mesure). La littérature sur ce sujet reste toutefois faible. En effet, seules 4 études traitent directement du sujet. C'est pourquoi, nous avons listé les différentes méthodes et paramètres modifiés lors d'études consacrées à la modification de la perception de l'effort vocal de la voix, qui elles, ne se concentrent pas directement sur la problématique de la pertinence.

Au cours de cette deuxième partie, nous avons alors mentionné des études qui modifient chacune d'entre elles des paramètres différents et qui, en finalité, correspondent à presque tous les aspects de la voix. On y trouve ainsi des modifications de la source glottique essentiellement, des modifications spectrales ou encore des méthodes de modification sur la base de modèles statistiques, ne permettant pas d'apporter d'informations majeures sur le sujet. En revanche, pour la voix chuchotée, la littérature s'accorde à attribuer le dévoisement comme caractéristique principale de la voix chuchotée et mentionne que certains ajustements spectraux sont nécessaires, afin d'apporter une bonne qualité de transformation.

Ainsi, au regard de cet état de l'art, la question que nous avons posée en début de ce chapitre : « Qu'est ce qui permet à l'Homme de différencier une voix parlée d'une voix criée ou encore d'une voix chuchotée ? » reste toujours sans réponse précise pour le cas des voix criées.

C'est pourquoi, afin d'éclaircir ce point crucial, nous avons mené une étude visant à déterminer qui d'entre les paramètres prosodiques ou les paramètres spectraux contribuent le plus dans la PAD ou plus précisément dans la perception de l'effort vocal.

Pour ce faire, les enregistrements faits pour des distances de communication de 5 m et 100 m, du corpus DB1, comprenant 3 locuteurs, ont été utilisés. Un vocodeur à prédiction linéaire avec un algorithme de *matching* (DTW) a été implanté, permettant ainsi de combiner les paramètres spectraux et prosodiques des deux types de voix. Dans une première voix ainsi synthétisée, on combine la prosodie de la voix de 100 m avec les paramètres spectraux de la voix normale (5 m), tandis que dans une seconde voix c'est l'inverse. Un test perceptif, basé sur des comparaisons en termes de ressemblance, a été effectué par 18 sujets. Il en ressort que, dans 86% des cas, les voix synthétiques basées sur les paramètres prosodiques de la voix à 100 m sont identifiées comme les voix distantes (100 m), tandis que les voix synthétiques avec les paramètres spectraux de la voix à 100 m sont perçues comme étant proches (5 m) et ce dans 93% des cas. Ce résultat montre ainsi clairement la prédominance de la prosodie dans la perception de la distance d'un locuteur. Cette étude confirme alors les conclusions faites par Tassa et Liénard (2000), Brungart et al. (2002) et Yi et al. (2000) :

*La prosodie constitue l'élément le plus pertinent dans la perception de l'effort vocal.*

# Dynamique des paramètres et effort vocal

---

— ” [...] loudness involves the spectrum of the sound, not just the amount of sound.”  
— *Ingo R. Titze*

La littérature concernant les éléments pertinents dans la PAD (perception auditive de distance) d'un locuteur (et ainsi que de la perception de l'effort vocal), ainsi que nos propres expériences nous indiquent que la prosodie (variations de la F0, de l'intensité et de la durée) joue un rôle plus pertinent que les paramètres spectraux (paramètres glottiques, formants, etc.). Le chapitre précédent nous a permis d'affirmer ce point. Cependant les études précédentes ne nous permettent pas de quantifier les variations prosodiques en fonction de l'effort vocal. Or, pour réaliser une transformation de voix parlée en voix criée, il nous faut des règles précises décrivant les variations des paramètres. Pour pallier le déficit concernant ce sujet dans la littérature, nous avons mené une étude approfondie à partir de nos corpus de voix. Notre objectif principal est de savoir quels sont les paramètres prosodiques pertinents dans la voix criée. Notons que dans ce chapitre nous ne nous intéressons pas à la voix chuchotée, dont les variations ainsi que les éléments pertinents à sa perception sont bien connus.

Ainsi, dans ce chapitre, nous nous intéressons dans un premier temps aux valeurs moyennes de l'intensité, ainsi que celle de la F0, pour l'ensemble des corpus que nous avons enregistrés (cf. CHAPITRE 6). Cette première étape a pour but de juger de la pertinence entre les différents corpus enregistrés (situation réelle de jour et de nuit vs. simulation de communication à distance en laboratoire). Ces relevés nous permettent d'ailleurs de valider le protocole de simulation que nous avons développé et permettent également de valider l'augmentation de l'effort vocal avec l'augmentation de la distance de communication.

Dans une deuxième étape, nous étudions les dynamiques de F0 et de l'intensité de ces corpus. Ce qui permettra, là encore d'apporter des informations quant à la validité du protocole de simulation. D'autre part, des différences significatives de ces dynamiques viendront également étayer l'hypothèse que la dynamique de la F0 et celle de l'intensité sont plus fortes en voix criée (Fux et al., 2011a, 2011b). En effet, l'augmentation de la dynamique de la F0 n'est pas systématiquement observée dans la littérature.

La dynamique de F0 est pour nous un paramètre important qui doit être considéré. Ainsi, nous avons mené plusieurs analyses permettant de faire nos propres conclusions sur le sujet. Nous mentionnons d'ailleurs une méthode de mesure (via la mesure de la différence entre les extrema) qui nous semble être plus révélatrice que celle mentionnée dans la littérature (mesure de l'écart-type, variance).

Les observations faites sur la dynamique n'apportent qu'une faible partie d'information sur l'évolution globale des contours de F0 et d'intensité, ce qui n'est pas encore suffisant pour leur description précise et ainsi leur modélisation. Afin de compléter ce point, une analyse prosodique plus poussée est réalisée sur le corpus enregistré dans cet objectif précis (i.e. le corpus en champ libre de nuit, DB2). Nous nous concentrons sur ces enregistrements car il s'agit d'une situation de communication réelle. De plus, parmi les deux corpus enregistrés en champ libre, le corpus DB2 contient 6 locuteurs alors que l'autre corpus, qui a également été enregistré en extérieur (DB1), n'en comporte que 3.

Enfin, nous évoquerons rapidement la dynamique de la fréquence des formants observés, et particulièrement la relation entre la méthode de mesure et la précision de l'estimation des formants.

## **8.1 Analyses de l'intensité et de la F0 globales**

---

Dans cette section, nous nous consacrons aux valeurs moyennes de l'intensité, de la F0, des différents corpus. Dans un premier temps ceci permet de juger de la vraisemblance des différents corpus que nous utilisons, et notamment de comparer le protocole de simulation avec les corpus en champ libre. De plus, ces analyses nous permettent de confirmer l'augmentation de l'effort vocal avec l'augmentation de la distance de communication. La mesure de l'intensité a été effectuée toutes les 10 ms puis la moyenne d'intensité pour chaque énoncé est retenue.

### **8.1.1 Intensité globale**

Comme nous le savons, l'intensité de la voix augmente à mesure que la distance augmente (cf Figure 8-1). On observe cependant des différences entre nos 4 corpus. Le corpus DB1 possède des valeurs d'intensité inférieures aux trois autres corpus, ce qui était prévisible. En effet, DB1 est le seul corpus dont la voix des locuteurs a été enregistrée à environ 1 mètre face à lui. Les voix, dans les autres corpus ont été enregistrées par un microphone de bouche. Ainsi, la différence d'intensité entre DB1 et les autres corpus, semble s'expliquer par la position du microphone. Pour DB1 les intensités sont moins stables d'une distance à l'autre, ce qui peut s'expliquer par la variation entre la distance du locuteur et du microphone.

La comparaison entre les corpus réalisés via le protocole de simulation et les corpus réalisés en champ libre de nuit, montre que l'intensité de DB2 n'est pas aussi forte que les intensités simulées pour la même distance. En première approche, ce phénomène est attribué au protocole d'enregistrement. Toutefois, rappelons que le corpus DB2 a été enregistré de nuit dans une prairie. Ainsi, les conditions atmosphériques étaient optimales pour la propagation de l'onde acoustique (peu de bruit de fond, gradient de température, cf. CHAPITRE 6). En effet, nous avons pu remarquer lors de ces enregistrements, que l'effort vocal nécessaire pour communiquer à une distance de 100 m, est moins important de nuit que de jour.

L'intensité des voix du corpus DB3 croît rapidement jusqu'à des intensités très fortes (100 dB SPL) mais semble toutefois être cohérent avec la distance de communication de 120 m. L'intensité pour le deuxième corpus simulé (DB4) croît encore plus rapidement. Cette augmentation s'explique par la difficulté de l'exercice lors de l'enregistrement. Rappelons que le locuteur, placé dans une pièce dont l'atténuation acoustique est de 36 dB, doit communiquer une suite de logatomes de type CV, CVC, VCV ou encore CVCV à un auditeur placé à l'extérieur. Ainsi, il semble naturel de devoir fournir plus d'effort afin de faire comprendre des mots qui n'existent pas et où l'ensemble des phonèmes doit être compréhensible.

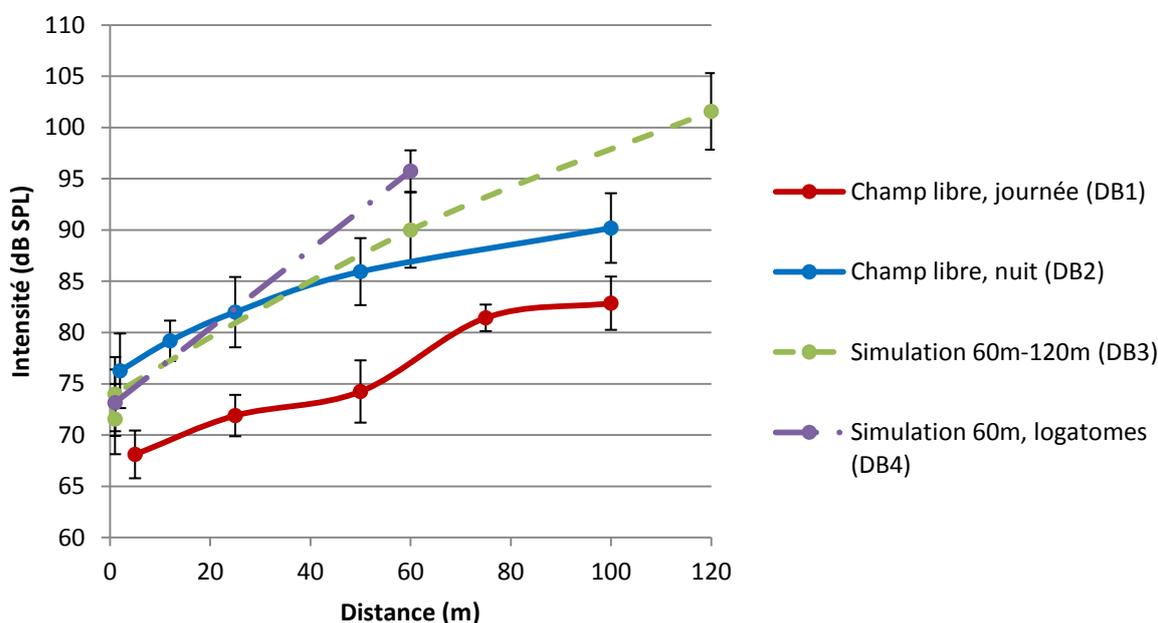


Figure 8-1: Intensité des voix en fonction de la distance de communication et du corpus

### 8.1.2 Fréquence fondamentale

Penchons-nous à présent sur l'évolution de la moyenne de F0 en fonction de la distance de communication (cf. Figure 8-2). Contrairement à la mesure de l'intensité, l'estimation de la F0 n'est pas sensible à la position du microphone.

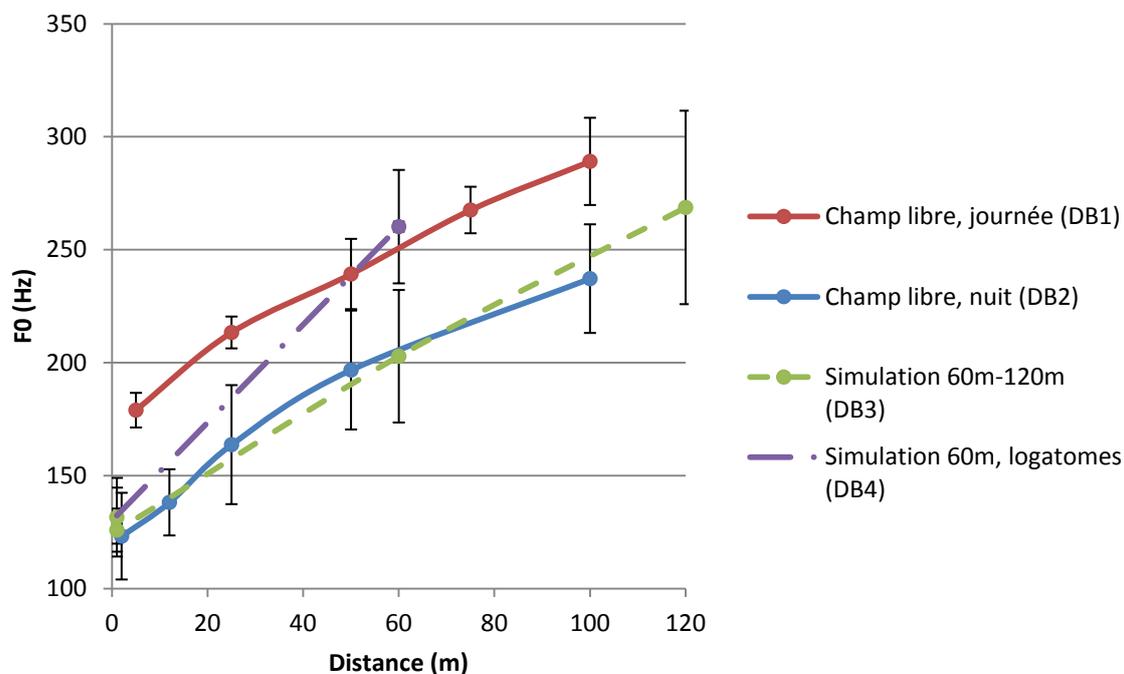


Figure 8-2 : F0 des voix en fonction de la distance de communication et du corpus

La première constatation est que la valeur mesurée pour le protocole en champ libre de nuit (DB2) correspond aux valeurs mesurées lors du protocole de simulation DB3. On constate cependant que la F0 du protocole en champ libre (DB1) est plus élevée. Ce phénomène peut s'expliquer en considérant le milieu de propagation. En effet, le protocole DB1 a été enregistré de jour en extérieur et le DB2 de nuit en extérieur. Ainsi, comme pour l'intensité, du fait d'une meilleure propagation du son la nuit, moins d'effort est nécessaire pour communiquer à distance. De ce fait, en pleine journée, dû au bruit de fond plus fort, les valeurs de DB1 sont plus intenses que pour DB2.

Une autre remarque concerne le protocole DB4. Dans ce corpus, qui utilise le protocole de simulation pour une distance de 60 m, on constate que les valeurs de F0 des voix criées sont aussi fortes que les valeurs obtenues pour une distance de communication en champ libre de jour (DB1) pour la même distance. Nous pensons que ce phénomène vient de l'utilisation des logatomes. En effet, les logatomes en absence totale de sens, sont plus difficiles à comprendre. Ainsi, pour le même protocole de simulation que DB3, les valeurs de F0 sont plus fortes du fait que les locuteurs fournissent plus

d'effort afin que les logatomes soient parfaitement intelligibles. En effet, pour le cas de DB3 il s'agissait de phrases qui étaient connues à l'avance. Toutefois, il est difficile de confirmer cet effet étant donné qu'aucune comparaison directe ne peut être effectuée entre ces corpus du fait que les messages et/ou les locuteurs sont différents d'un corpus à l'autre. Néanmoins, au regard de ces valeurs, le protocole de simulation semble être approprié et correspondre à la réalité.

## 8.2 Dynamique des paramètres vs effort vocal

La littérature mentionne des variations de la dynamique de la F0. Mais il semble que cette augmentation de la dynamique ne soit pas systématique (Jessen et al., 2005). Au regard de nos enregistrements, nous constatons rapidement des dynamiques plus fortes sur la F0 ainsi que sur l'intensité. Sur Figure 8-3, sont représentés les cinq signaux correspondant aux cinq distances de communication pour la phrase « Lève le bras. » du locuteur 1 du corpus nocturne enregistré en extérieur DB2. Sur cette figure, sont également représentés les contours de F0 ainsi que les contours de l'intensité. On constate alors de fortes variations sur l'évolution de la F0 à partir d'une distance de communication de 25 m. Les contours de F0 pour de grandes distances, présentent des forts ventres et creux. De la même manière, l'intensité présente également des variations plus fortes pour des grandes distances de communication. On se demande toutefois si ces variations sont significatives.

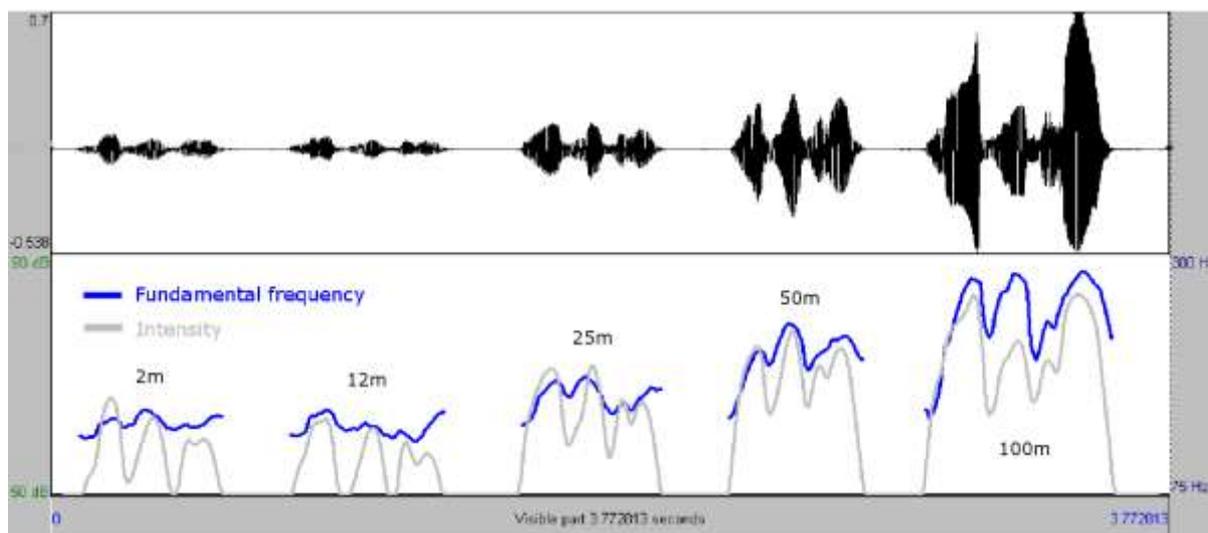


Figure 8-3 : Exemple pour la phrase 1 et le locuteur 1 pour les 5 distances de communication. En haut on observe le signal de la parole, en bas le suivi du contour de F0 en bleu et de l'intensité en gris (les valeurs de l'intensité sont arbitraires)

Pour répondre à cette question nous avons alors mesuré la dynamique de F0 et de l'intensité pour l'ensemble des phrases des corpus dont nous disposons (hormis DB4 qui ne contient pas de phrases et

dont les dynamiques seront mesurées plus précisément dans le CHAPITRE 9). D'autre part, la dynamique est couramment mesurée par le biais de l'écart-type des contours de F0 ou de I. Des variantes consistent également à normaliser ces écarts-types par la moyenne du contour (Jessen et al., 2005). Cependant, cette méthode nous paraît peu appropriée pour refléter réellement la dynamique de ces contours. C'est pourquoi nous avons proposé une nouvelle méthode de mesure.

### 8.2.1 Méthode proposée pour la mesure de la dynamique

Afin de mesurer la dynamique des paramètres nous faisons le choix de considérer uniquement les valeurs extrêmes le long d'un contour. Théoriquement, il suffit alors de localiser ces valeurs par le biais de la dérivée (instants correspondant au passage par zéro de la dérivée). Cependant, dans le cas des contours d'une voix réelle, ces contours présentent beaucoup de micro-variations. Dans le but de réduire ces micro-variations, les contours sont préalablement lissés par une fenêtre glissante d'une largeur de 100 ms (i.e. par un filtre moyenneur glissant). Par la suite, les maxima et minima du contour sont localisés (cf. Figure 8-4). Ceci permet alors de ne considérer que les macro-variations du contour. A partir de ces valeurs, la dynamique correspond à la moyenne des différences entre les points adjacents. Ainsi, si N extrema ont été localisés ( $f0(x_1) \dots f0(x_N)$ ) la dynamique  $\delta x$  se calcul de la manière suivante :

$$\delta x = \frac{1}{N} \sum_{i=1}^{N-1} |f0(x_i) - f0(x_{i+1})| \quad (\text{Eq. 8-1})$$

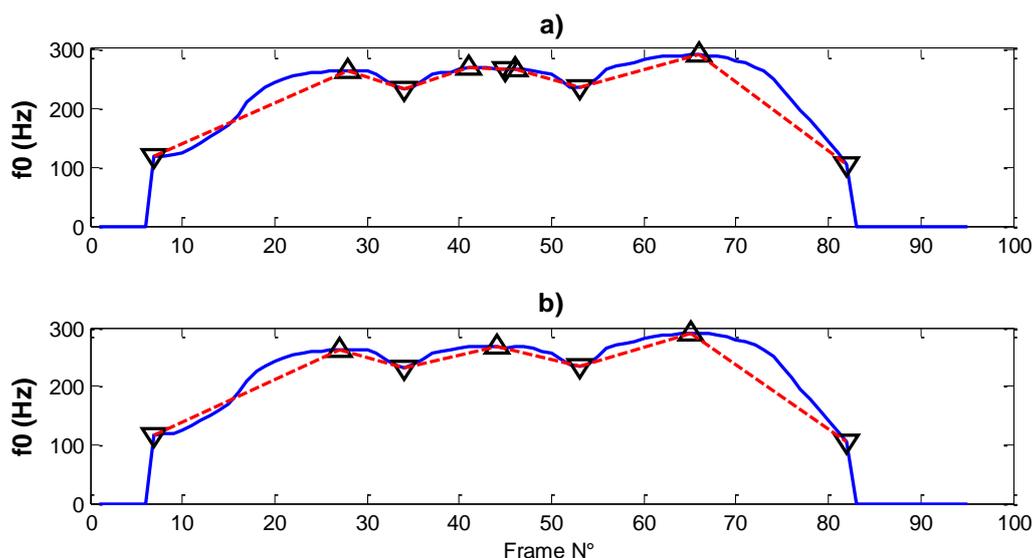


Figure 8-4: Exemple de localisation des extrema d'un contour de F0 avant (en haut) et après (en bas) le lissage. Les pointillés représentent les écarts entre deux valeurs successives prises en compte pour le calcul de la dynamique.

Notons également que pour la mesure de la dynamique de F0 les valeurs inférieures à 50 Hz ne sont pas considérées. De la même manière, pour l'intensité, les valeurs inférieures à 40 dB SPL ne sont pas prises en compte.

Rappelons que cette méthode de mesure est fondamentalement différente de la mesure de l'écart-type. Cette méthode de mesure ne tient pas compte d'éventuel plateau sur le contour de la F0. On mesure donc bien la dynamique de la courbe et non pas son écart-type par la variabilité d'une série de données.

Prenons l'exemple d'une fonction sinus définie entre 0 et  $\pi$  et d'amplitude maximale égale à 1. L'écart-type de cette courbe vaut 0,3. Par la mesure de dynamique que nous proposons la valeur est égale à 1. Ainsi cette méthode permet une mesure efficace de la dynamique moyenne d'un contour de F0 (ou d'intensité) qui ne tiens pas compte du nombre de points sur la courbe et qui, de ce fait, donne des résultats différents de la mesure de l'écart-type. En effet, on ne considère pas ici une série de valeurs dont nous voudrions connaître la variation moyenne, mais bien un contour dont nous souhaitons connaître la variation moyenne.

## 8.2.2 Analyses prosodique de DB2

Nous avons réalisé l'enregistrement DB2 dans le but de réaliser des analyses concernant les variations prosodiques globales. Ainsi nous avons effectué une étude regroupant plusieurs mesures dans le but de mieux comprendre les phénomènes globaux d'augmentation de la dynamique observée sur l'ensemble des corpus (Fux et al., 2011b). Pour mieux comprendre ces phénomènes nous analysons séparément les phrases interrogatives et les phrases déclaratives.

### 8.2.2.1 Dynamique de la F0 et de l'intensité

On observe sur la Figure 8-5 que la dynamique de l'intensité ainsi que la dynamique de F0 augmentent avec la distance de communication. La dynamique d'intensité est presque doublée entre la distance de 5 m et la distance de 100 m. Concernant la dynamique de F0, la valeur de la dynamique est également presque 2 fois plus élevée. D'autre part, la dynamique de F0 est plus intense pour les phrases interrogatives. Ces valeurs sont représentatives de l'intonation d'une phrase interrogative. En effet, le contour de F0, pour une phrase interrogative, présente une montée en fin d'énoncé qui permet l'identification de la forme interrogative (Delattre, 1966).

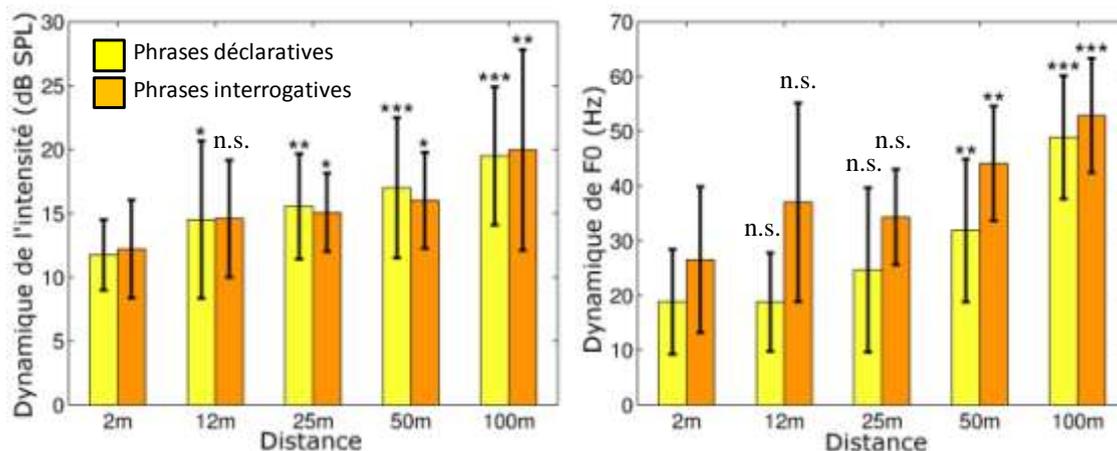


Figure 8-5: Valeurs moyennes de la dynamique de l'intensité (à gauche) et de la F0 (à droite) pour les phrases de DB2. Les mesures sont réalisées séparément pour les phrases interrogatives et les phrases exclamatives. Les tests de significativité ont pour référence les valeurs pour la distance de 2 m.

### 8.2.2.2 Analyses de l'évolution de la fréquence fondamentale

La dynamique de F0 peut en partie être expliquée par la mesure des valeurs minimales et maximales de F0 au cours d'une phrase. On observe alors sur la Figure 8-6 que les valeurs minimales et maximales de F0 augmentent avec la distance de communication. Cependant, les évolutions sont moins prononcées pour les phrases déclaratives. D'autre part, les valeurs maximales augmentent plus rapidement que les valeurs minimales. Ce phénomène peut d'ailleurs s'observer sur la Figure 8-3. En effet, les différences entre  $F0_{max}$  et  $F0_{min}$  sont pour les phrases interrogatives, et les distance de 2 m, 12 m, 25 m, 50 m et 100 m respectivement de 50, 60, 74, 86 et 132 Hz, alors que pour les phrases déclaratives les différences sont de 33, 40, 44, 68 et 96 Hz.

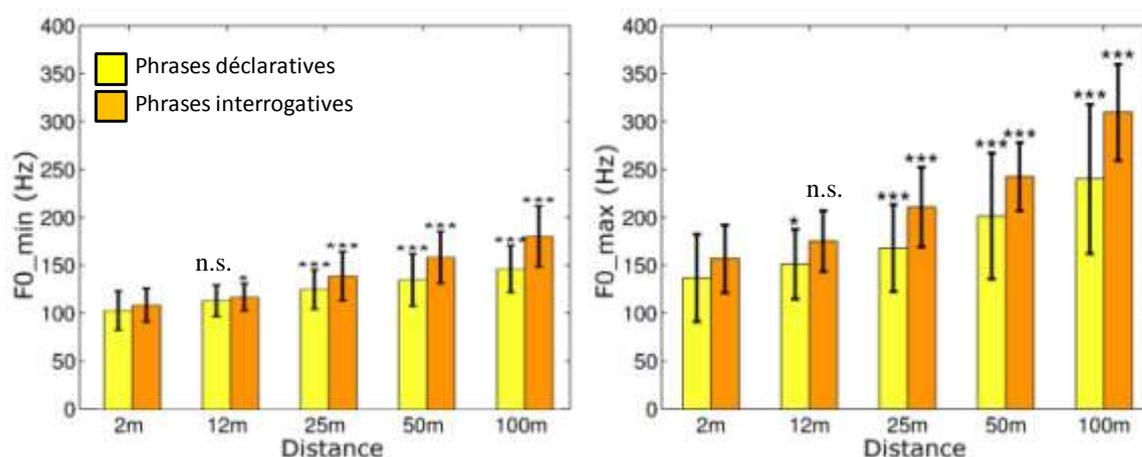


Figure 8-6: Valeurs minimales et maximales de F0 pour les phrases de DB2. Les mesures sont réalisées séparément pour les phrases interrogatives et les phrases exclamatives. Les tests de significativité ont pour référence les valeurs pour la distance de 2 m.

La Figure 8-7 représente la déclinaison de F0 pour le corpus DB2. Ces valeurs correspondent aux coefficients directeurs d'une droite de régression calculée sur les contours de F0 (en excluant les parties non voisées de chaque phrase). On note ainsi, que pour les phrases déclaratives en voix modale (2 m), la déclinaison est nulle et que comme attendu, pour les phrases interrogatives, les valeurs sont positives. Ce qui indique une montée progressive de F0 le long de l'énoncé. Ces valeurs sont parfaitement en accord avec la littérature (Delattre, 1966). Néanmoins, on constate que plus la distance de communication est élevée, plus la droite de régression est raide. Entre la voix modale (2 m) et la voix la plus criée (100 m), la raideur de cette droite est augmentée d'un facteur 2,9 pour les phrases interrogatives et d'un facteur 11,6 pour les déclaratives ! Ces relevés signifient alors que si pour une phrase déclarative modale la droite de régression de la F0 est constante, en voix criée, la F0 augmente tout au long de l'énoncé. Pour les interrogatives le phénomène de montée de F0 finale est accentué en voix criée.

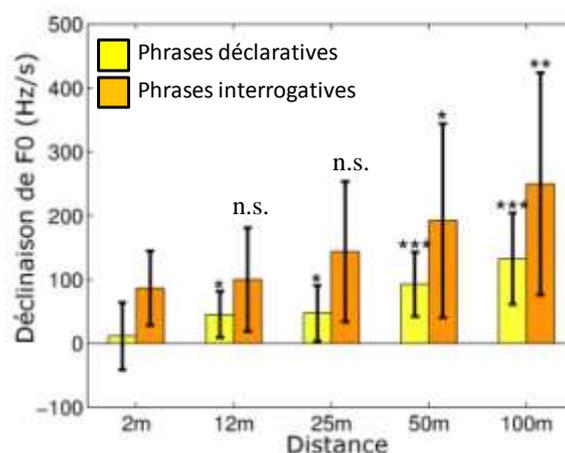


Figure 8-7: Évolution générale de F0 pour le corpus DB2. Les tests de significativité ont pour référence les valeurs pour la distance de 2 m.

### 8.2.2.3 Analyses de l'évolution de l'intensité

Dans l'objectif d'interpréter la dynamique de l'intensité observée nous avons mesuré l'intensité de chaque phonème séparément. La Figure 8-9 représente ainsi les variations de l'intensité des cinq phrases, par rapport à la voix modale. On constate sur cette figure que l'intensité augmente progressivement le long de la phrase pour la distance de 100 m. La raideur de l'évolution de l'intensité semble également être fonction de la distance de communication. En effet, la Figure 8-8 représente les valeurs du coefficient directeur d'une droite de régression calculée sur le contour de l'intensité de chaque phrase. On constate ainsi, que pour les phrases déclaratives, l'intensité reste constante tout au long de la phrase sauf pour la distance de 100 m. Toutefois, pour les phrases interrogatives, des variations significatives interviennent à partir d'une communication de 25 m. Ainsi pour des phrases criées à une distance de 100 m, l'intensité augmente au cours de la phrase. D'autre part, l'intensité des

voyelles varie plus que celle des consonnes. L'amplitude des variations augmente avec la distance de communication. On relève ainsi que pour une distance de 100 m, les voyelles augmentent en moyenne de 15,2 dB SPL et que les consonnes augmentent seulement de 11,8 dB SPL.

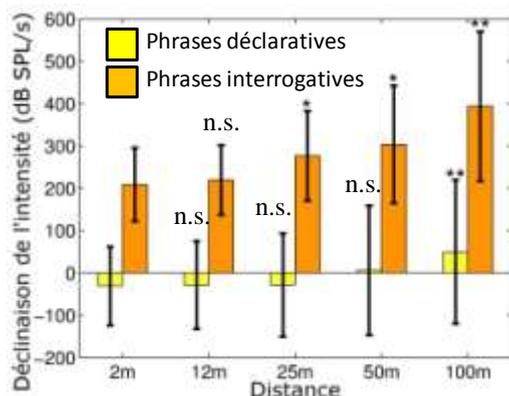


Figure 8-8: Évolution générale de l'intensité. Les tests de significativité ont pour référence les valeurs pour la distance de 2 m.

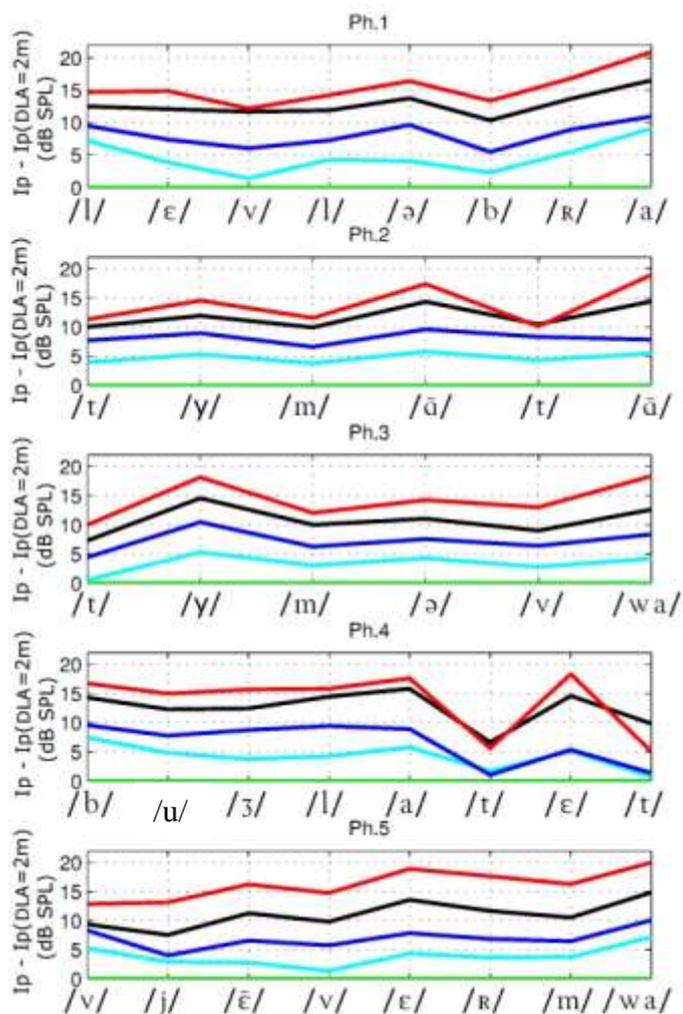


Figure 8-9: Évolution temporelle de l'intensité pour le corpus DB2

#### 8.2.2.4 Analyses des variations de durées

Le Tableau 8-1 donne les valeurs de la durée totale des énoncés en fonction de la distance de communication. Ces valeurs correspondent à la moyenne pour les 6 locuteurs. D'une part, les grandeurs des écarts-types indiquent une forte dépendance du locuteur sur la durée totale des énoncés. D'une manière générale, on observe une augmentation de la durée des énoncés avec la distance de communication.

Intéressons-nous aux variations de la durée des phonèmes pour expliquer cette augmentation de durée totale. La Figure 8-10 représente les variations de durée de chaque phrase, en fonction de la durée du phonème ( $T_p$ ) en voix modale (2 m), pour chaque phonème et distance. On observe une similarité entre les différentes courbes de distance pour chaque phrase. La durée des voyelles, même si elle varie beaucoup pour les phrases 2 et 3, ne permet pas d'en déduire une tendance générale. Toutefois, seule une règle peut être issue de cette représentation. La durée de la dernière voyelle de chaque phrase, pour une distance de 100 m augmente significativement (environ +40 %). Cette tendance est en accord avec la littérature bien que l'ordre de grandeur soit inférieur (+66% pour (Rostolland, 1982a)). On pourrait alors interpréter ce phénomène de la manière suivante : les locuteurs, pour des voix criées, ayant des difficultés à augmenter la  $F_0$  en fin de phrase, et notamment pour les phrases interrogatives, accentuent alors la dernière syllabe par le biais de l'intensité et de la durée.

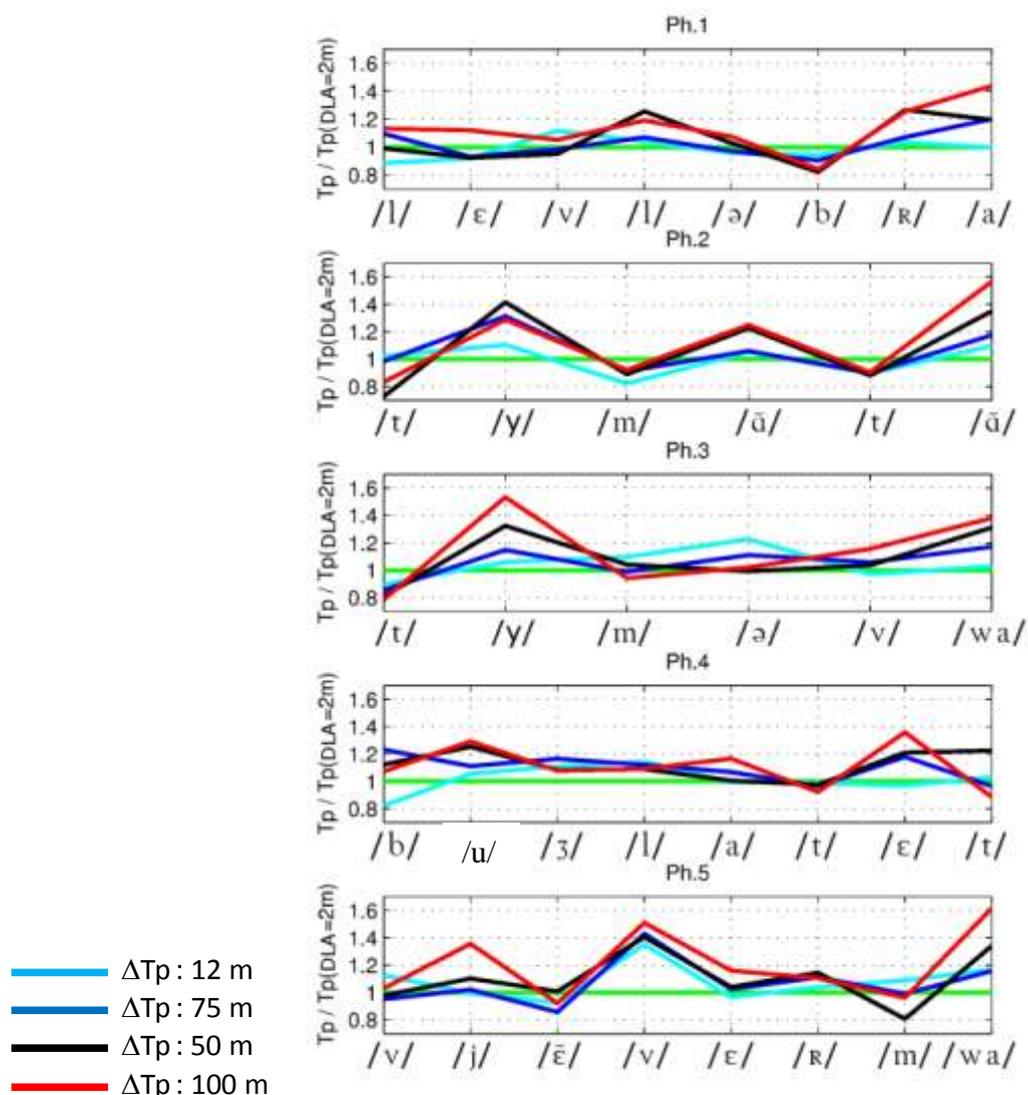


Figure 8-10: Variations moyennes de la durée des phonèmes pour le corpus DB2

Tableau 8-1: Durées totales des énoncés de DB2

	2m	12m	25m	50m	100m
$T_{PH1}$ (ms) (sd.)	582 (40)	570 (56)	605 (95)	613 (116)	678 (89)
$T_{PH2}$ (ms) (sd.)	463 (36)	467 (69)	489 (69)	513 (87)	550 (74)
$T_{PH3}$ (ms) (sd.)	466 (67)	486 (33)	499 (67)	517 (99)	539 (88)
$T_{PH4}$ (ms) (sd.)	663 (60)	674 (81)	722 (122)	751 (111)	733 (114)
$T_{PH5}$ (ms) (sd.)	558 (74)	599 (91)	586 (110)	615 (113)	682 (118)
Mean (ms) p-value	546 /	559 ns	580 *	602 **	636 ***

### 8.2.3 Évaluation du protocole de simulation (DB3) par comparaisons avec le corpus enregistré en champ libre de jour (DB1)

Deux corpus ont été enregistrés à l'extérieur où le locuteur et l'auditeur ont été séparés par une distance donnée (DB1 et DB2). Ces expériences à l'extérieur sont difficiles à réaliser car le bruit de fond est toujours présent. En outre, il est difficile de gérer les enregistrements avec un nombre important de sujets, tout en gardant les mêmes conditions d'enregistrements. Par conséquent, nous avons proposé un protocole d'enregistrement qui simule la communication à distance en utilisant une atténuation acoustique qui sépare le locuteur de l'auditeur (cf. §6.4). Ce protocole permet d'enregistrer plus de sujets dans d'excellentes conditions acoustiques de manière reproductible. Nous savons toutefois qu'il s'agit d'une situation artificielle qui peut ne pas correspondre à une communication à distance (par exemple la distance visuelle très faible entre locuteur et l'auditeur ne correspond pas à une situation réelle)

Nous avons alors réalisé une étude permettant de valider d'une part, le protocole de simulation et d'autre part, l'augmentation de la dynamique de F0 et de l'intensité (Fux et al., 2011a)<sup>1</sup>. Pour valider ce protocole nous nous basons sur des mesures statistiques qui sont les moyennes, les dynamiques ainsi que les raideurs des montées initiales et finales de F0.

Sur la Figure 8-11 sont représentées la dynamique de l'intensité et celle de la F0 pour le corpus de simulation de distance ainsi que pour le corpus réel enregistré en champ libre de jour (DB1). Sur ces

<sup>1</sup> Notons que dans l'article référence (Fux et al., 2011a) la dynamique de l'intensité et de F0 sont noté  $\Delta I$  et  $\Delta F0$ . Toutefois, dans ce manuscrit afin de ne pas confondre ces notations avec celles utilisées pour les variations de valeurs moyennes nous noterons ces dynamiques  $\delta I$  et  $\delta F0$ .

figures sont représentées les valeurs correspondant aux voix modales de chaque corpus ainsi qu'aux voix criées. Pour le corpus DB1 nous ne considérons cependant que la voix criée pour une distance de 100 m dans le but de comparer ces valeurs extrêmes avec le protocole de simulation. Nous séparons également pour le corpus DB3, les valeurs pour les sujets masculins et féminins.

On remarque alors que la dynamique de la F0 augmente avec la distance de communication. La dynamique de F0 des femmes évolue moins fortement que celle des hommes. On note également une corrélation positive entre la F0 moyenne et la dynamique pour les hommes ( $r=0,72$ , pente=0,25) mais cette corrélation est moins forte pour les femmes ( $r=0,52$ , pente=0,24). La dynamique de l'intensité augmente également significativement avec la distance de communication pour les hommes. Toutefois cette augmentation de dynamique est moins prononcée pour les femmes. De la même manière que précédemment une corrélation positive est observée entre les valeurs moyennes de l'intensité et sa dynamique pour les hommes ( $r=0,69$ , pente=0,43) et pour les femmes ( $r=0,51$ , pente=0,26).

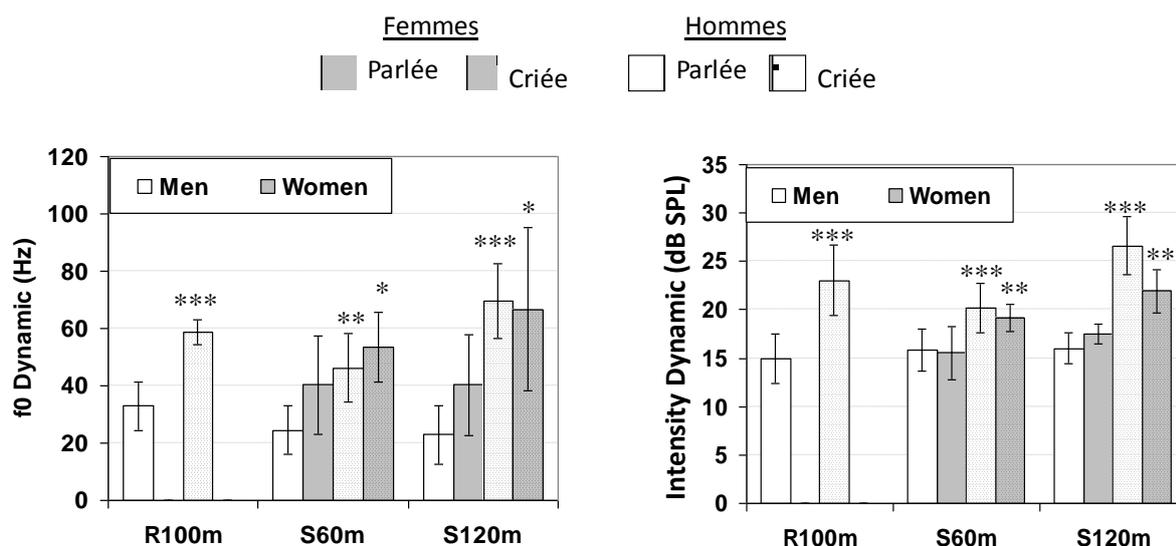


Figure 8-11: Dynamique de F0 (à gauche) et de l'intensité (à droite) pour le corpus de simulation de distance ainsi que pour les distances extrêmes du corpus enregistré en champ libre de jour DB1. R100m correspond au corpus réalisé dans une situation réelle à 100 m, S60m le protocole de simulation à 60 m et S120m le protocole de simulation de 120 m. Les tests de significativité ont pour référence les valeurs de la voix parlée associée.

Pour chaque phrase de chaque corpus, les pentes initiales et les pentes finales de F0 ont été déterminées. Ces pentes sont mesurées par une régression linéaire en utilisant une méthode semi-automatique. Nous veillons à ce que les mesures soient toujours réalisées sur la même région de chaque phrase. Les pentes initiales ont été mesurées sur les transitions entre la première consonne et la première voyelle de chaque phrase. Les pentes finales ont été mesurées à la fin de la voyelle finale. D'une façon générale les relevés ont été effectués entre la valeur initiale et le premier maxima du

contour de F0 pour les pentes initiales, et entre le dernier maxima et le dernier point de F0 pour les pentes finales. La Figure 8-12 donne les valeurs moyennes des pentes initiales et finales de F0 pour tous les corpus. Le Tableau 8-2 montre l'évolution de ces valeurs moyennes ainsi que la significativité des variations. On observe que la pente initiale de F0 augmente de façon significative avec la distance de communication (sauf pour les femmes-S60m). L'intonation modale implique généralement une diminution de F0 en fin de phrase déclarative. Nous observons que ce phénomène est considérablement amplifié pour les voix criées (sauf pour les femmes-S60m).

Il est donc clair que les trois protocoles présentent tous une augmentation de l'effort vocal ainsi que de la dynamique de la F0 et de l'intensité en fonction de la distance de communication. Des différences peuvent toutefois être observées entre ces trois corpus. La valeur moyenne et la dynamique de F0 pour la voix parlée est plus élevée pour la distance réelle que pour le protocole de simulation. Ceci est probablement lié à la première distance de communication utilisée pour le corpus en champ libre. En effet pour ce corpus la première distance de communication était de 5 m tandis que pour le protocole de simulation celle-ci était de 1 mètre. Cette différence quant à la distance de communication peut alors être la cause de ces valeurs plus intenses. Le sexe des locuteurs semble également affecter les résultats obtenus. On observe moins de variations sur l'intensité et la F0 des femmes que pour les hommes. Ceci peut être lié à la différence de mécanismes laryngés utilisés entre les hommes et les femmes (Holmberg et al., 1988). Les variations de dynamique de l'intensité et de la F0 sont également moins prononcées chez les femmes que chez les hommes. Toutefois comme la dynamique et les valeurs moyennes semblent être corrélées, les variations moins fortes de dynamique semblent être cohérentes avec les variations moins fortes observées pour les femmes (cf. Tableau 8-2). D'autre part, malgré une précision limitée de mesure, les pentes initiales et finales de F0 montrent des variations significatives avec la distance de communication. Pour les locuteurs féminins, ces variations sont également moins importantes, ce qui pourrait alors traduire des efforts vocaux moins importants chez les femmes (ce qui est en accord avec les variations et les dynamiques plus faibles observées).

Au regard de ces valeurs moyennes ainsi que de dynamique nous remarquons que le protocole de simulation de communication de distance que nous avons élaboré semble être cohérent avec les enregistrements réalisés en champ libre. Nous considérons donc que notre protocole est valable pour simuler les différentes distances de communication. Ceci nous a permis par la suite d'enregistrer facilement le corpus DB4 qui aurait été difficilement réalisable en extérieur.

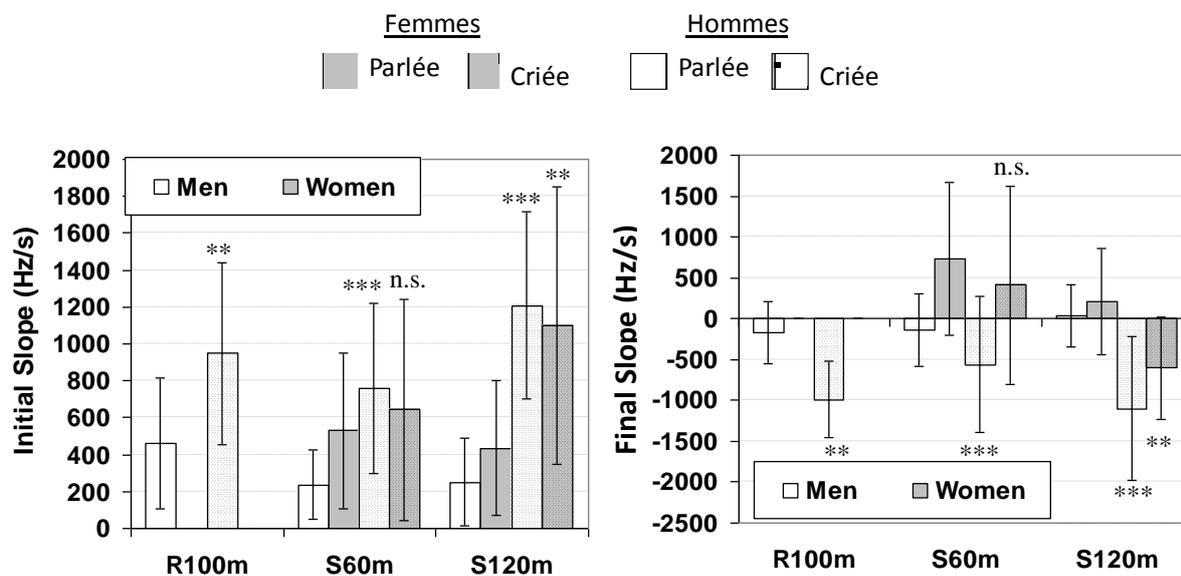


Figure 8-12: Valeur du coefficient directeur des droite de régression sur les pente initiale (à gauche) et finale (à droite) de la F0 pour le corpus de simulation de distance ainsi que pour les distances extrêmes du corpus enregistré en champ libre de jour DB1. R100m correspond au corpus réalisé dans une situation réelle à 100 m, S60m le protocole de simulation à 60 m et S120m le protocole de simulation de 120 m. Les tests de significativité ont pour référence les valeurs de la voix parlée associée.

Tableau 8-2: Évolutions de a) I et F0, b) de  $\delta I$  et  $\delta f_0$ , c) les pentes initiale et finale de  $f_0$ , entre les voix parlées et criées pour les trois expériences, pour les hommes et les femmes. df: degré de liberté, t: t-valeur, Diff: différence entre la voix parlée et la voix criée.

a)	I (dB SPL)			f0 (Hz)		
	df	t	Diff.	df	t	Diff.
Men S60m	88	-16,64	+14,2 ***	88	-16,86	+79 ***
Men R100m	28	-18,84	+18,3 ***	28	-21,63	+122 ***
Men S120m	88	-19,75	+22,7 ***	88	-20,74	+149 ***
Women S60m	28	-4,26	+9,0 ***	28	-5,27	+49 ***
Women S120m	28	-7,17	+14,6 ***	28	-6,81	+94 ***

b)	$\delta I$ (dB SPL)			$\delta f_0$ (Hz)		
	df	t	Diff.	df	t	Diff.
Men S60m	88	-4,07	+4,4 ***	88	-6,81	+21,7 ***
Men R100m	28	-4,32	+8,1 ***	28	-2,97	+25,7 **
Men S120m	88	-10,75	+10,6 ***	88	-12,49	+46,1 ***
Women S60m	28	-3,39	+3,6 **	28	-1,93	+13,2 *
Women S120m	28	-2,74	+4,5 **	28	-2,40	+26,5 *

c)	Initial f0 slope (Hz/s)			Final f0 slope (Hz/s)		
	df	t	Diff.	df	t	Diff.
Men S60m	88	-7,01	+520 ***	88	3,06	-429 **
Men R100m	28	-3,07	+488 **	28	5,31	-818 ***
Men S120m	88	-11,45	+957 ***	88	7,96	-1139 ***
Women S60m	28	-0,59	+112 n.s.	28	0,80	-318 n.s.
Women S120m	28	-3,07	+662 **	28	3,48	-812 **

\*\*\*  $p < 0.001$ , \*\*  $p < 0.01$ , \*  $p < 0.1$ , n.s.  $p > 0.1$

---

## 8.3 Dynamique des formants

---

Il a été reporté dans la littérature une augmentation de la fréquence du premier formant pour les voix criées. Cependant, les valeurs reportées sont variées et les explications divergent (distance F0-F1, *formant tuning*, ...). C'est pourquoi, nous avons voulu nous faire notre propre opinion sur le sujet en relevant les valeurs de déplacement du premier formant. Pour évaluer le déplacement des formants nous nous concentrons sur le corpus DB4 car la recherche d'intelligibilité dans ce corpus est, *a priori*, plus prononcée. En effet, dans ce corpus, l'utilisation de logatomes a forcé les sujets à être le plus intelligible possible. Ainsi, dans ce corpus, le déplacement des formants devrait être plus révélateur que pour les autres corpus où le besoin d'intelligibilité est moins fort, dû à l'utilisation de phrases. Nous utilisons ainsi les logatomes CV du corpus DB4 pour mesurer le déplacement des formants.

Etant donné que de nombreux auteurs évoquent la variation concordante de F1 en fonction de F0 (cf. chapitre précédent), nous jugeons utile de présenter une variation de F1 en fonction de F0. Les Figure 8-13, Figure 8-14 et Figure 8-15 représentent les valeurs mesurées du premier formant en fonction de la F0. Ces mesures ont été effectuées sur la totalité de la voyelle et non pas à un instant précis de celle-ci. Chacune de ces figures représentent la F1 en fonction de F0 de chaque locuteur du corpus DB4. Les voyelles parlées sont représentées par des ronds et les voyelles criées par des croix. Sur ces figures on constate clairement que le formant F1, en voix parlée se concentre vers une zone plus ou moins diffuse autour des valeurs théoriques de F1 pour ces voyelles (représentées par des ellipses sur les figures). On constate naturellement des valeurs qui s'écartent sensiblement de cette zone mais qui semblent correspondre à des erreurs d'estimation et notamment en début ou fin de voyelle.

Pour les voix criées, on constate que les points s'étalent le long de l'axe F0. Ceci correspond alors parfaitement aux mesures de dynamique que nous avons effectuées. En effet, la dynamique étant plus grande, la plage de variation de F0 en voix criée est plus grande ; d'où l'étirement du nuage de points pour F1 des voyelles criées. D'autre part, on constate que F1, pour les voix criées se concentre autour d'une harmonique de F0. Pour les voyelles /i/ et /u/, F1 se concentre autour de la première harmonique alors que pour les voyelles /a/, F1 se concentre autour de la troisième harmonique (sauf pour le locuteur 3 où cette tendance n'est pas clairement observée). Ce phénomène pourrait correspondre à la théorie du « *formant tuning* » (Garnier et al., 2010; Joliveau et al., 2004a, 2004b; Sundberg, 1977).

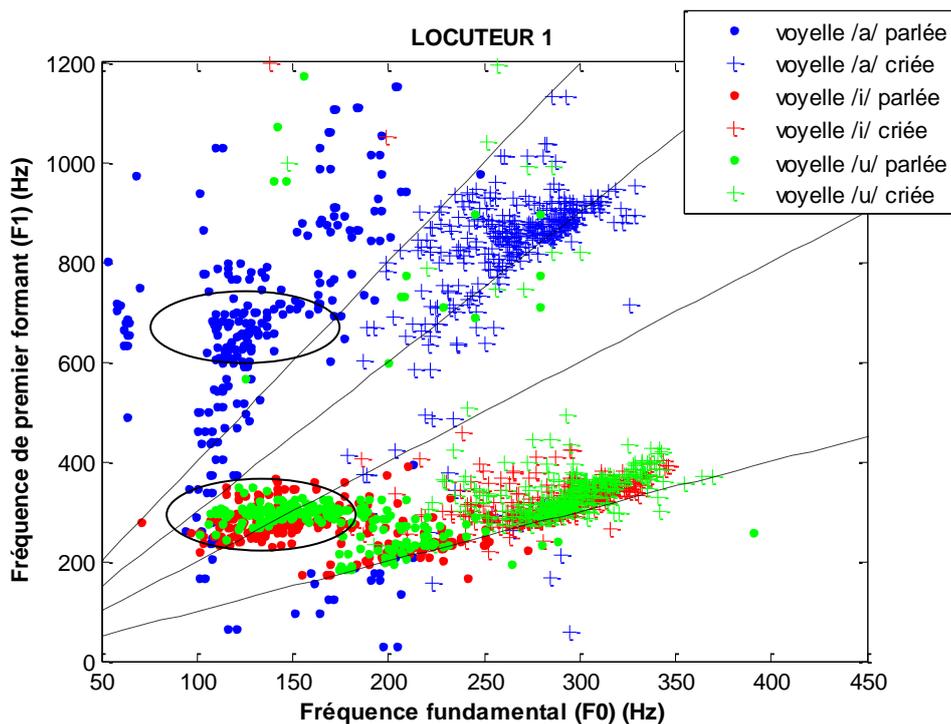


Figure 8-13: Relevés des valeurs du premier formant sur l'ensemble de la voyelle parlée et criée, des logatomes CV, en fonction de la F0 pour le locuteur 1. Les droites en pointillé correspondent aux harmoniques de F0.

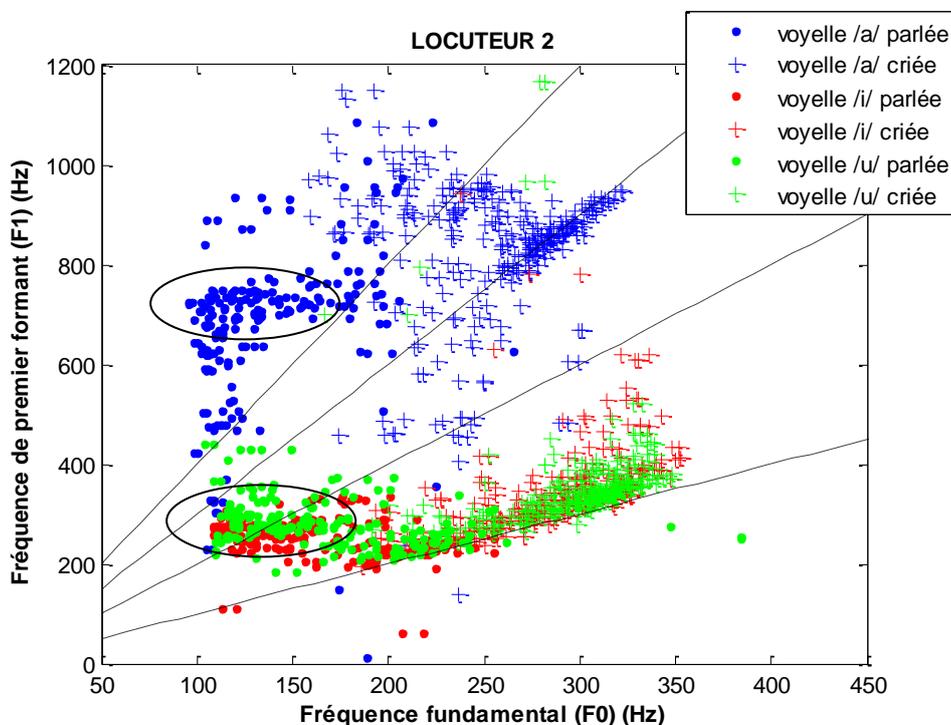


Figure 8-14: Relevés des valeurs du premier formant sur l'ensemble de la voyelle parlée et criée, des logatomes CV, en fonction de la F0 pour le locuteur 2. Les droites en pointillé correspondent aux harmoniques de F0.

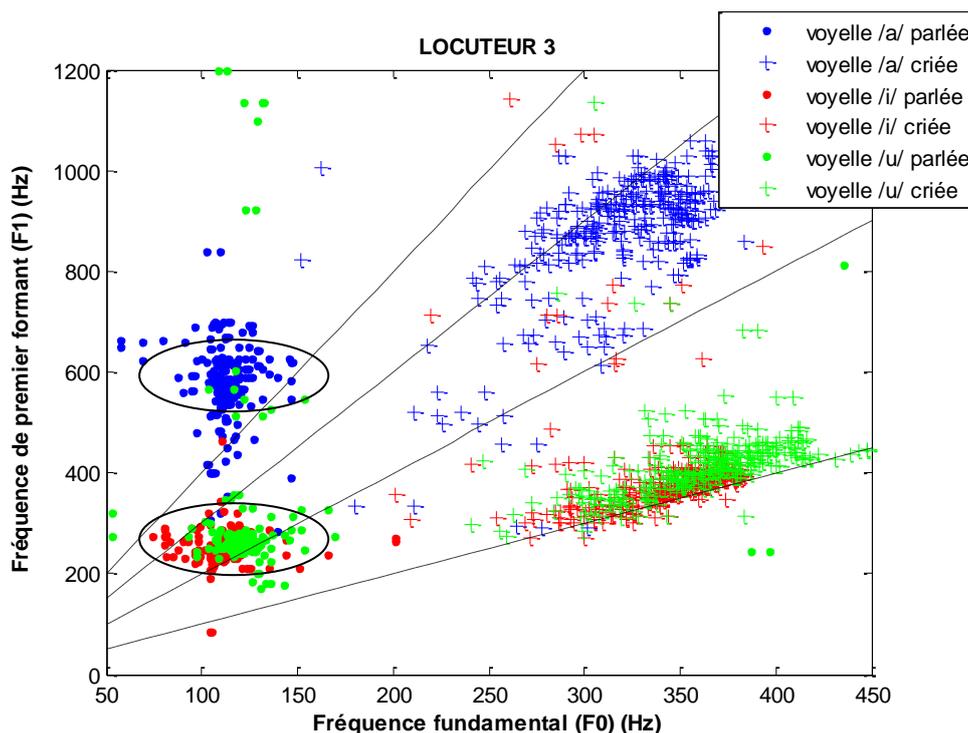


Figure 8-15: Relevés des valeurs du premier formant sur l'ensemble de la voyelle parlée et criée, des logotomes CV, en fonction de la F0 pour le locuteur 3. Les droites en pointillé correspondent aux harmoniques de F0.

Cependant il faut souligner que la mesure des formants par la méthode de la prédiction linéaire souffre d'erreurs d'estimation importantes pour des F0 élevées. En effet le spectre d'une voix ayant une F0 élevée (i.e. voix criée, chantée, etc...) possède un spectre dont les raies harmoniques sont plus espacées. La mesure des formants devient alors difficile du fait du manque d'énergie entre les harmoniques (Atal et Schroeder, 1974; Monsen et Engebretson, 1983; Traunmüller et Eriksson, 1997; Vallabha et Tuller, 2002).

Nous montrons ici une simple expérience pour mettre en évidence ce phénomène. Nous avons créé un signal d'excitation à partir d'un peigne de Dirac où la F0 varie progressivement de 75 Hz à 500 Hz. Le filtre utilisé est un filtre issu d'une analyse d'un /œ/ qui possède une F1=420 Hz et F2=1470 Hz (cf. Figure 8-16). Nous effectuons ensuite une mesure des formants à l'aide de méthode « standard » (par prédiction linéaire). On observe que, malgré le fait que le filtre soit constant tout au long de la voyelle synthétisée, les valeurs mesurées de F1 et F2 varient. On observe en effet sur Figure 8-17, que ces derniers sont influencés par les harmoniques de F0. Les relevés montrent que le suivi de F1 et F2 « saute » d'une harmonique à l'autre. Les erreurs maximales mesurées sur cet exemple sont de 70 Hz pour F1 et de 130 Hz pour F2. Par ailleurs on observe, en accord avec la théorie de l'efficacité de la voix, que lorsque F1 se situe sur l'harmonique l'intensité de la voix est plus forte.

Bien que ces erreurs n'expliquent pas entièrement les fortes variations des formants observées, elles compliquent considérablement la caractérisation du déplacement de ceux-ci. Sans vouloir remettre

complètement en cause la théorie du *formant tuning*, nous mentionnons simplement la difficulté de mesurer la position des formants et les erreurs que ces mesures peuvent engendrer. Soulignons également que le phénomène du *formant tuning* est principalement observé en voix chantée. Rares sont les études qui observent des phénomènes aussi flagrants en voix criée. Notons également que dans nos analyses les formants ne se placent pas exactement sur les harmoniques de F0.

N'ayant pas réalisé nos corpus dans l'objectif précis d'effectuer la mesure de formants, nous ne disposons que du signal acoustique à partir duquel il est difficile d'estimer la position des formants. Ainsi, nous ne réalisons pas ici une étude de caractérisation des formants en voix criée. Signalons qu'il existe une méthode permettant de mesurer la position des formants par une approche qui n'utilise pas le signal de la parole. En effet, la difficulté d'estimation des formants provient d'un manque de résolution spectrale du fait du fort écartement des harmoniques de F0 pour les voix ayant une F0 élevée. Ainsi, une méthode dite « *broadband excitation* » a été développée par Epps (Epps et al., 1997). Dans ce type de mesure un bruit blanc est injecté dans la bouche du locuteur pendant la prononciation d'une voyelle. Un microphone enregistre ensuite le signal, qui est modulé par le conduit vocal et qui est additionné à la parole, permettant de déterminer la réponse en fréquence du conduit vocal. Des post-traitements permettent alors de localiser avec précision la position des résonances du conduit vocal. Cependant, les résultats obtenus par cette méthode semblent similaires à ceux obtenus par les méthodes plus traditionnelles. En effet, cette méthode se base sur une hypothèse forte, à savoir que le conduit vocal est acoustiquement réversible. En d'autres termes que l'excitation du conduit vocal au niveau de la bouche engendrerait la même réponse fréquentielle que l'excitation faite par les cordes vocales qui se situe à l'autre bout du conduit vocal. La vérification de cette hypothèse de départ n'est pas triviale.

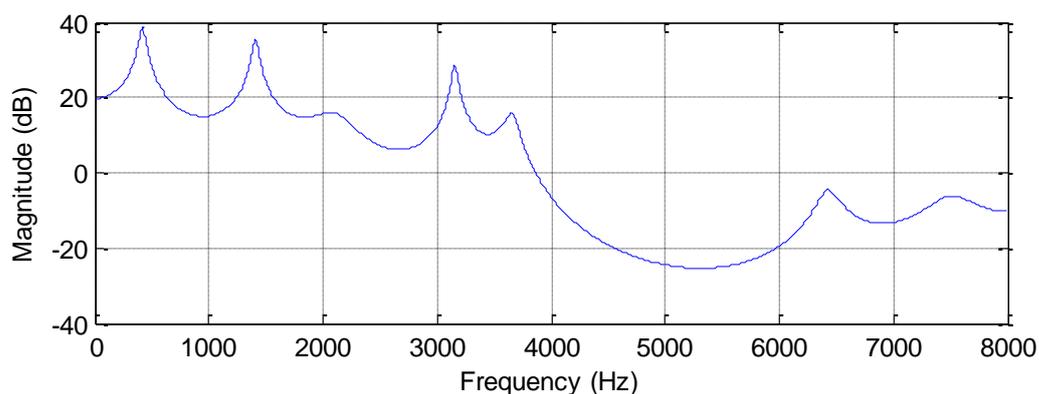


Figure 8-16 : Fonction de transfert du filtre utilisé pour la mise en évidence des erreurs de mesure des formants.

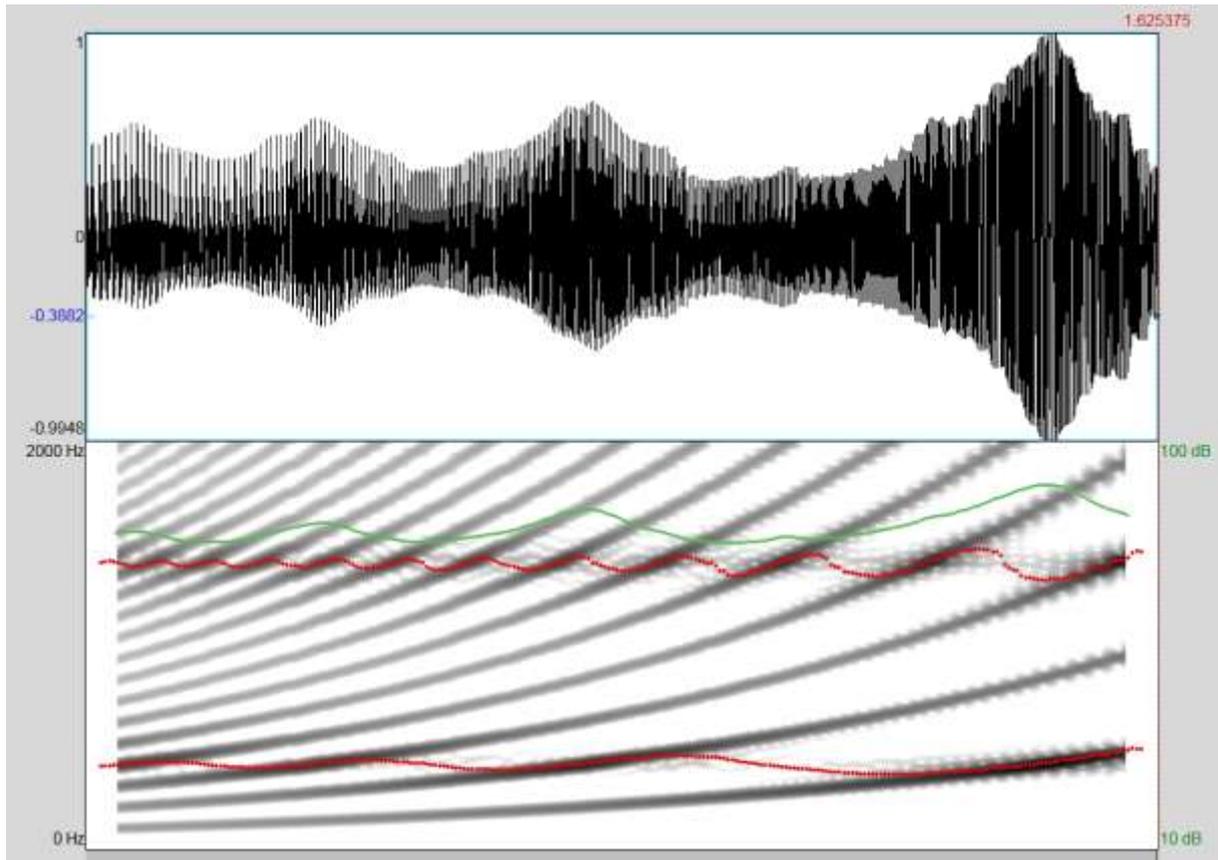


Figure 8-17: exemple d'erreur d'estimation de la position des formants sur une voyelle synthétisé (i.e. filtre du conduit vocal constant)

## 8.4 Conclusion du chapitre 8

L'objectif principal de ce chapitre est d'extraire les variations des paramètres pertinents directement liés à l'effort vocal sur le signal de parole. Les résultats confirment la relation entre les variations de F0 et d'intensité et de DLA, mais ils montrent aussi une forte corrélation entre la dynamique de F0, et de la dynamique d'intensité et de DLA. Ceci suggère, que les dynamiques de I et F0 sont de bons indicateurs de l'augmentation de l'effort vocal du locuteur. Les mesures des dynamiques de F0 et d'intensité montrent clairement que pour une voix criée, les contours d'intensité et de F0 ont des valeurs moyennes supérieures ainsi que des dynamiques plus fortes (cf. Figure 8-18 et Figure 8-19). Par conséquent, dans le but de transformer une voix naturelle permettant d'indiquer la distance du locuteur, les évolutions dynamiques des paramètres prosodiques doivent être prises en compte. De plus, la littérature mentionne également l'importance de ces paramètres. Par ailleurs, la cohérence entre les mesures obtenues dans des conditions simulées de DLA et dans les conditions extérieures,

valide le protocole d'enregistrement en laboratoire que nous avons proposé. Celui-ci permet de réaliser de bons enregistrements acoustiques avec de nombreux participants.

Les analyses effectuées dans ce chapitre sont très globales. Celles-ci permettent de mettre en évidence des variations significatives de la dynamique de F0 et de l'intensité. Mais pour réaliser une transformation de la voix criée, nous avons besoin d'avantage de précisions sur le contour prosodique. C'est pourquoi suite à cette étude, nous avons réalisé des enregistrements de logatomes (DB4) dans l'objectif de mener des analyses précises des variations prosodiques qui permettront d'expliquer les phénomènes observés dans ce chapitre. Le chapitre suivant permet également de « modéliser » un contour prosodique qui conduit à définir les règles de transformation.

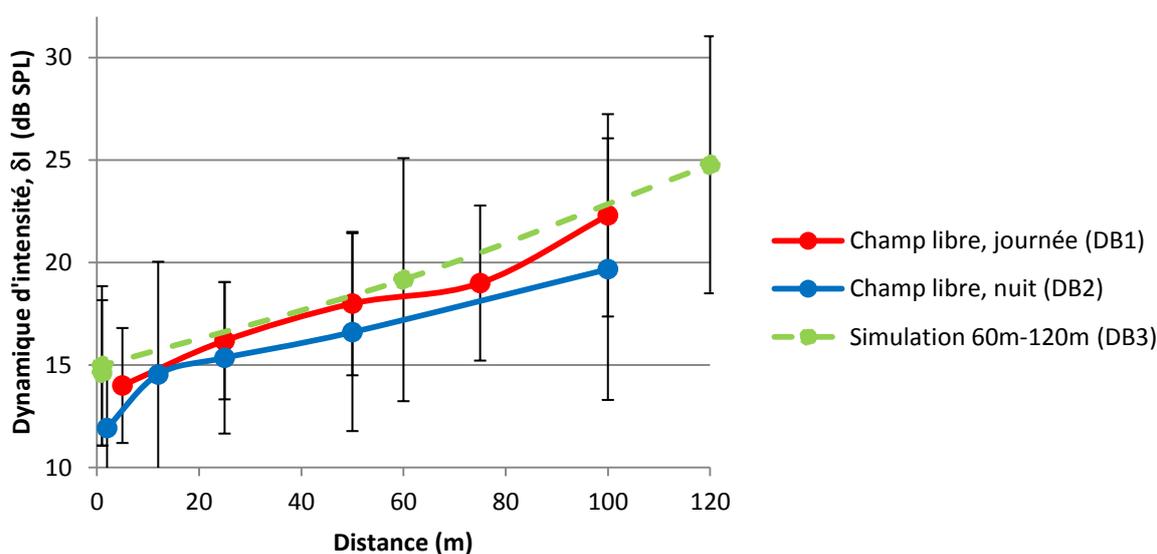


Figure 8-18: Dynamique de l'intensité des voix en fonction de la distance de communication et du corps

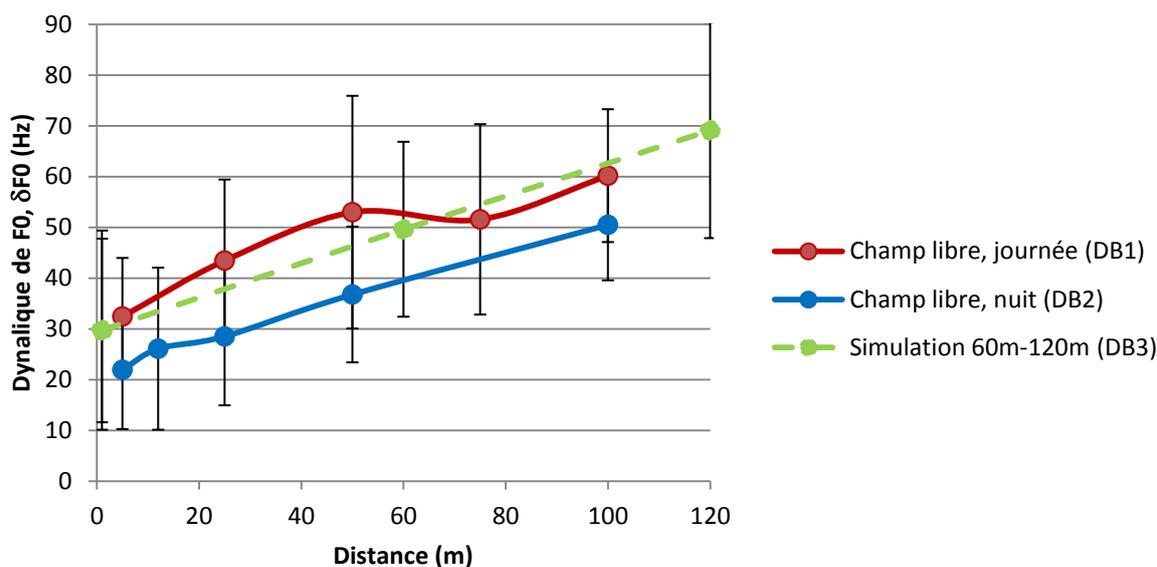


Figure 8-19 : Dynamique de F0 des voix en fonction de la distance de communication et du corps



# La prosodie et la micro-prosodie de l'effort vocal

---

*« La phrase la plus excitante à entendre en science, celle qui annonce de nouvelles découvertes, n'est pas «Eureka» (j'ai trouvé!), mais plutôt «Tiens, c'est marrant...»  
— Isaac Asimov*

Dans le chapitre précédent, nous avons pu constater des évolutions des paramètres prosodiques qui semblent dépendre de la distance de communication et ainsi de l'effort vocal. En effet, on a pu observer des dynamiques de F0 et d'intensité plus fortes pour des voix criées que pour des voix parlées. De plus, la littérature s'accorde à dire que la prosodie est l'élément le plus pertinent pour la perception de l'effort vocal (Brungart et al., 2002; Fux et al., 2010; Tassa et Liénard, 2000).

C'est tout naturellement que dans ce chapitre nous nous intéressons à l'étude de la prosodie des voix criées dans le but d'une communication à distance. Rappelons que la prosodie correspond aux variations temporelles de la F0, de l'intensité et de durée. La prosodie décrit les évolutions de ces paramètres au cours d'un énoncé. La prosodie constitue ainsi une macro-analyse des évolutions de ces paramètres. Toutefois, comme nous le verrons, il existe également des variations de paramètres au sein d'un même phonème. Ces variations au sein d'un même phonème (micro-analyse) constituent la micro-prosodie. Ce chapitre se consacre également à l'étude de la micro-prosodie de l'effort vocal. Le paramètre le plus pertinent pour la perception de l'effort vocal et ainsi de la distance semble être la F0. C'est pourquoi dans la partie destinée à l'étude de la micro-prosodie, nous nous penchons uniquement sur l'étude de la micro-mélodie (i.e les variations locales de F0). De plus, seuls les aspects liés à l'augmentation de l'effort vocal seront traités dans ce chapitre. Nous n'évoquerons pas la prosodie ou la micro-prosodie des voix chuchotées, *a priori*, mieux comprises à ce jour que celles des voix criées.

Afin de réaliser ces études prosodiques et micro-prosodique nous avons réalisé un corpus dédié : DB4. Le corpus DB4 est constitué de structures phonétiques de base CV, CVC, VCV et CVCV construites à partir des 17 consonnes du français ainsi que des 3 voyelles cardinales /a/, /i/, /u/. En effet, l'analyse des variations prosodiques est trop complexe à réaliser sur des énoncés entiers. De plus, l'interprétation de ces variations sur des phrases trop longues et complexes sont également difficilement interprétables. C'est pourquoi nous avons pris la décision de réaliser des études sur 4

types de logatomes. Pour réaliser cette étude nous effectuons ainsi des relevés de durée, d'intensité et de F0 sur chaque phonème des logatomes de DB4. Nous comparons ainsi l'influence des catégories des phonèmes ainsi que la structure phonétique sur l'évolution de ces paramètres. Nous nous intéressons également tout particulièrement aux formes du contour de F0 des voix criées. Nous nous concentrons principalement sur les phénomènes invariants et pertinents qui apparaissent en voix criée. L'aspect analytique ne s'oriente pas vers une analyse précise visant à évaluer l'influence de tel ou tel type de phonème sur les variations, mais plutôt d'essayer de déduire une stratégie globale de ces analyses.

Nous abordons tout d'abord dans ce chapitre les variations de la durée de ces structures phonétiques en voix criée. Nous nous intéressons particulièrement à la durée totale des énoncés, aux variations entre les consonnes et voyelles, ainsi qu'à certaines constances entre la durée des voyelles au sein d'un même mot. La section suivante est consacrée à l'étude des variations d'intensité sur ces énoncés et notamment du rapport entre l'augmentation de l'intensité des consonnes et des voyelles. Enfin nous traitons l'aspect mélodique de l'effort vocal. De la même manière nous cherchons à mettre en évidence les différences qui existent entre l'évolution de la F0 sur les consonnes voisées et les voyelles. Enfin, une dernière section est consacrée à la micro-prosodie et particulièrement à la micro-mélodie.

Nous cherchons dans cette section à mettre en avant les différences de forme du contour mélodique entre la voix modale et la voix criée. Enfin, nous finissons ce chapitre par des hypothétiques interprétations des phonèmes observés. Soulignons que nous n'avons pas l'intention de mener ici une étude complète sur la prosodie des voix criées. Notre étude a un but plus pragmatique : extraire les règles permettant une meilleure transformation. L'objectif est alors de mieux caractériser une voix criée liée à l'augmentation de l'effort vocal ou à la communication à distance.

## 9.1 Analyses de la durée

---

Cette section regroupe les analyses de durées réalisées sur le corpus DB4 dédié à cette étude prosodique. L'analyse de la durée des phonèmes permet d'observer des stratégies d'accentuation sur un énoncé. D'autre part, la littérature sur les variations de durée en voix criée mentionne des variations systématiques sur la durée des voyelles mais pas sur les consonnes. Nous souhaitons ainsi à travers cette section éclaircir ce point afin de définir les règles de transformation de voix parlée en voix criée. Nous abordons tout d'abord la question de la durée totale des énoncés. Une deuxième étape est l'analyse séparée de la variation de durée des consonnes et des voyelles.

### 9.1.1 Analyse de la variation de durée totale entre les structures étudiées

Avant d'observer les variations de durée sur les phonèmes, nous nous intéressons tout d'abord aux variations de la durée totale des logatomes de DB4.

La Figure 9-1 représente les variations de la durée des énoncés pour l'ensemble du corpus DB4 par rapport à la voix parlée. Sur cette figure chaque graphique correspond à une structure phonétique. Sur ces graphiques, la variation de la durée et l'écart-type pour chaque locuteur ainsi que la moyenne des 3 locuteurs sont représentés.

La première constatation faite sur la durée totale des énoncés, est une augmentation significative pour la majorité des énoncés, bien que le locuteur 1 et 3 ne rallongent pas de manière significative la durée des énoncés CVC. D'autre part on observe également, d'un locuteur à l'autre, des stratégies différentes qui se traduisent par des augmentations plus ou moins prononcées. En effet, hormis pour les VCV où les augmentations sont similaires pour les 3 locuteurs, les variations de la durée sont différentes d'un locuteur à l'autre et d'un énoncé à l'autre. D'une façon plus générale, la durée des énoncés augmente en moyenne de 24 %, et de 15 %, 19 % et 12 % pour les énoncés CV, CVC, VCV, et CVCV. Ceci correspond à une augmentation générale de l'ordre de 17 %.

Ainsi, une augmentation systématique de la durée des énoncés est observée entre une voix parlée et une voix criée. Plusieurs auteurs (Rostolland, 1982a; Traunmüller et Eriksson, 2000) ont observé des variations de durée des voyelles et des consonnes qui sont différentes (cf § 5.2.5.1). La section suivante est ainsi consacrée à l'étude séparée des variations de la durée des consonnes et des voyelles des voix criées par rapport aux voix parlées.

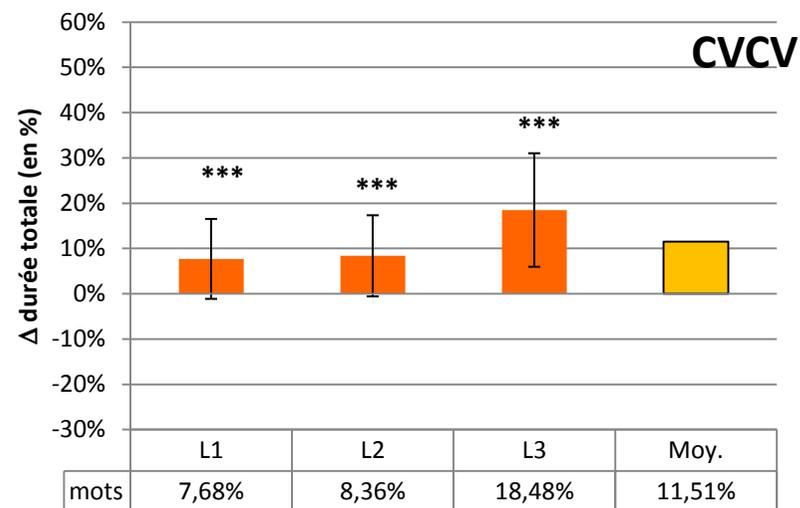
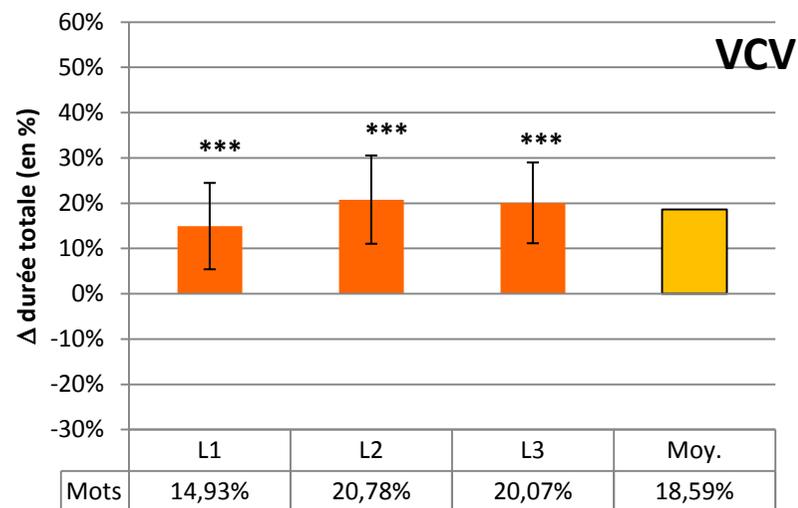
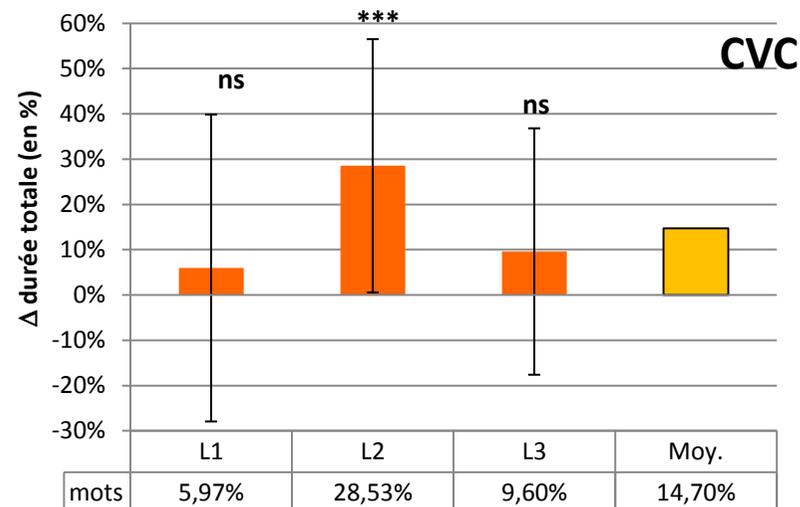
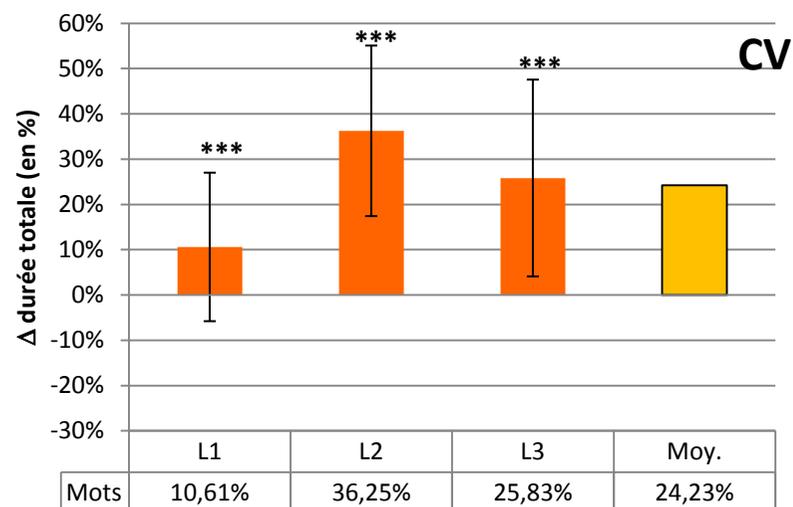


Figure 9-1: Variations de la durée des logatomes de DB4 en fonction des locuteurs. Les graphiques représentent les valeurs pour chacun des locuteurs ainsi que les valeurs moyennes des 3 locuteurs confondus

### 9.1.2 Analyse de la variation de durée des consonnes et des voyelles

Dans le but de mieux comprendre les variations de la durée globale, une analyse plus précise considérant les variations des consonnes et des voyelles séparément a été réalisée. Nous analysons dans cette section les variations de durée pour chaque structure phonétique du corpus DB4 et pour chaque locuteur.

#### 9.1.2.1 Variations relatives de la durée des consonnes et des voyelles pour les logatomes CV

La Figure 9-2 représente l'évolution des consonnes et des voyelles des mots CV pour les 3 locuteurs. Concernant les CV, les variations des consonnes sont fortement dépendantes du locuteur : L1 diminue de manière significative la durée des consonnes, L2 augmente les durées de celles-ci tandis que L3 n'effectue pas de changements significatifs de la durée des consonnes. Cependant, concernant les voyelles, les 3 locuteurs augmentent la durée de celles-ci de manière significative et dans les mêmes proportions. En effet, les 3 locuteurs augmentent la durée des voyelles dans les énoncés CV de l'ordre de 44%. Ainsi, pour les CV, les variations de la durée globale traduisent sans doute des stratégies différentes lors de la production des consonnes.

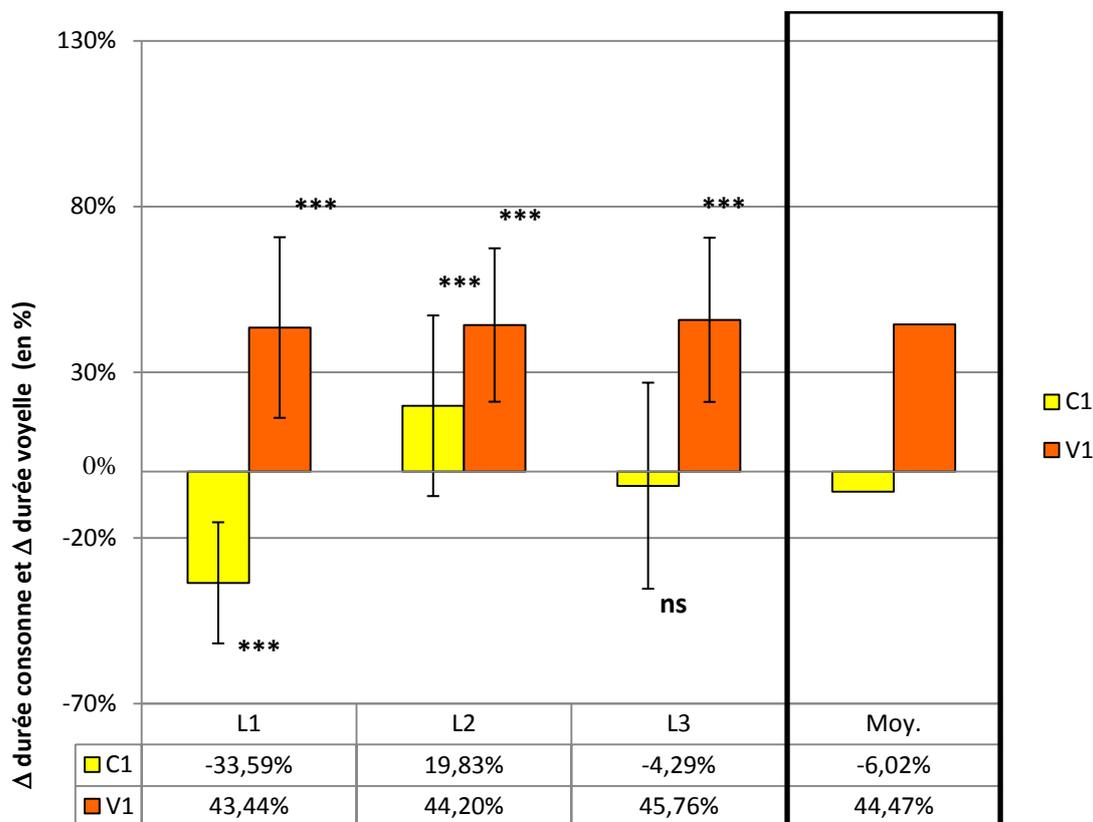


Figure 9-2 : Variations de durée entre la voix parlée et la voix criée des consonnes et voyelles pour les logatomes CV

### 9.1.2.2 Variations relatives de la durée des consonnes et des voyelles pour les logatomes CVC

La Figure 9-3 représente l'évolution des consonnes et des voyelles des mots CVC pour les 3 locuteurs. La durée des consonnes des CVC ne varie que faiblement, hormis pour la dernière consonne du locuteur L3 qui elle diminue de 25 %. On observe toutefois, une tendance générale à la diminution de la durée des consonnes et notamment les dernières (-12% en moyenne). La durée des voyelles quant à elle, est fortement augmentée. Les locuteurs L2 et L3 rallongent les voyelles de prêt de 80 % et de 50 % pour L1. En moyenne, la durée des voyelles des logatomes CVC est augmentée de 68 %.

Les augmentations de la durée des voyelles observées dans ce cas sont beaucoup plus importantes par rapport à celles observées dans le cas des logatomes CV. Par exemple, l'augmentation de durée observée pour L2 pour des logatomes CVC est près de 2 fois plus élevée que celle observée sur les logatomes CV.

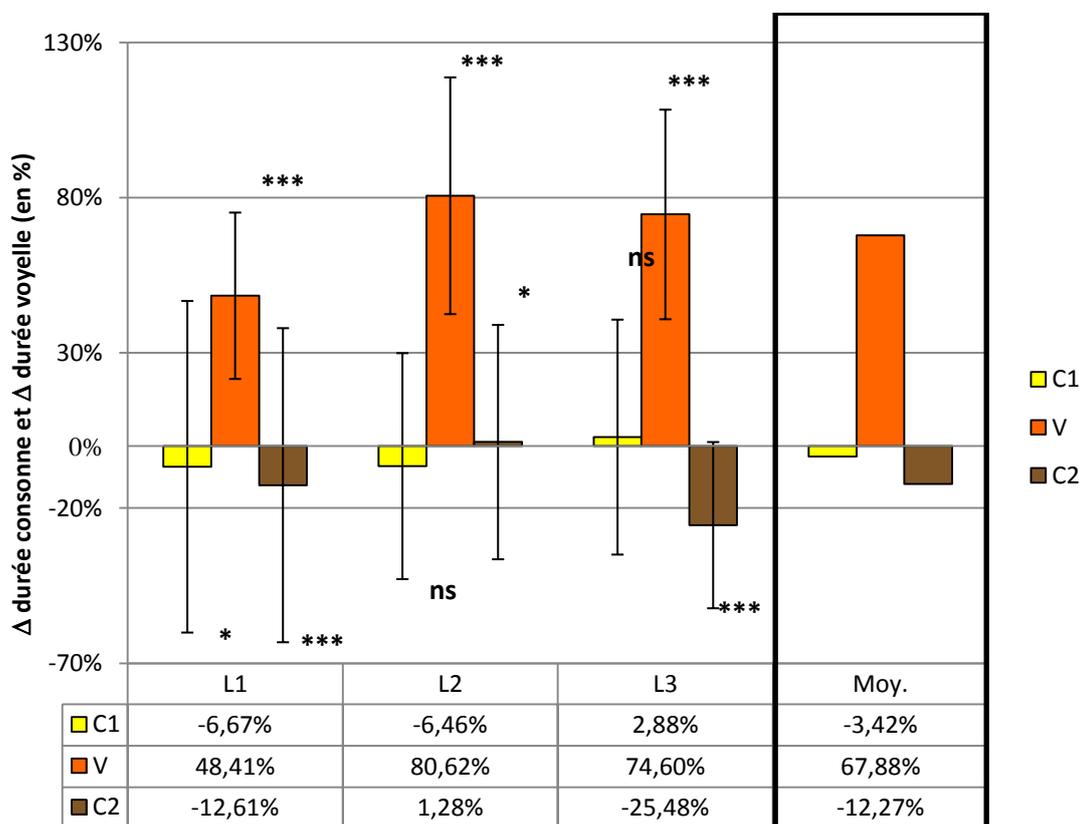


Figure 9-3 : Variations de durée entre la voix parlée et la voix criée des consonnes et voyelles pour les logatomes CVC

### 9.1.2.3 Variations relatives de la durée des consonnes et des voyelles pour les logatomes VCV

Considérons à présent le cas des énoncés VCV. La Figure 9-4 représente l'évolution des consonnes et des voyelles des mots VCV pour les 3 locuteurs. On constate une diminution significative de la consonne intervocalique pour L1 et L2. Mais cette diminution n'est pas observée pour L3 qui ne

modifie pas significativement la durée des consonnes intervocaliques. Comme pour les énoncés précédents, les voyelles augmentent également mais dans des proportions moins importantes que celle déjà observées (+30% en moyenne dans le cas présent). De plus, on constate, que les secondes voyelles (V2) augmentent plus (37% en moyenne) que les premières (V1) (25% en moyenne) et ceci quel que soit le locuteur.

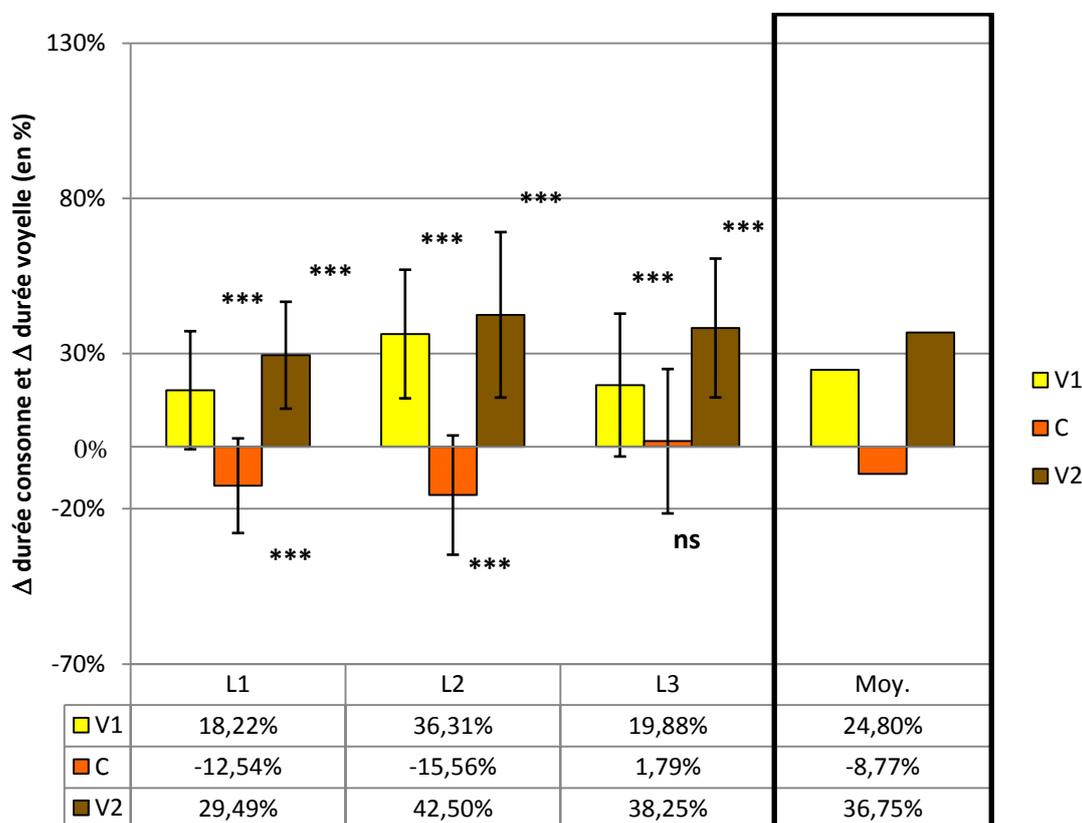


Figure 9-4 : Variations de durée entre la voix parlée et la voix criée des consonnes et voyelles pour les logatomes VCV

#### 9.1.2.4 Variations relatives de la durée des consonnes et des voyelles pour les logatomes CVCV

Enfin, les énoncés CVCV montrent des modifications plus complexes (cf. Figure 9-5). Pour 2 des 3 locuteurs, bien que les variations ne soient pas significatives, la durée des consonnes initiales (C1) augmentent tandis qu'elle diminue significativement pour L3. Toutefois les écarts-types mesurés sont très grands. Ces écarts-types semblent directement être liés à la difficulté de l'annotation des consonnes initiales non voisées. Nous avons également observé ce phénomène pour les logatomes CVC. Les variations des consonnes intervocaliques (C2) sont plus homogènes que celles des premières consonnes (C1) pour les 3 locuteurs, et ont une tendance à diminuer significativement (-20% en moyenne) ; notamment pour L1 et L2. On retrouve également des augmentations de la durée des voyelles dans des proportions identiques que celles observées pour les énoncés VCV (+30% en

moyenne). Toutefois, contrairement aux VCV et hormis pour L3, la tendance à allonger plus fortement V2 que V1 n'est ici pas observée.

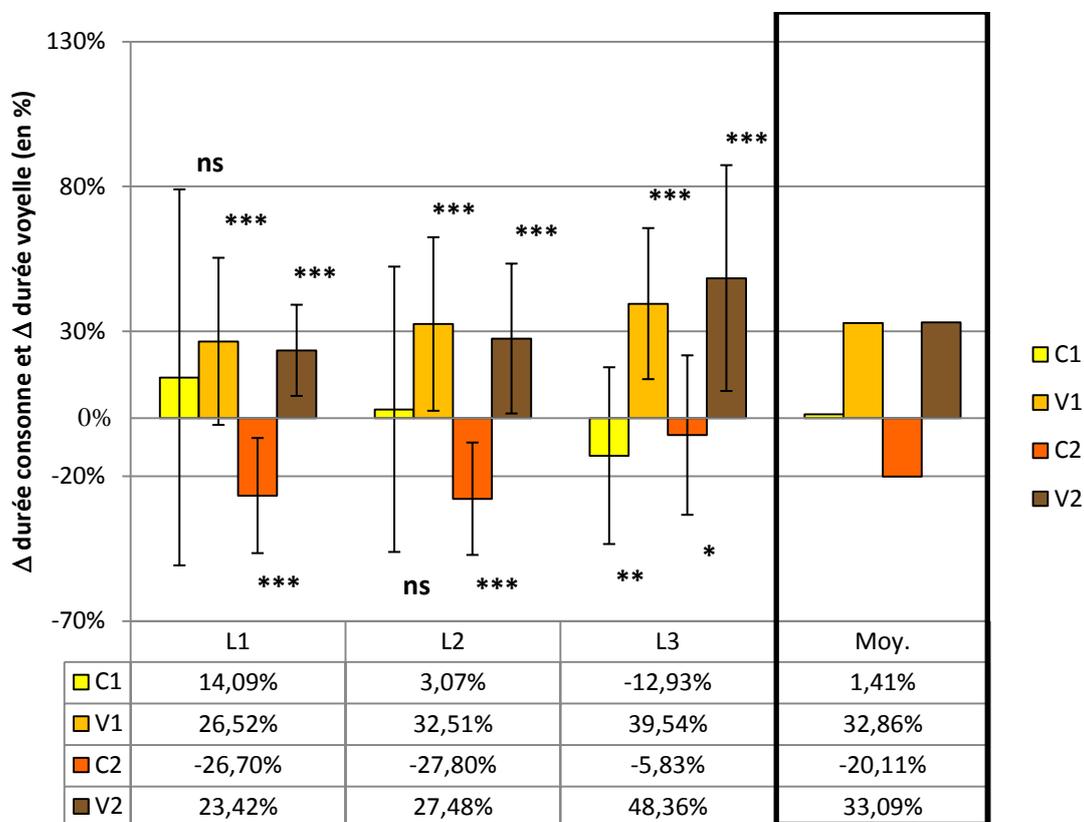


Figure 9-5 : Variations de durée entre la voix parlée et la voix criée des consonnes et voyelles pour les logatomes CVCV

### 9.1.3 Rapport entre la durée des voyelles pour les logatomes VCV et CVCV

Dans le cas des logatomes CVCV, les variations de la durée des voyelles sont identiques pour V1 et V2 pour les locuteurs 1 et 2 mais sont différentes pour le locuteur 3. Dans le cas des logatomes VCV, la seconde voyelle est allongée plus fortement que la première et quelque soit le locuteur. Nous avons souhaité savoir s'il existe une certaine constance quant au rapport de durée entre V2 et V1 pour ces deux types de logatome. Ainsi la Figure 9-7 et Figure 9-6 représentent le rapport entre la durée (en seconde) des secondes voyelles (V2) et des premières voyelles (V1) pour la voix parlée et la voix criée pour les logatomes VCV et CVCV.

Pour les VCV, le rapport de durée entre V2 et V1 varie significativement pour L1 et L3 mais est constant entre la voix parlée et criée pour L2. Dans le cas des CVCV, bien que la valeur du rapport soit dépendante du locuteur, il n'existe pas d'évolution significative de celui-ci entre la voix parlée et la voix criée.

Toutefois, il est difficile de tirer des conclusions générales de ces graphiques. En effet, la constance ou non du rapport peut être liée au fait que nous comparons la durée de voyelles qui ne se situent pas à la même place dans le mot. Par exemple, pour les CVCV les rapports sont plus grands peut être du fait que l'on compare des voyelles inter-consonantiques et des finales, et que pour les VCV on compare des initiales et des finales. Ainsi dans le cas d'une voyelle en position initiale ou finale les temps d'établissements et de relâchements sont comptés contrairement à la voyelle intervocalique où l'établissement ou le relâchement se fait par coarticulation et se propage dans la consonne précédente ou suivante.

Nous penchons toutefois pour une explication tendant à estimer que le rapport entre V2 et V1 est constant, et que sa valeur dépend du locuteur.

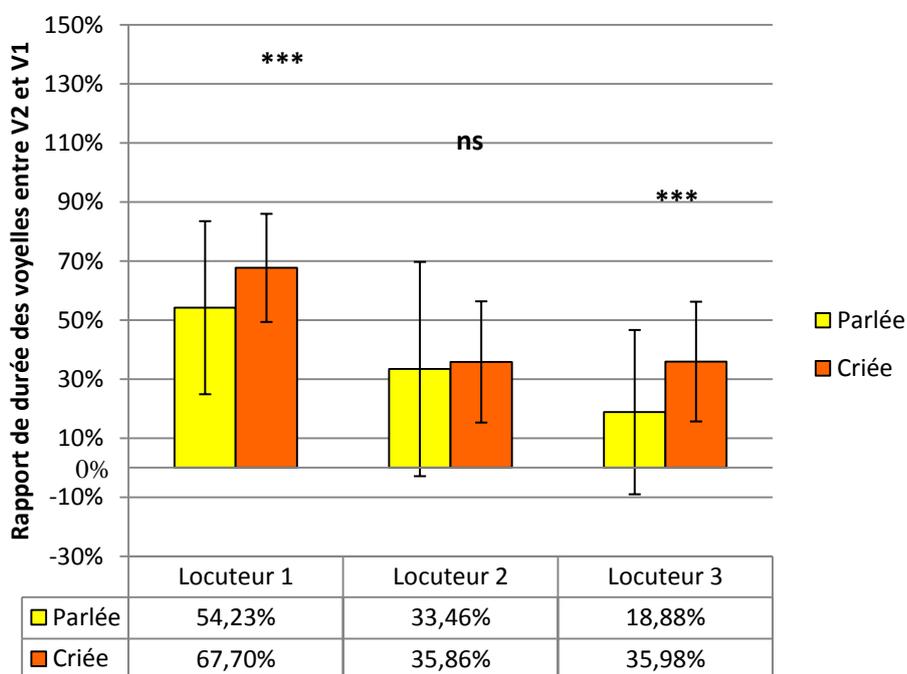


Figure 9-6 : Rapport de la longueur de V2 par rapport à celle de V1 pour les logatomes VCV

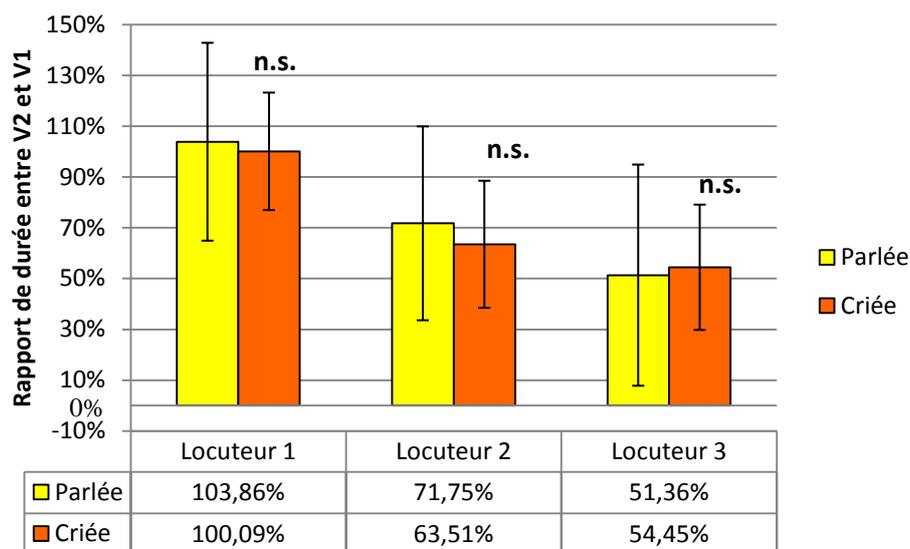


Figure 9-7 : Rapport de la longueur de V2 par rapport à celle de V1 pour les logatomes CVCV

#### 9.1.4 Résumé

De ces analyses de durée globale et pour chaque phonème, plusieurs conclusions peuvent être citées. D'une part, on observe des augmentations significatives (sauf pour les CVC de L1 et L3) de la durée totale des énoncés. En moyenne cette augmentation est de l'ordre de 17 %.

D'autre part pour l'ensemble des locuteurs et des 4 structures étudiées, la durée des voyelles est systématiquement rallongée de manière significative. Le taux de variation est cependant assez varié. Aussi bien en fonction de la structure ainsi que du locuteur. En effet l'augmentation des voyelles est plus forte pour le cas des logatomes CVC et CV. Par contre, lorsque deux voyelles sont présentes dans le mot des variations moins importantes sont observées sur ces voyelles. Le fait que les voyelles sont identiques dans un même mot peut être une explication. En effet, on peut penser que moins d'effort est fourni sur chacune des voyelles étant donné que celle-ci est présente deux fois dans le mot. On estime toutefois qu'une voyelle est en moyenne rallongée de l'ordre de 40%.

Bien qu'une tendance à la diminution de la longueur des consonnes est observée et notamment pour les consonnes intervocaliques les variations sont moins marquées que pour les voyelles. D'autre part, la difficulté liée à segmentation des consonnes non voisées en position initiale et finale, ne permet pas de réaliser une étude précise de ces phénomènes.

## 9.2 Analyses de l'intensité des logatomes

Nous avons constaté dans le chapitre précédent, des différences concernant l'évolution de l'intensité en fonction de la distance de communication (i.e. la dynamique de l'intensité  $\delta I$ ). Cette section se consacre alors à l'analyse des intensités entre les consonnes et les voyelles des structures de base étudiées CV, CVC, VCV et CVCV dans le but de mieux comprendre le phénomène de l'augmentation de la dynamique de l'intensité.

### 9.2.1 Variations relatives de l'intensité des consonnes et des voyelles pour les logatomes CV

La Figure 9-9 représente les moyennes de l'intensité pour les trois locuteurs, en fonction de l'effort vocal (parlée ou criée) ainsi qu'en fonction des consonnes et des voyelles. Pour chaque locuteur, les moyennes et les écarts-types de la voix parlée (traits discontinus) et de la voix criée (trait plein) sont représentés. On constate sur cette figure que pour les voix parlées et criées, les consonnes possèdent des intensités plus faibles que les voyelles. En effet, la différence de l'intensité moyenne entre les consonnes (C1) et les voyelles (V1) en voix parlée est de 5,6 dB SPL, 5,7 dB SPL, et 6,7 dB SPL pour les locuteurs 1, 2 et 3. Pour les voix criées ces différences sont toutefois plus intenses et sont respectivement de 14,3 dB SPL, 13,4 dB SPL et 17 dB SPL (cf. Tableau 9-1)

Tableau 9-1: Tableau récapitulatif des différences d'intensité entre les phonèmes pour les logatomes CV (en dB SPL)

	L1		L2		L3	
	Parlée	Criée	Parlée	Criée	Parlée	Criée
$I_{V1}-I_{C1}$	5,6	14,3	5,7	13,4	6,7	17

On constate alors sur la Figure 9-8 que les variations d'intensité entre consonnes et voyelles en voix parlée et criée ne sont pas identiques. En effet, cette figure montre les variations d'intensité ( $\Delta I$ ) des consonnes et des voyelles pour chaque locuteur ainsi que la moyenne des trois locuteurs. On remarque que les voyelles sont plus amplifiées que les consonnes et que cette différence est significative. En effet l'augmentation de l'intensité des voyelles est plus intense que celle des consonnes de 8,7 dB SPL, 7,7 dB SPL et 10,3 dB SPL pour les locuteurs 1, 2 et 3. (cf. Tableau 9-2)

Tableau 9-2: Tableau récapitulatif des différences de variation d'intensité entre les phonèmes pour les logatomes CV (en dB SPL)

	L1	L2	L3
$\Delta I_{V1}-\Delta I_{C1}$	8,7	7,7	10,3

On remarquera que si les valeurs obtenues pour L1 et L2 sont relativement identiques, L3 semble produire plus d'effort. En effet, les variations obtenues pour L3 sont plus intenses. Ceci s'explique par le fait que ce locuteur 3 produit d'une part, des voix criées plus intenses que les 2 autres locuteurs, mais produit également une voix parlée moins forte que les deux autres locuteurs, impliquant ainsi des variations plus intenses entre voix parlée et criée.

Ainsi, outre l'augmentation de l'intensité globale des CV, lors d'un cri, on observe également des variations d'intensité plus fortes pour les voyelles que les pour les consonnes. Les voyelles sont en moyenne plus intenses que les consonnes de 8,9 dB SPL (soit 2,8 fois plus intense).

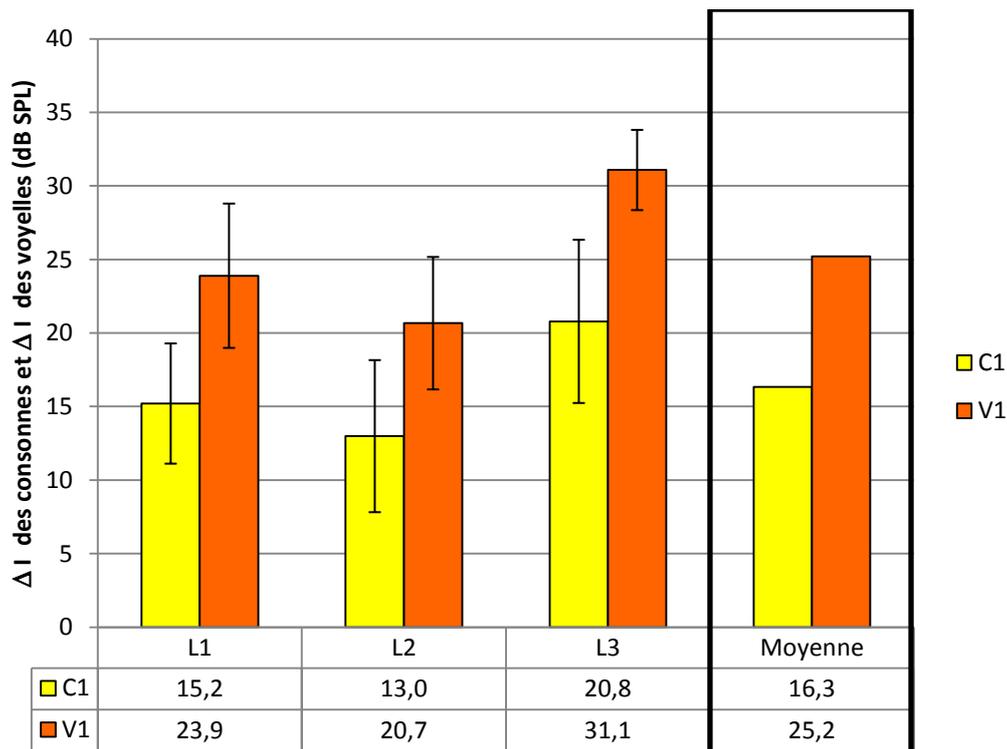


Figure 9-8: Variations d'intensité entre voix modale et voix projetée pour les consonnes et voyelles des logatomes CV

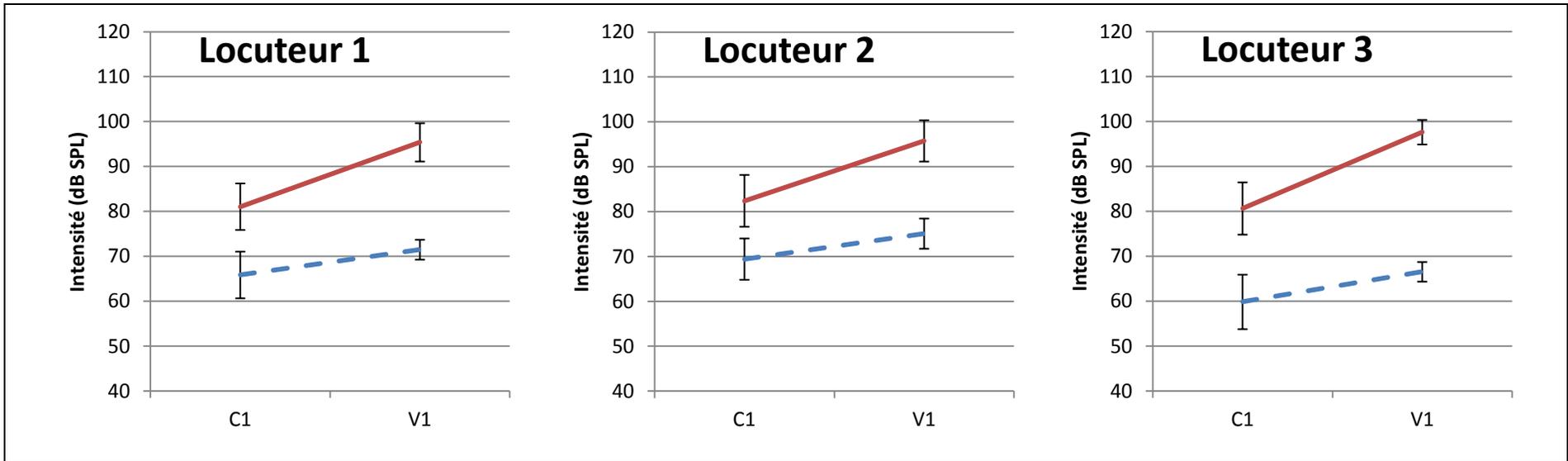


Figure 9-9 : Valeur moyenne de l'intensité pour chaque phonème et locuteur des logotomes CV. Les traits en pointillés correspondent à la voix parlée.

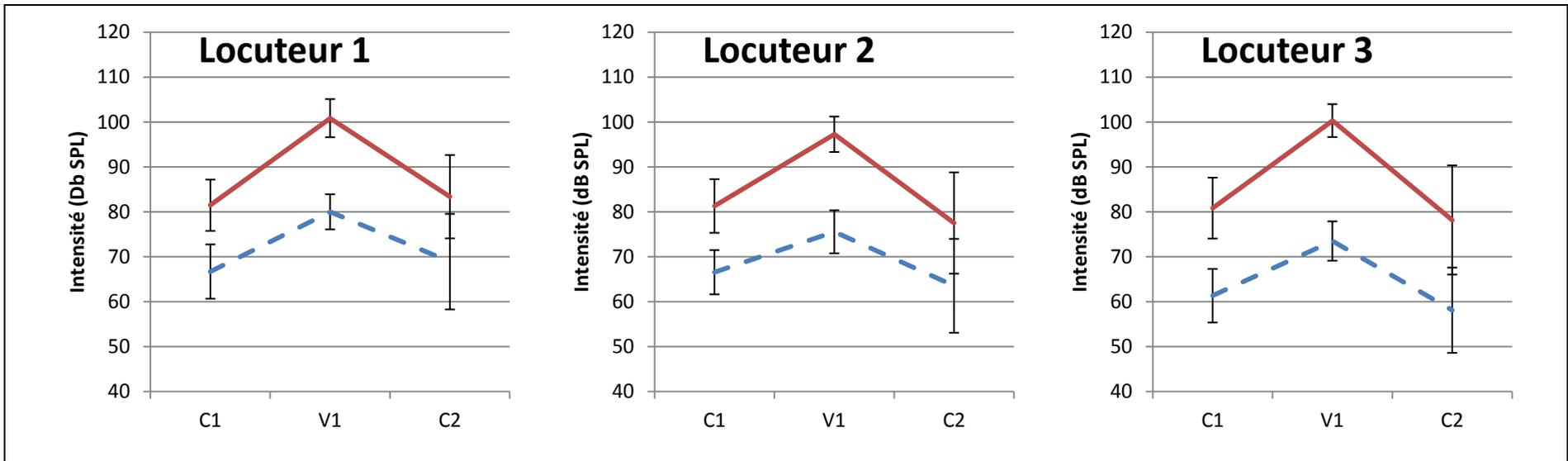


Figure 9-10: Valeur moyenne de l'intensité pour chaque phonème et locuteur des logotomes CVC. Les traits en pointillés correspondent à la voix parlée.

### 9.2.2 Variations relatives de l'intensité des consonnes et des voyelles pour les logatomes CVC

Dans le cas des CVC on observe des faits similaires à ceux observés pour les logatomes CV. La Figure 9-10 représente l'intensité moyenne en voix parlée et en voix criée pour les 3 locuteurs. On observe clairement des intensités plus fortes sur les voyelles que sur les consonnes dans les deux types de phonation.

En effet dans le cas des voix parlées l'intensité moyenne des voyelles (V1) est plus forte que celle des premières consonnes (C1) de 13,3 dB SPL, 9 dB SPL et 12,2 dB SPL pour L1, L2 et L3. En voix criée ces écarts sont plus larges et sont respectivement de 19,4 dB SPL, 16 dB SPL et 19,5 dB. (cf. Tableau 9-3)

De la même manière, l'intensité moyenne des voyelles (V1), en voix parlée, est plus intense que celle des dernières consonnes (C2) de 11 dB SPL, 12 dB SPL et 15,4 dB SPL pour L1, L2 et L3. Ces écarts d'intensité moyenne, sont dans le cas des voix criées, respectivement de 17,5 dB SPL, 16,7 dB SPL et 22,1 dB SPL. (cf. Tableau 9-3)

Ainsi, en voix criée, les écarts entre l'intensité moyenne des consonnes et des voyelles augmentent d'environ 6 dB SPL. Toutefois l'écart d'intensité entre la première consonne (C1) et la seconde (C2) ne varie pas ou peu. (cf. Tableau 9-3)

Tableau 9-3: Tableau récapitulatif des différences d'intensité entre les phonèmes pour les logatomes CVC (en dB SPL)

	L1		L2		L3	
	Parlée	Criée	Parlée	Criée	Parlée	Criée
$I_{V1}-I_{C1}$	13,3	19,4	9	16	12,2	19,5
$I_{V1}-I_{C2}$	11	17,5	12	16,7	15,4	22,1
$I_{C2}-I_{C1}$	2,2	1,9	-3	-3,8	-3,2	-2,6

Ainsi, d'un point de vue relatif, la variation d'intensité des voyelles est plus forte que celle des consonnes. Ce phénomène peut également s'observer sur la Figure 9-11 qui représente les variations d'intensité moyenne pour chaque phonème et chaque locuteur. Bien que pour le locuteur 3 ces variations soit plus intenses, les proportions de variation entre consonnes et voyelles sont similaires pour les trois locuteurs.

En effet, pour le locuteur 1 les voyelles (V1) augmentent de 6,1 dB de plus que les premières consonnes (C1) et de 6,4 dB SPL de plus que les deuxièmes consonnes (C2). Pour le locuteur 2 ces valeurs sont de 7 dB et de 7,7 dB et pour le locuteur 3 elles sont de 7,3 dB et 6,7 dB. En moyenne, les voyelles augmentent de **6,9 dB** (soit des voyelles plus intenses de 2,2 fois par rapport aux consonnes).

Ainsi dans le cas des CVC, les variations sont légèrement moins intenses que celle observée pour les logatomes CV, et ce même sur la voix parlée.

Tableau 9-4: Tableau récapitulatif des différences de variation d'intensité entre les phonèmes pour les logatomes CVC (en dB SPL)

	L1	L2	L3
$\Delta I_{V1}-\Delta I_{C1}$	6,1	7	7,3
$\Delta I_{V1}-\Delta I_{C2}$	6,4	7,7	6,7
$\Delta I_{C2}-\Delta I_{C1}$	-0,3	-0,8	0,6

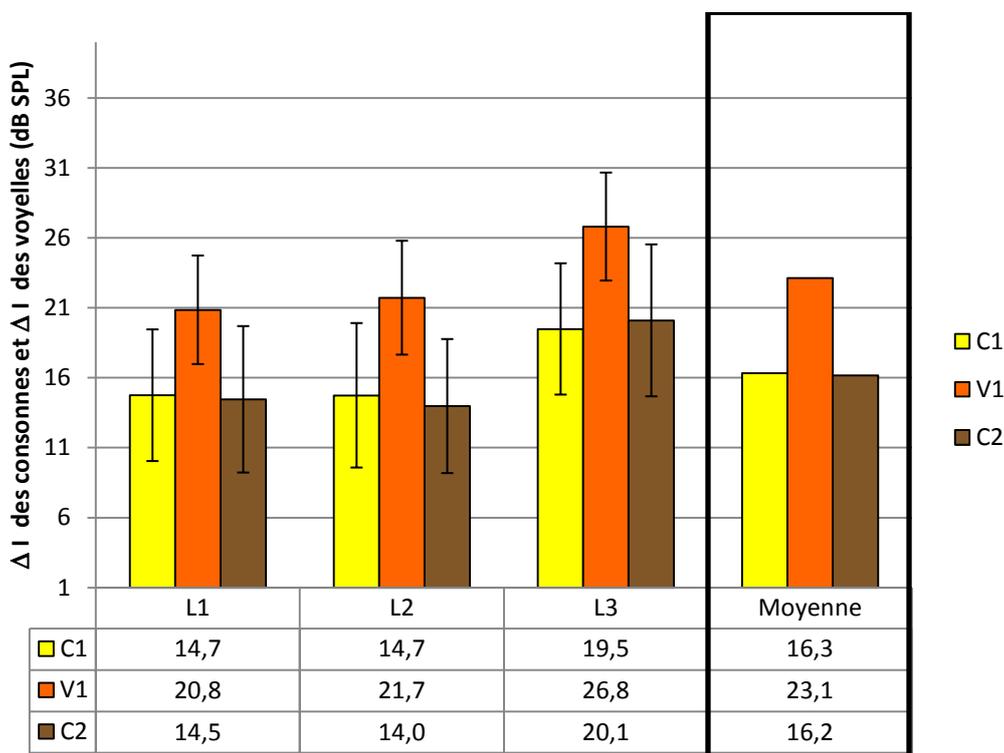


Figure 9-11 : Variations d'intensité entre voix modale et voix projetée pour les consonnes et voyelles des logatomes CVC

### 9.2.3 Variations relatives de l'intensité des consonnes et des voyelles pour les logatomes VCV

Pour le cas des logatomes VCV, les variations d'intensité entre consonnes et voyelles sont moins flagrantes. Toutefois on retrouve une logique quant aux intensités des voyelles. La Figure 9-13 représente les valeurs moyennes de l'intensité pour les deux niveaux de phonation ainsi que pour les trois locuteurs. Comme précédemment les voyelles semblent augmenter plus fortement que les consonnes.

En effet la différence d'intensité moyenne entre les premières voyelles (V1) et les consonnes intervocaliques (C1), en voix parlée, est de 4,1 dB SPL, 3,8 dB SPL et 3,6 dB SPL pour L1, L2 et L3, alors qu'en voix criée ces valeurs sont de 7 dB SPL, 9 dB SPL et de 7,4 dB SPL. (cf. Tableau 9-5)

De la même manière, entre les deuxièmes voyelles (V2) et les consonnes (C1), en voix parlée, les écarts sont de 5,6 dB SPL de 2 dB SPL et de 3,1 dB alors qu'en voix criée ils sont de 10,7 dB 8,5 dB SPL et de 12,4 dB SPL. (cf. Tableau 9-5)

De plus, l'intensité des deuxièmes voyelles (V2) par rapport à celle des premières (V1) tend à augmenter, ce qui signifie que l'intensité tend à augmenter au cours de la prononciation contrairement aux voix parlées (cf. Tableau 9-5). Ceci est sans doute dû à une stratégie du locuteur visant à mieux marquer l'accent final.

Tableau 9-5: Tableau récapitulatif des différences d'intensité entre les phonèmes pour les logatomes VCV (en dB SPL)

	L1		L2		L3	
	Parlée	Criée	Parlée	Criée	Parlée	Criée
$I_{V1}-I_{C1}$	4,1	7	3,8	9	3,6	7,4
$I_{V2}-I_{C1}$	5,6	10,7	2	8,5	3,1	12,4
$I_{V2}-I_{V1}$	1,5	3,7	-1,7	-0,5	-0,5	5

On constate d'ores et déjà des différences plus marquées entre les locuteurs. En effet les variations de chaque phonème sont moins homogènes que dans les cas précédents (cf. Figure 9-12).

Sur la Figure 9-12, on observe que les premières voyelles (V1) augmentent de façon plus intense que les premières consonnes (C1) de 2,9 dB SPL, 5,2 dB SPL et 3,7 dB SPL pour L1, L2 et L3 tandis que les variations entre les consonnes C1 et les deuxièmes voyelles (V2) sont respectivement de 5 dB SPL, 6,4 dB SPL et de 9 dB. (cf. Tableau 9-6)

Ainsi, pour les trois locuteurs on constate une augmentation plus forte pour les secondes voyelles (V2) que pour les premières (V1). Les différences sont de l'ordre de 3,1 dB SPL pour L1, 1,2 dB SPL pour L2 et de 5,3 dB SPL pour L3. Sachant que l'augmentation est observée dans des propositions différentes pour les trois locuteurs, il est difficile de conclure sur l'augmentation de l'intensité des deuxièmes voyelles par rapport aux premières. Néanmoins une augmentation est observée. Le cas des logatomes CVCV nous apportera peut être plus de réponse à ce sujet.

Par rapport aux logatomes précédents, les variations observées ici sont bien moins intenses que celles observées pour les logatomes CVC et se rapprochent plus de celles observées pour les logatomes CV.

Tableau 9-6 : Tableau récapitulatif des différences de variation d'intensité entre les phonèmes pour les logatomes VCV (en dB SPL)

	L1	L2	L3
$\Delta I_{V1}-\Delta I_{C1}$	2,9	5,2	3,7
$\Delta I_{V2}-\Delta I_{C1}$	5	6,4	9
$\Delta I_{V2}-\Delta I_{V1}$	3,1	1,2	5,3

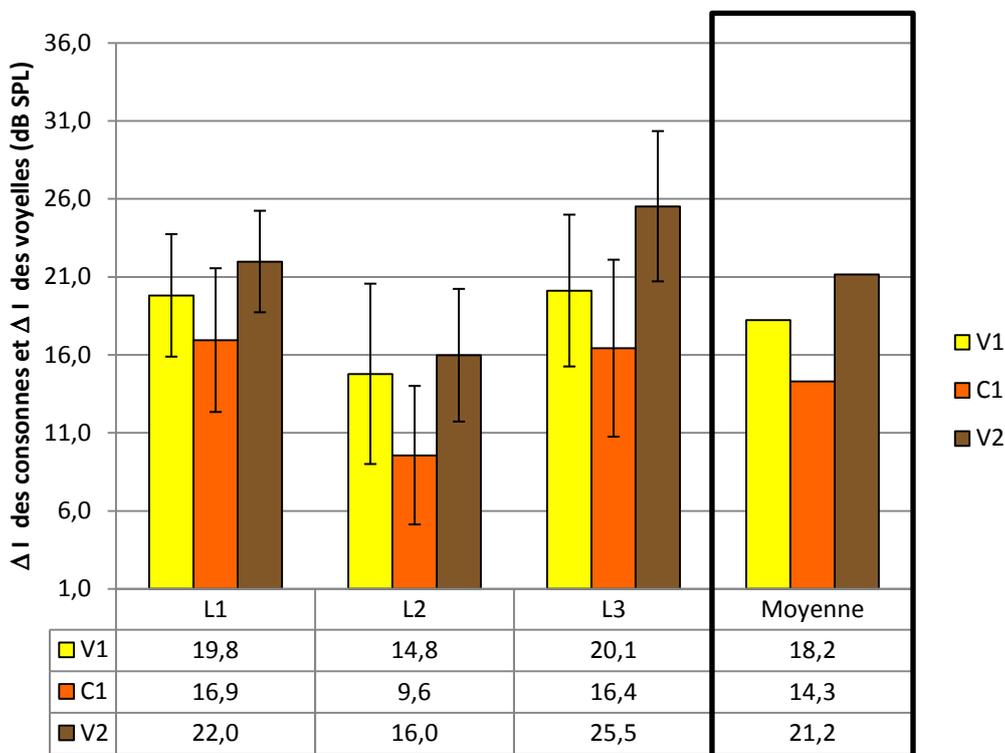


Figure 9-12: Variations d'intensité entre voix modale et voix projetée pour les consonnes et voyelles des logatomes VCV

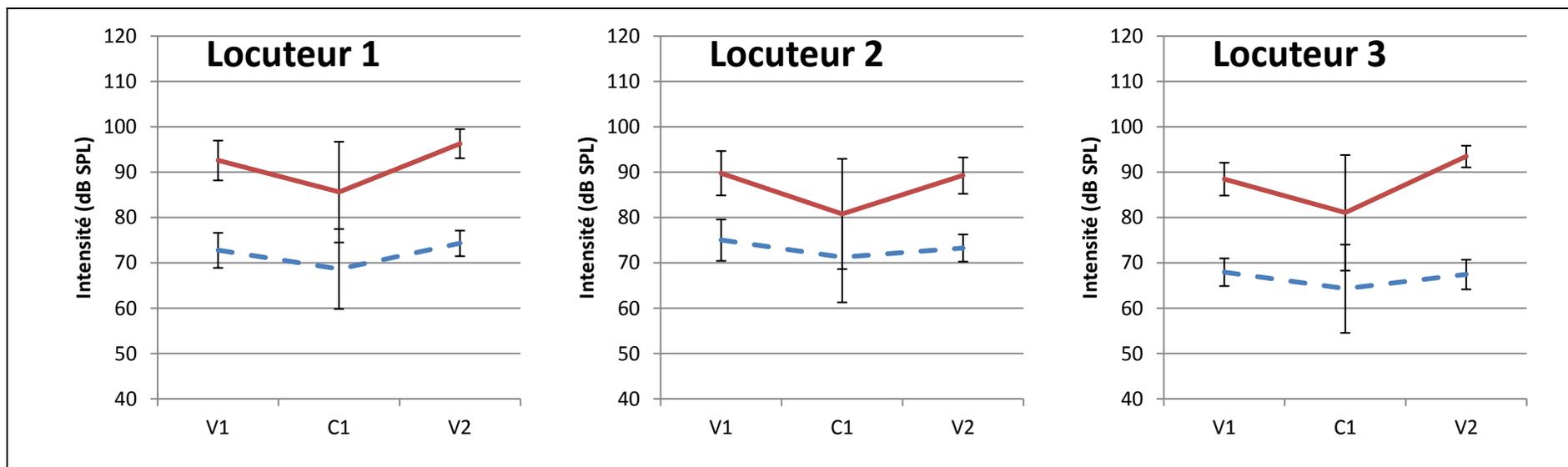


Figure 9-13 : Valeur moyenne de l'intensité pour chaque phonème et locuteur des logatomes VCV. Les traits en pointillés correspondent à la voix parlée.

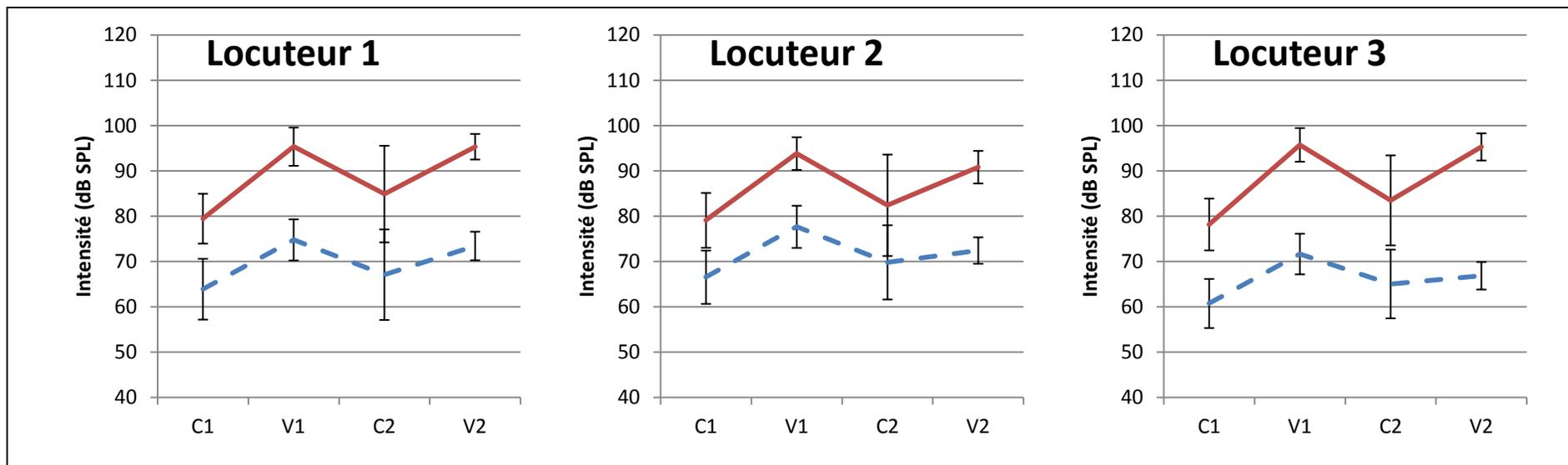


Figure 9-14 : Valeur moyenne de l'intensité pour chaque phonème et locuteur des logatomes CVCV. Les traits en pointillés correspondent à la voix parlée.

### 9.2.4 Variations relatives de l'intensité des consonnes et des voyelles pour logatomes CVCV

Considérons enfin le cas des logatomes CVCV. La Figure 9-14 représente les moyennes de l'intensité pour la voix parlée et la voix criée pour les trois locuteurs. On observe à nouveau, des variations plus fortes sur les voyelles que sur les consonnes.

En effet, la différence d'intensité en voix parlée entre les premières consonnes (C1) et les premières voyelles (V1) est de 10,9 dB SPL, 11,1 dB SPL et de 10,9 dB SPL pour L1, L2 et L3 alors qu'en voix criée elle est de 15,9 dB SPL, 14,7 dB SPL et de 17,6 dB SPL. (cf. Tableau 9-7)

De la même manière, entre les premières voyelles (V1) et les deuxièmes consonnes (C2) les différences en voix parlée sont de 7,7 dB SPL, 7,8 dB SPL et de 6,6 dB SPL tandis qu'en voix criée elles sont de 10,5 dB SPL, 11,4 dB SPL et de 12,2 dB SPL pour les 3 locuteurs. (cf. Tableau 9-7)

Entre les deuxièmes voyelles (V2) et les deuxièmes consonnes (C2) on note toutefois des variations moins constantes. En voix parlée elles sont de 6,3 dB SPL, 2,6 dB SPL et de 1,8 dB SPL mais augmentent également en voix criée pour atteindre des valeurs de 10,4 dB SPL, 8,4 dB SPL et de 11,8 dB SPL. On note également que l'écart d'intensité entre les premières et les deuxièmes voyelles est plus faible en voix criée et que l'écart entre les deux consonnes tend à augmenter légèrement. (cf. Tableau 9-7)

Tableau 9-7: Tableau récapitulatif des différences d'intensité entre les phonèmes pour les logatomes CVCV (en dB SPL)

	L1		L2		L3	
	Parlée	Criée	Parlée	Criée	Parlée	Criée
$I_{V1}-I_{C1}$	10,9	15,9	11,1	14,7	10,9	17,6
$I_{V1}-I_{C2}$	7,7	10,5	7,8	11,4	6,6	12,2
$I_{V2}-I_{C2}$	6,3	10,4	2,6	8,4	1,8	11,8
$I_{V2}-I_{V1}$	-1,3	0,0	-5,2	-3,0	-4,8	-0,5
$I_{C2}-I_{C1}$	3,2	5,4	3,3	3,3	4,3	5,3

Ainsi des variations entre intensité des consonnes et des voyelles sont observées entre une voix parlée et une voix criée (cf. Figure 9-15). Notons entre autre, que les premières voyelles (V1) augmentent de 5 dB SPL, 3,6 dB SPL et de 6,7 dB SPL de plus que les premières consonnes (C1).

Entre les premières voyelles (V1) et les deuxièmes consonnes (C2) les écarts de variations d'intensité sont de 2,8 dB SPL, 3,6 dB SPL, et 5,6 dB SPL pour les trois locuteurs. En toute logique les deuxièmes voyelles (V2) par rapport aux deuxièmes consonnes (C2) augmentent de 4,1 dB SPL, 5,8 dB SPL et de 10 dB SPL de plus.

Notons également que la variation des deuxièmes voyelles (V2) est bien plus forte que celle des premières (V1). On trouve des écarts de variation de 6,3 dB SPL 5,8 dB SPL et de 11 dB SPL pour les locuteurs. Hormis pour le locuteur 1 les variations entre l'intensité des deuxièmes consonnes (C2) sont égale à celles des premières consonnes (C1).

Une fois encore, la variation d'intensité des voyelles est plus forte que celle des consonnes. La variation des deuxièmes voyelles (V2) est également plus intense que la variation des premières (V1) ce qui tend à rendre la ligne de déclinaison d'intensité nulle voir même positive.

Tableau 9-8 : Tableau récapitulatif des différences de variation d'intensité entre les phonèmes pour les logatomes CVCV (en dB SPL)

	L1	L2	L3
$\Delta I_{V1} - \Delta I_{C1}$	5,0	3,6	6,7
$\Delta I_{V1} - \Delta I_{C2}$	2,8	3,6	5,6
$\Delta I_{V2} - \Delta I_{C2}$	4,1	5,8	10,0
$\Delta I_{V2} - \Delta I_{V1}$	2,2	0,1	1,0
$\Delta I_{C2} - \Delta I_{C1}$	6,3	5,8	11,0

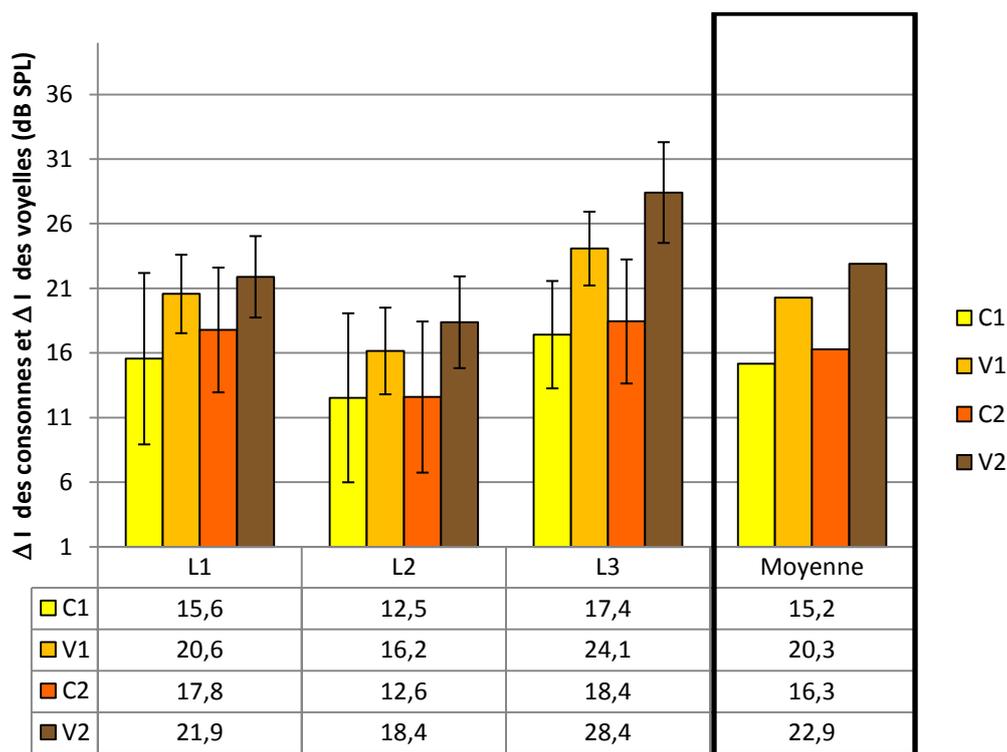


Figure 9-15: Variations d'intensité entre voix modale et voix criée pour les consonnes et voyelles des logatomes CVCV

### 9.2.5 Résumé

Deux phénomènes ont été mis à jour par l'étude des variations d'intensité. Le premier est que les voyelles subissent une augmentation en intensité bien supérieure à celle des consonnes. En effet, en moyenne, les voyelles sont augmentées d'environ 6 dB SPL de plus que les consonnes. Le rapport d'intensité entre la voyelle et la consonne est alors, en voix criée, fortement augmenté par rapport à la voix parlée. D'autre part, l'augmentation de la deuxième voyelle est régulièrement plus intense que l'augmentation de la première voyelle. En effet on observe des variations de 0 à 5 dB SPL. Toutefois on observe que les intensités des voyelles V1 et V2 tendent à s'égaliser et que dans certain cas V2 est plus intense que V1 (cas des logatomes VCV).

## 9.3 Analyses de la fréquence fondamentale des logatomes

---

Nous avons déjà pu constater, comme bon nombre d'études que la valeur moyenne de F0 augmente pour une voix criée. D'autre part, étant donné que la dynamique de F0 change également avec l'effort vocal, une analyse précise des variations de F0 a été faite. Cette analyse est effectuée sur les structures phonétiques de base en considérant les valeurs de F0 des consonnes voisées séparément des voyelles. Nous tentons ici de mieux comprendre les phénomènes de l'augmentation de la dynamique de F0. Les valeurs sont exprimées en demi-tons avec pour référence une fréquence de 50 Hz.

### 9.3.1 Les variations relatives de F0 des consonnes et des voyelles pour les logatomes CV

La Figure 9-17 représente les valeurs moyennes de F0 des énoncés CV pour les trois locuteurs. Les moyennes sont représentées en fonction du phonème (consonnes ou voyelles). Pour chaque locuteur, les moyennes et les écarts-types de la voix parlée (traits discontinus) et de la voix criée (traits pleins) sont représentés. La première constatation concerne les voix parlées.

On remarque qu'en voix parlée les locuteurs 1 et 2 augmentent considérablement la F0 des voyelles (V1) par rapport à celle des consonnes (C1) (7,5 demi-tons pour L1, 4,9 demi-tons pour L2). Le locuteur 3 en revanche, conserve une F0 relativement constante entre la consonne et la voyelle (0,3 demi-ton, n.s.). Les moyennes observées pour les voix criées sont en revanche homogènes. (cf. Tableau 9-17)

En effet, dans le cas d'un effort vocal, les voyelles (V1) ont des F0 plus élevées que les consonnes (C1) pour les 3 locuteurs. On observe des variations de 10,2 demi-tons, 9,2 demi-tons et 9,3 demi-tons, pour le locuteur 1, 2 et 3. On remarque que le locuteur 3 réalise des F0 moyennes plus élevées que les deux autres locuteurs mais que les variations entre consonnes et voyelles sont identiques. (cf. Tableau 9-17)

Tableau 9-9: Tableau récapitulatif des différences de F0 entre les phonèmes pour les logatomes CV (en demi-ton)

	L1		L2		L3	
	Parlée	Criée	Parlée	Criée	Parlée	Criée
$F0_{V1}-F0_{C1}$	7,5	10,2	4,9	9,2	0,3	9,3

Ainsi, pour les logatomes CV en voix criée, on constate une plus forte augmentation de la F0 sur les voyelles que sur les consonnes voisées par rapport aux voix parlées (cf. Figure 9-16). Ces augmentations sont similaires pour les locuteurs L1 et L2. Pour le locuteur L3, la F0 des consonnes et des voyelles augmente plus. Alors que pour les premiers locuteurs les variations étaient de l'ordre de 7 demi-tons pour les consonnes et 10 demi-tons pour les voyelles, le troisième locuteur quant à lui augmente les consonnes de 10 demi-tons et près de 19 demi-tons pour les voyelles. Les différences de variations entre locuteurs sont également différentes. On note que la F0 des voyelles augmente plus que celle des consonnes de 2,6 demi-tons pour L1, 4,3 demi-tons pour L2 et 5,3 demi-tons pour L3. (cf. Tableau 9-18).

La variation de la F0 des voyelles est bien plus intense pour L3. En effet, en voix parlée, L3 n'a pas réalisé une montée de F0 en fin d'énoncé contrairement aux locuteurs 1 et 2 (i.e. courbe intonative). Ce locuteur a également produit des voyelles ayants des F0 plus élevées que les 2 autres locuteurs. C'est pourquoi, la variation de F0 des voyelles est plus élevée pour le locuteur 3.

Ainsi pour les 3 locuteurs nous observons des variations plus intenses de F0 pour les voyelles que pour les consonnes entre une voix parlée et une voix criée.

Tableau 9-10: Tableau récapitulatif des différences de variation de F0 entre les phonèmes pour les logatomes CV (en demi-ton)

	L1	L2	L3
$\Delta F0_{V1}-\Delta F0_{C1}$	2,6	4,3	5,3

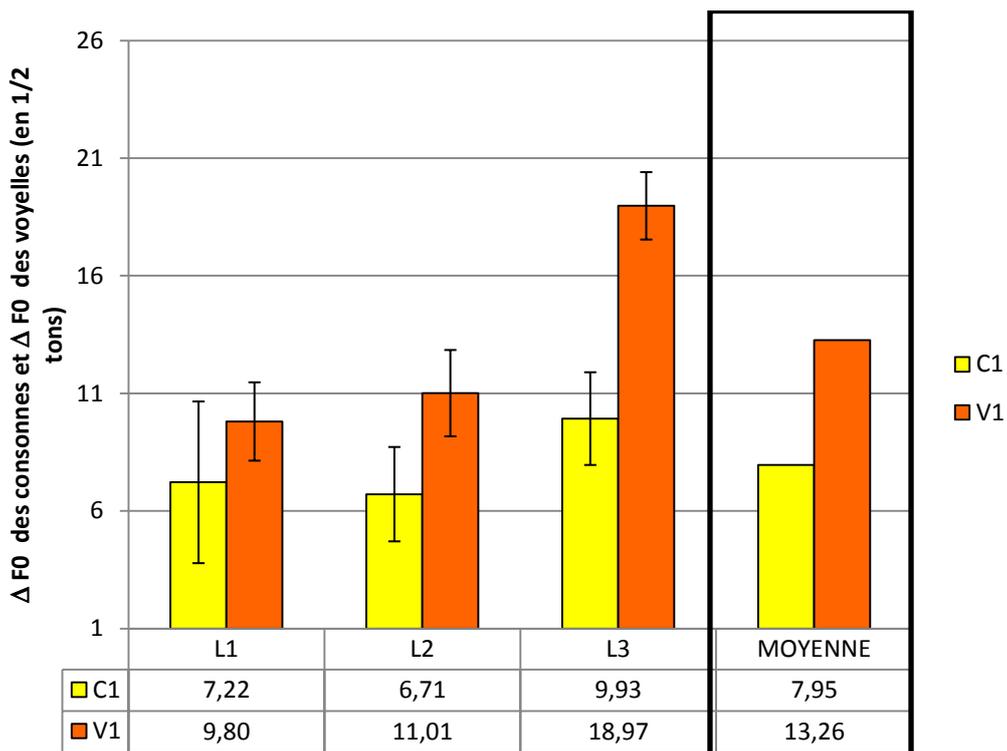


Figure 9-16: Variations de F0 entre voix modale et voix criée pour les consonnes et voyelles des logatomes CV

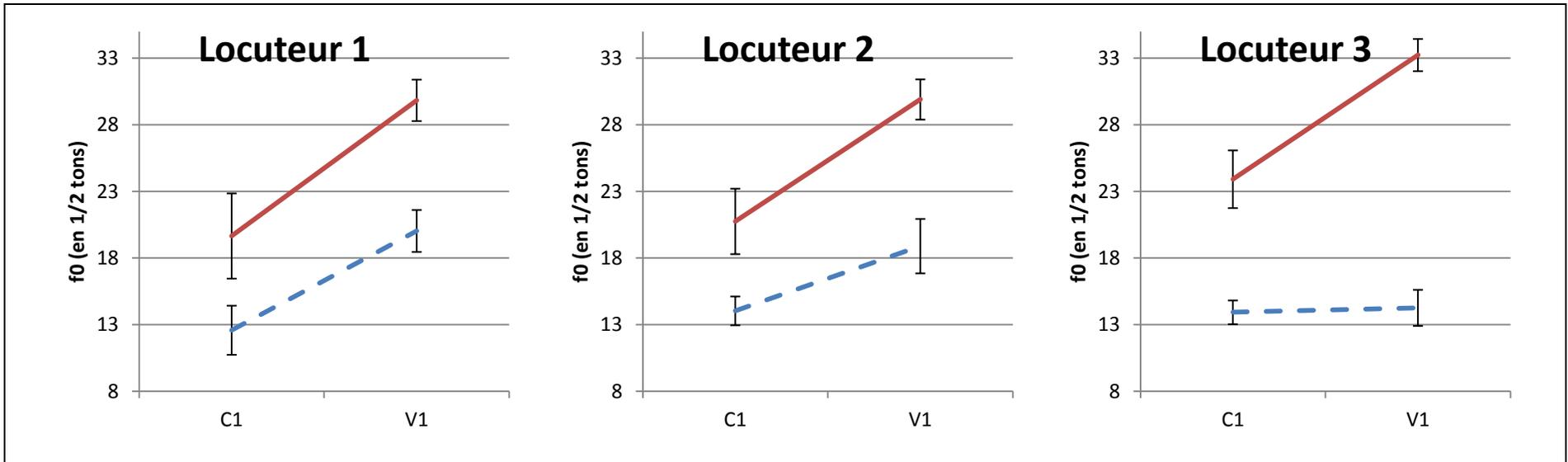


Figure 9-17 : Valeur moyenne de F0 pour chaque phonème et locuteurs des logatomes CV. Les traits en pointillés correspondent à la voix parlée.

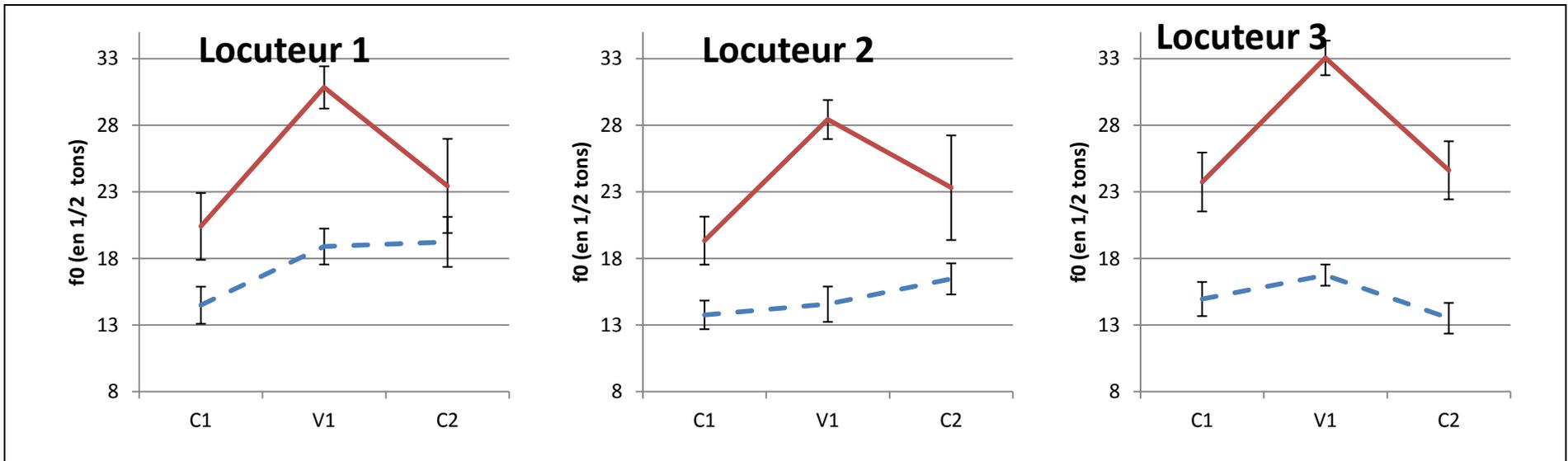


Figure 9-18 : Valeur moyenne de F0 pour chaque phonème et locuteurs des logatomes CVC. Les traits en pointillés correspondent à la voix parlée.

### 9.3.2 Les variations relatives de F0 des consonnes et des voyelles pour les logatomes CVC

Considérons à présent le cas des énoncés CVC. Sur Figure 9-18 nous pouvons faire les mêmes remarques que précédemment. A savoir que les moyennes pour les voix parlées sont différentes entre les locuteurs mais que les voix criées se ressemblent étrangement.

En effet, on observe pour les voix parlées une augmentation forte de F0 entre les consonnes initiales (C1) et les voyelles inter-consonantiques (V1) pour L1 (+4,4 demi-tons) et des augmentations moins prononcées pour L2 (+0,8 demi-ton) et L3 (+1,8 demi-ton). Les consonnes finales (C2) quant à elles, par rapport aux consonnes initiales (C1) augmentent de 4,8 demi-tons pour L1 et de 2,7 demi-tons pour L2. En revanche, pour L3, F0 diminue de 1,4 demi-ton (cf. Tableau 9-3).

Tableau 9-11: Tableau récapitulatif des différences de F0 entre les phonèmes pour les logatomes CVC (en demi-ton)

	L1		L2		L3	
	Parlée	Criée	Parlée	Criée	Parlée	Criée
$F0_{V1}-F0_{C1}$	4,4	10,4	0,8	9,1	1,8	9,3
$F0_{V1}-F0_{C2}$	-0,3	7,4	-1,9	5,1	3,2	8,4
$F0_{C2}-F0_{C1}$	4,8	3	2,7	4	-1,4	0,9

Pour les voix criées en revanche, les voyelles (V1) sont fortement accentuées par rapport aux consonnes initiales (C1). On relève des variations de 10,4 demi-tons, 9,1 demi-tons et de 9,3 demi-tons pour L1, L2 et L3. Les consonnes finales sont également plus faibles que les voyelles mais ne diminuent pas autant que les premières consonnes (C1) (7,4 demi-tons, 5,1 demi-tons et 8,4 demi-tons). Les différences entre les consonnes initiales (C1) et finales (C2) sont en voix criée de 3 demi-tons, 4 demi-tons et 0,9 demi-ton, pour L1, L2 et L3 (cf. Tableau 9-3).

Pour les CVC, on constate des évolutions de la F0 des consonnes, qui sont similaires à celles observées pour les CV (cf. Figure 9-19). Les voyelles sont toutefois légèrement plus affectées que lors des CV. On trouve des différences entre consonnes et voyelles d'un écart moyen de 6 demi-tons (cf. Tableau 9-4). De plus, les consonnes finales (C2) augmentent plus que les consonnes initiales (C1) pour 2 locuteurs.

Tableau 9-12: Tableau récapitulatif des différences de variation de F0 entre les phonèmes pour les logatomes CVC (en demi-ton)

	L1	L2	L3
$\Delta F0_{V1}-\Delta F0_{C1}$	6	8,3	7,5
$\Delta F0_{V1}-\Delta F0_{C2}$	7,8	6,8	5,2
$\Delta F0_{C2}-\Delta F0_{C1}$	-1,7	1,5	2,3

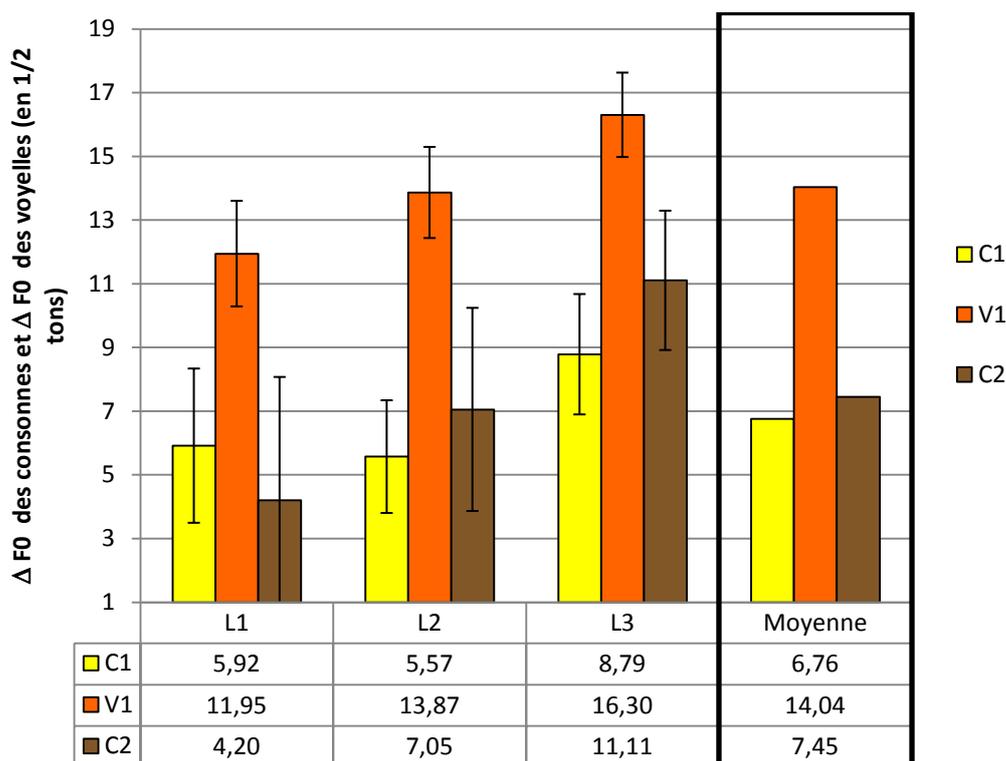


Figure 9-19 : Variations de F0 entre voix modale et voix criée pour les consonnes et voyelles des logatomes CVC

### 9.3.3 Les variations relatives de F0 des consonnes et des voyelles pour les logatomes VCV

La Figure 9-21 montre les valeurs moyennes de F0 pour les énoncés VCV. Concernant les valeurs moyennes des voix parlées, celles-ci sont différentes pour les trois locuteurs. Le locuteur 1, entre les premières voyelles (V1) et les consonnes (C1), maintient la F0 aux mêmes niveaux. Les locuteurs 2 et 3 possèdent des F0 plus faibles sur les consonnes. Il s'agit toutefois de faibles variations, de l'ordre de 1 demi-ton entre C1 et V1. Les dernières voyelles (V2) semblent toutefois être plus accentuées pour au moins deux locuteurs (L1 et L2). En effet par rapport aux premières voyelles, les deuxièmes possèdent une F0 plus élevée de 7,5 demi-tons pour L1 mais de seulement 1,4 demi-ton pour L2.

Les voix criées contrairement à ce qui a été observé sur les structures CV, possèdent également quelques différences. Les consonnes intervocaliques augmentent légèrement pour L1 et L2 (environ 1 demi-ton) par rapport (C1) aux premières voyelles (V1). En revanche, les voyelles finales (V2) sont accentuées et notamment pour deux locuteurs (L1 et L3). Les accentuations sont de l'ordre de 4,9 demi-tons, 1,4 demi-ton et 4,2 demi-tons, pour L1, L2 et L3. Ainsi, hormis le niveau des voyelles finales (V2) de L2, les trois courbes moyennes des voix criées sont très ressemblantes pour les trois

locuteurs considérés. A savoir une F0 moyenne identique pour V1 et C1 et une augmentation réalisée sur V2, ce qui n'est pas le cas pour les voix parlées.

Tableau 9-13: Tableau récapitulatif des différences de F0 entre les phonèmes pour les logatomes VCV (en demi-ton)

	L1		L2		L3	
	Parlée	Criée	Parlée	Criée	Parlée	Criée
$F0_{V1}-F0_{C1}$	-0,2	0,9	1,5	1	1,2	0,1
$F0_{V2}-F0_{C1}$	7,2	4	2,9	0,4	1,4	4,1
$F0_{V2}-F0_{V1}$	7,5	4,9	1,4	1,4	0,2	4,2

Dans le cas des énoncés **VCV**, on observe cette fois une forte variation de la F0 pour les consonnes intervocaliques (cf. Figure 9-20). Les consonnes augmentent de 13 demi-tons. Pour les locuteurs L1 et L2, les voyelles V1 et V2 augmentent moins que les consonnes intervocaliques (C1). On note également que les deuxièmes voyelles (V2) augmentent moins fortement que les premières (V1) pour L1 et que des variations similaires sont observées pour V1 et V2 pour le locuteur 2. Pour L3 en revanche, V2 est fortement augmentée (15 demi-tons). Toutefois, au regard des valeurs moyennes de la Figure 9-21, les variations observées pour L2 ou L3 sont plus révélatrices que celles de L1 qui présentent une forte forme interrogative en voix parlée. Ainsi les variations relevées pour L2 et L3 sont plus correctes.

Contrairement aux cas présentés jusqu'ici, les valeurs et les courbes moyennes (en voix criée notamment) sont moins révélatrices d'une stratégie qui pourrait être partagée par l'ensemble des locuteurs. L'étude des logatomes CVCV nous permettra peut être d'apporter des réponses concernant ce type de divergences interlocuteurs.

Tableau 9-14 : Tableau récapitulatif des différences de variation d'intensité entre les phonèmes pour les logatomes VCV (en dB SPL)

	L1	L2	L3
$\Delta F0_{V1}-\Delta F0_{C1}$	-0,7	-2,5	-1,4
$\Delta F0_{V2}-\Delta F0_{C1}$	-3,3	-2,5	2,6
$\Delta F0_{V2}-\Delta F0_{V1}$	-2,5	0	4

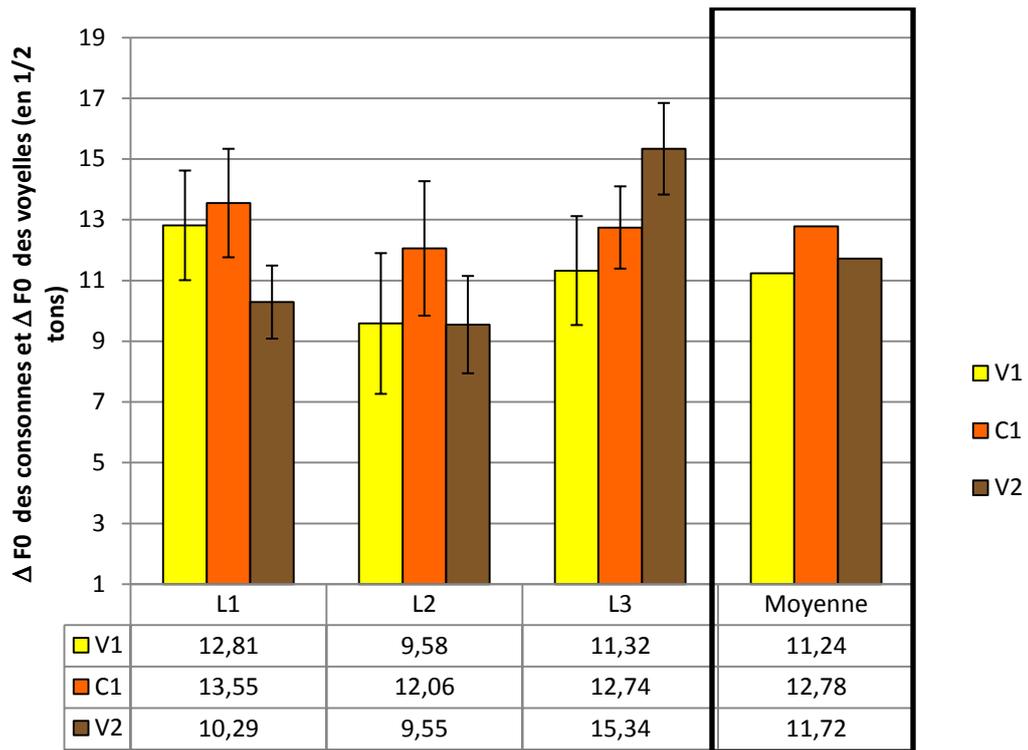


Figure 9-20 : Variations de F0 entre voix modale et voix criée pour les consonnes et voyelles des logatomes VCV

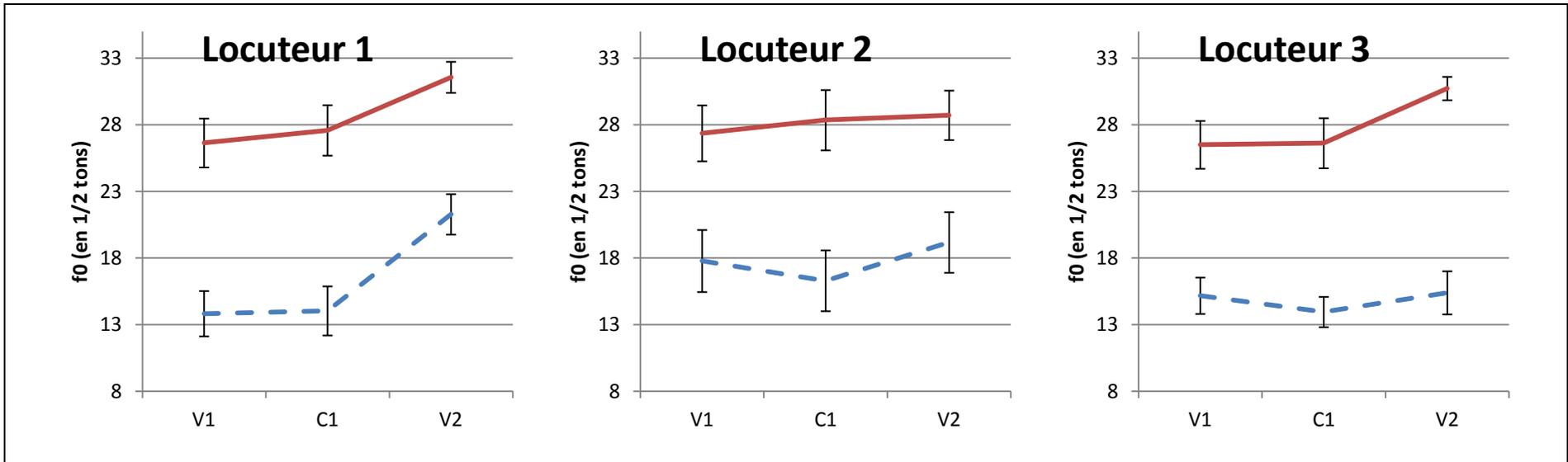


Figure 9-21 : Valeur moyenne de F0 pour chaque phonème et locuteurs des logatomes VCV. Les traits en pointillés correspondent à la voix parlée.

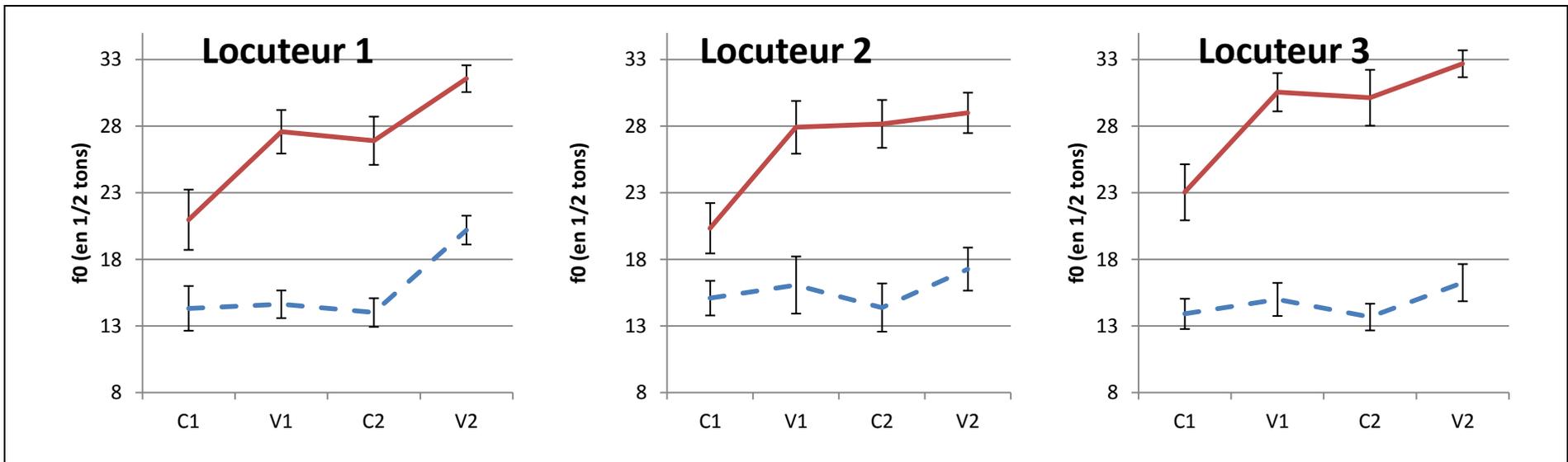


Figure 9-22 : Valeur moyenne de F0 pour chaque phonème et locuteurs des logatomes CVCV. Les traits en pointillés correspondent à la voix parlée.

### 9.3.4 Les variations relatives de F0 des consonnes et des voyelles pour les logatomes CVCV

Considérons enfin le cas des énoncés CVCV., La Figure 9-22, donne les valeurs moyennes de F0 pour la voix parlée et criée. Elle montre également des similarités avec les cas précédents.

Les voix parlées, sont assez semblables en ce qui concerne les 3 premiers phonèmes (C1, V1, C2) des 3 locuteurs. Toutefois, comme déjà observé sur les énoncés précédents, les voyelles finales sont différentes d'un locuteur à l'autre. Une forte augmentation est observée entre les premières voyelles (V1) et les deuxièmes (V2) pour le locuteur 1 (+5,6 demi-tons), tandis que les variations sont moins prononcées pour L2 (+1,2 demi-ton) et L3 (+1,3 demi-ton). On retrouve ici des intonations interrogatives en voix parlée et notamment pour L1.

Pour les voix criées, on relève une forte augmentation entre les consonnes initiales (C1) et les premières voyelles (V1) pour les trois locuteurs comme dans le cas des logatomes CV et CVC. (6,6 demi-tons, 7,6 demi-tons et 7,5 demi-tons pour L1, L2 et L3). Les secondes consonnes (C2), par rapport aux premières voyelles (V1) ne varient que peu. En revanche les voyelles finales (V2) par rapport aux premières voyelles (V1) augmentent pour L1 (4 demi-tons) et L3 (2,1 demi-tons) mais peu pour L2 (1,1 demi-ton).

Concernant les énoncés CVCV, on retrouve une faible augmentation de la F0 pour les consonnes en positions initiales de 7 demi-tons, similaire à celle observée pour les consonnes initiales sur les logatomes CV et CVC. La suite des énoncés CVCV est fortement augmentée avec une augmentation plus forte des consonnes intervocaliques (14,4 demi-tons) comparée aux voyelles (13,5 demi-tons pour V1 et 13,2 demi-tons pour V2). On constate de la même manière que pour les VCV, que L3 augmente plus les voyelles V2 que les autres locuteurs.

Tableau 9-15: Tableau récapitulatif des différences de F0 entre les phonèmes pour les logatomes CVCV (en demi-ton)

	L1		L2		L3	
	Parlée	Criée	Parlée	Criée	Parlée	Criée
$F0_{V1} - F0_{C1}$	0,3	6,6	1	7,6	1,1	7,5
$F0_{V1} - F0_{C2}$	0,6	0,7	1,7	-0,3	1,3	0,4
$F0_{V2} - F0_{C2}$	6,2	4,7	2,9	0,8	2,6	2,5
$F0_{V2} - F0_{V1}$	5,6	4	1,2	1,1	1,3	2,1
$F0_{C2} - F0_{C1}$	-0,3	5,9	-0,7	7,8	-0,2	7,1

Ainsi, on observe de fortes différences entre les premières consonnes (C1) et le reste de l'énoncé (comme pour les logatomes CV et CVC). On constate également que les consonnes intervocaliques sont augmentées aussi fortement que les consonnes intervocaliques des logatomes VCV. De plus cette augmentation est similaire (voir plus intense) que celle des voyelles adjacentes.

Tableau 9-16 : Tableau récapitulatif des différences de variations de F0 entre les phonèmes pour les logatomes CVCV (en demi-ton)

	L1	L2	L3
$\Delta F0_{V1}-\Delta F0_{C1}$	6,3	6,6	6,4
$\Delta F0_{V1}-\Delta F0_{C2}$	0,1	-1,9	-0,9
$\Delta F0_{V2}-\Delta F0_{C2}$	-1,6	-2,1	0
$\Delta F0_{V2}-\Delta F0_{V1}$	-1,6	-0,1	0,9
$\Delta F0_{C2}-\Delta F0_{C1}$	6,3	8,5	7,3

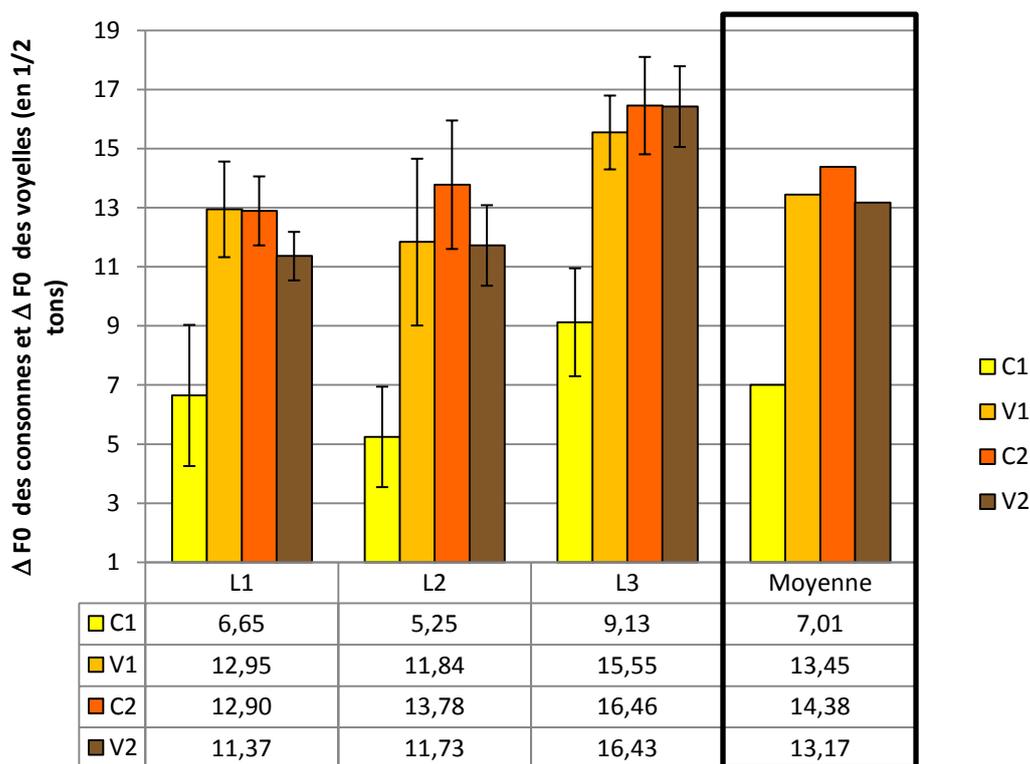


Figure 9-23 : Variations de F0 entre voix modale et voix criée pour les consonnes et voyelles des logatomes CVCV

### 9.3.5 Résumé

On constate que sur l'ensemble du corpus les locuteurs 1 et 2 ont une tendance, en voix parlée, à générer une structure prosodique similaire à celle d'une phrase interrogative. La fin des énoncés est quasi-systématiquement accentuée. Les variations relatives de F0 observées pour ces deux locuteurs, et notamment pour les phonèmes finals, sont biaisées. Le locuteur 3 ne présente toutefois pas cette caractéristique sur les voix parlées. Il apparaît ainsi que les résultats du locuteur 3 semblent plus proches de la réalité, du fait que l'intonation faite par ce dernier semble plus proche d'une intonation déclarative (Delattre, 1966). On constate ainsi, pour L3, de fortes variations de F0 entre la voix parlée et criée en fin d'énoncé par rapport à L1 et L2. Toutefois les analyses des valeurs moyennes de la F0

en voix criée nous ont permis de mettre à jour des stratégies communes aux 3 locuteurs. Des ressemblances quant à l'évolution de la F0 sont également observées en fonction de la position et du type de phonème considéré (consonne/voyelle). En résumé on constate que la F0 des consonnes en positions initiales ou en positions finales augmente moins fortement que pour les autres phonèmes des énoncés. La F0 des consonnes en positions intervocaliques, quant à elle, augmente plus que celle des voyelles. Un dernier constat est que la F0 des voyelles est moins augmentée sur les voyelles des énoncés VCV.

## 9.4 Micro-mélodie des énoncés CV, CVC, VCV et CVCV

---

Nous avons pu voir qu'il existe des différences entre les augmentations de la F0 en fonction du type de phonème (voyelles ou consonnes) mais également en fonction de la position de ces derniers. Dans cette section nous nous intéressons à la micro-mélodie de ces structures phonétiques. C'est-à-dire aux variations de F0 intrinsèques aux phonèmes. Nous nous intéressons dans cette section à la quantification des contours mélodiques des énoncés CV, CVC, VCV et CVCV en voix criée. Nous ne traitons pas ici de la micro-intonation (i.e. les variations intrinsèques de l'intensité). En effet, l'étude de ces phénomènes n'a rien apporté de plus que l'étude des valeurs moyennes de l'intensité. De plus la F0 semble être le paramètre qui contient le plus d'information pour la perception de la distance et de l'effort vocal (Brungart et al., 2002).

### 9.4.1 Les contours mélodiques des logatomes CV

La Figure 9-25 représente les contours de F0 superposés des logatomes CV en voix modale pour lesquels les consonnes sont voisées pour les trois locuteurs, et la Figure 9-26, les contours de ces mêmes mots en voix criée. Ces contours ont été normalisés en temps. A savoir que la F0 des consonnes a été ramenée entre 0 et 0,5 et la F0 des voyelles ramenée entre 0,5 et 1. Ainsi la limite consonnes-voyelles se situe systématiquement à 0,5 s. Comme nous l'avons déjà évoqué auparavant, on remarque pour les voix modales, que les locuteurs 1 et 2 augmentent considérablement la F0 en fin de phonation tandis que L3 génère une intonation plus neutre. Il est largement admis que le locuteur 3 présente une intonation naturelle alors que les locuteurs 1 et 2 ont utilisé des intonations de type interrogative (cf. (Delattre, 1966)). Toutefois, bien que les stratégies soient différentes en voix modale, les mêmes formes de F0 sont observées pour les trois locuteurs en voix criée.

En voix criée on observe donc un lobe sur les voyelles, correspondant à un focus, et une montée progressive de F0 sur la consonne. On remarque que pour les voix criées, la F0 des voyelles présente systématiquement un focus. Caractéristiques que l'on n'observe pas ou pas clairement en voix modale. On constate également que la F0 des consonnes est beaucoup plus basse que celle des voyelles en voix criée. En voix modale par contre, la F0 est relativement identique pour les consonnes et voyelles (pour le locuteur 3 tout du moins). Ainsi on remarque rapidement que dans le but de transformer les énoncés CV d'une voix modale en une voix criée, une simple multiplication, ou décalage des contours de F0, n'est pas suffisant pour obtenir le type de courbes observées.

La Figure 9-27 et la Figure 9-28 représentent les mêmes types de courbe mais pour les mots CV parlée et criée ayant des consonnes non-voisées. Contrairement à ce que l'on pourrait imaginer, la F0 des voyelles en voix criée ne débute pas au même niveau que les consonnes voisées, mais semble débiter à un niveau peu dépendant du caractère voisé ou non des consonnes qui précèdent. On retrouve donc les mêmes formes de F0 pour les mots CV avec C1 voisée ou non voisée.

Pour mieux d'écrire les évolutions de la F0 nous considérons 4 points caractéristiques de ces contours :

1. **F0<sub>C1</sub>-init** : la valeur de F0 en début de consonne,
2. **F0<sub>V1</sub>-init** : la valeur de F0 en début de voyelle,
3. **F0<sub>V1</sub>-max** : la valeur maximale de F0 et
4. **F0<sub>V1</sub>-finale** : la valeur de F0 en fin de voyelle.

Le Tableau 9-17 représente les valeurs moyennes de ces points caractéristiques pour l'ensemble des logatomes CV (en voix criée uniquement) et en fonction des classes de consonne. L'ensemble des valeurs classées par classe de phonème, par locuteur ou encore par voyelle est donné en Annexe E.

A partir des valeurs de ce tableau, on remarque toutefois certaines différences en ce qui concerne les valeurs caractéristiques de F0 des voyelles, par rapport au critère voisé ou non de la consonne qui précède. En effet, la valeur initiale de la voyelle (**F0<sub>V1</sub>-init**) est plus élevée quand elle est précédée d'une consonne occlusive ou fricative non voisé ou encore d'une nasale. On note également dans ce tableau que la valeur initiale des nasales (**F0<sub>C1</sub>-init**) est plus élevée que les autres consonnes voisées. Les valeurs maximales de la voyelle (**F0<sub>V1</sub>-max**) ainsi que les valeurs finales (**F0<sub>V1</sub>-finale**) sont identiques pour l'ensemble des logatomes CV.

Tableau 9-17: Valeurs caractéristiques moyennes des contours de F0 des logatomes CV pour les 3 locuteurs classés par famille de consonne. Les valeurs sont exprimées en demi-ton

	F0 <sub>c1</sub> -init	F0 <sub>v1</sub> -init	F0 <sub>v1</sub> -max	F0 <sub>v1</sub> -finale
<b>Fricatives voisées</b>	16,8	26,7	32,5	29,1
<b>Fricatives non-voisées</b>	/	28,7	32,3	29,0
<b>Occlusives voisées</b>	17,1	25,7	32,5	29,3
<b>Occlusives non-voisées</b>	/	28,7	32,4	29,0
<b>Nasales</b>	18,7	29,0	32,9	29,6
<b>Liquides</b>	17,1	26,2	33,0	28,9
<b>MOYENNE</b>	<b>17,4</b>	<b>27,5</b>	<b>32,6</b>	<b>29,2</b>

Afin de mieux décrire les contours de F0 des logatomes, nous considérons également la position du maximum de F0 sur la voyelle. Ainsi, la Figure 9-24 représente la position de ce maximum en pourcentage de la durée de la voyelle, pour les voix parlées et criées, ainsi que pour les trois locuteurs de notre corpus.

La position du maximum de la F0 change significativement en fonction du type de phonation. On constate naturellement que pour les locuteurs 1 et 2 ayant fait des intonations interrogatives, que le maximum se situe au alentour des 90% de la durée de la voyelle. En revanche, le maximum de F0 du locuteur 3 varie beaucoup entre la première partie et la dernière partie de la voyelle. En voix criée, on remarque toutefois une tendance à placer le maximum de la F0 aux alentours de la moitié de la durée de la voyelle. Les locuteurs 1 et 3 le place à 64 % alors que le locuteur 2 le place à 46%. Ainsi, le pic de F0 de la voyelle apparaît plus tôt en voix criée pour les locuteurs 1 et 2.

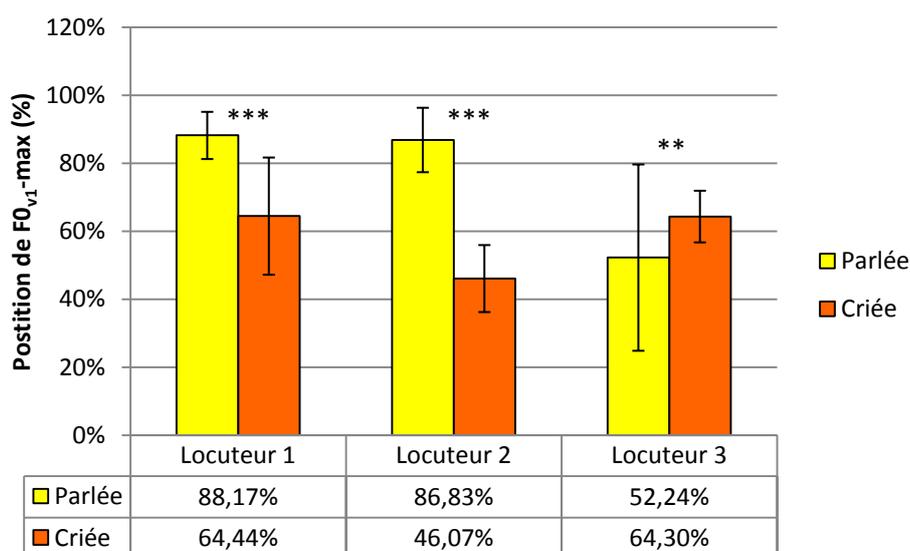


Figure 9-24: Position du maximum de F0 sur la voyelle des logatomes CV pour les 3 locuteurs (F0<sub>v1</sub>-max). La position est exprimée en % de la durée de la voyelle.

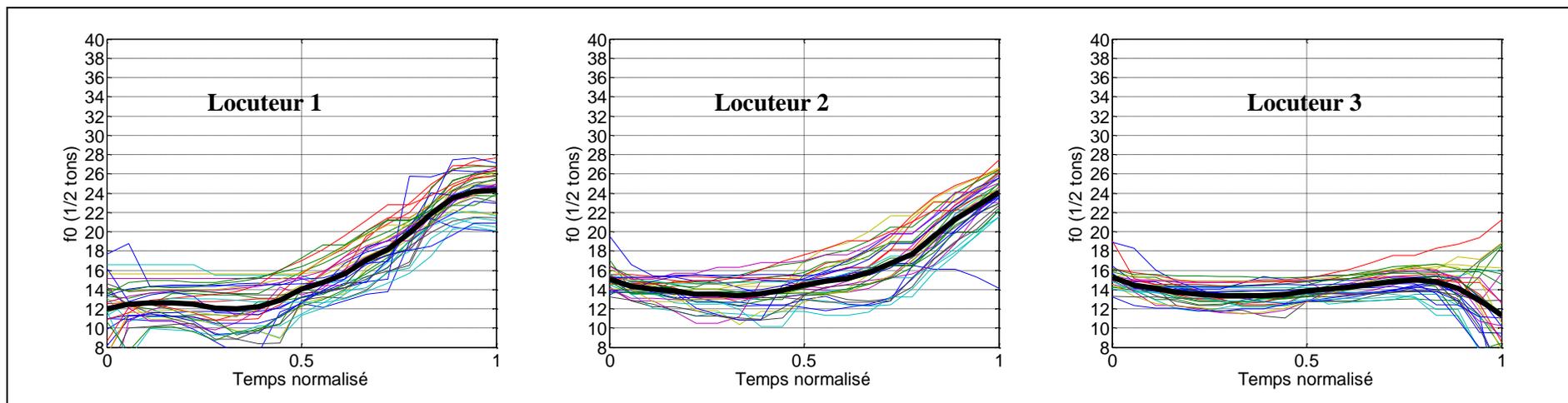


Figure 9-25 : Contours de F0 des logatomes CV en voix modale des 3 locuteurs dans le cas des consonnes voisées. Le trait gras représente la moyenne des courbes.

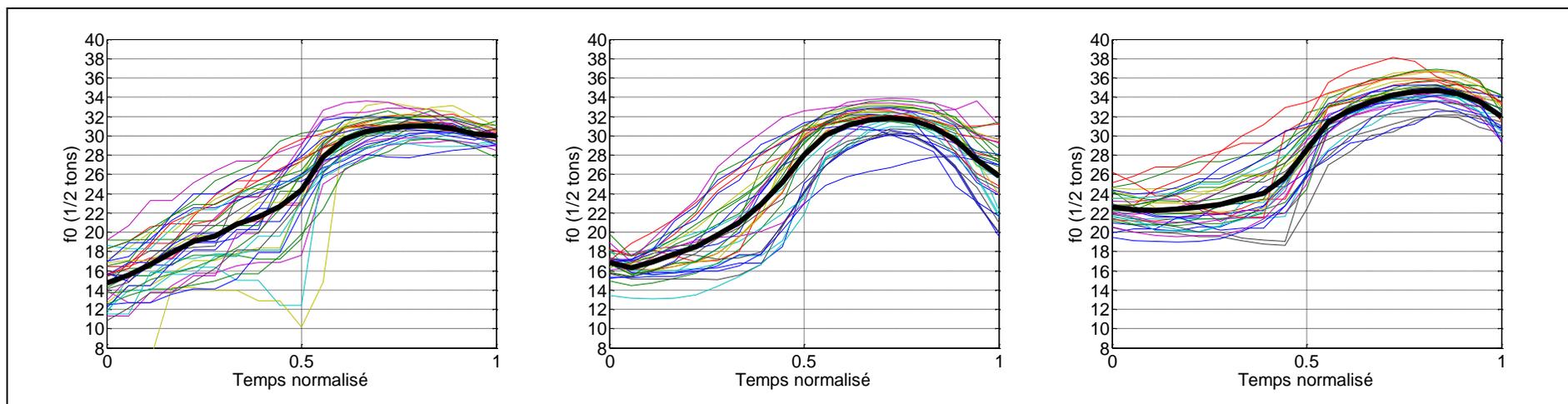


Figure 9-26 : Contour de F0 des logatomes CV en voix crée des 3 locuteurs dans le cas des consonnes voisées. Le trait gras représente la moyenne des courbes.

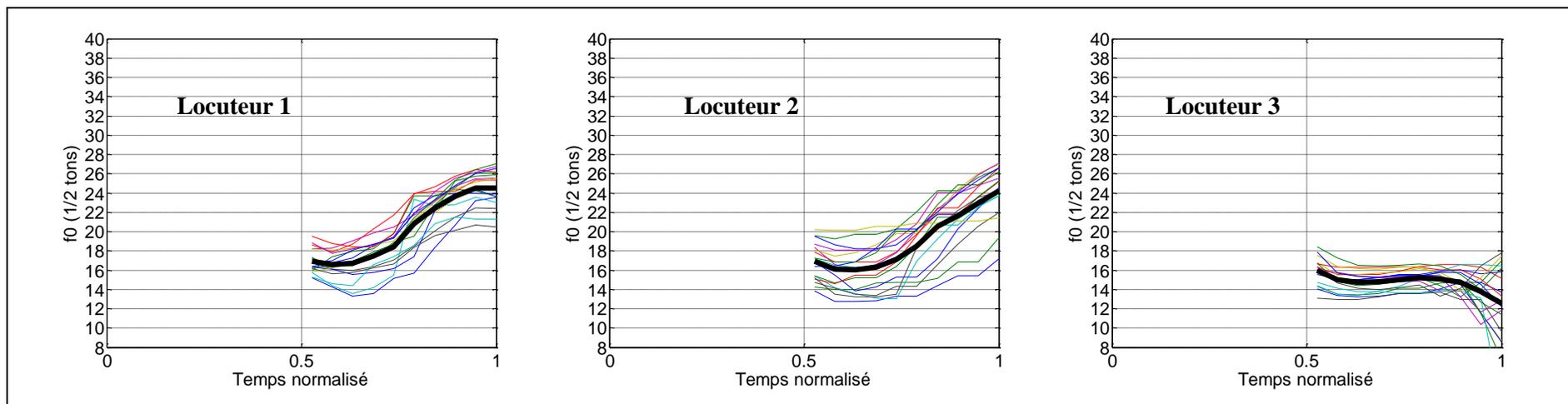


Figure 9-27: Contour de F0 des logatomes CV en voix modale des 3 locuteurs dans le cas des consonnes non voisées. Le trait gras représente la moyenne des courbes.

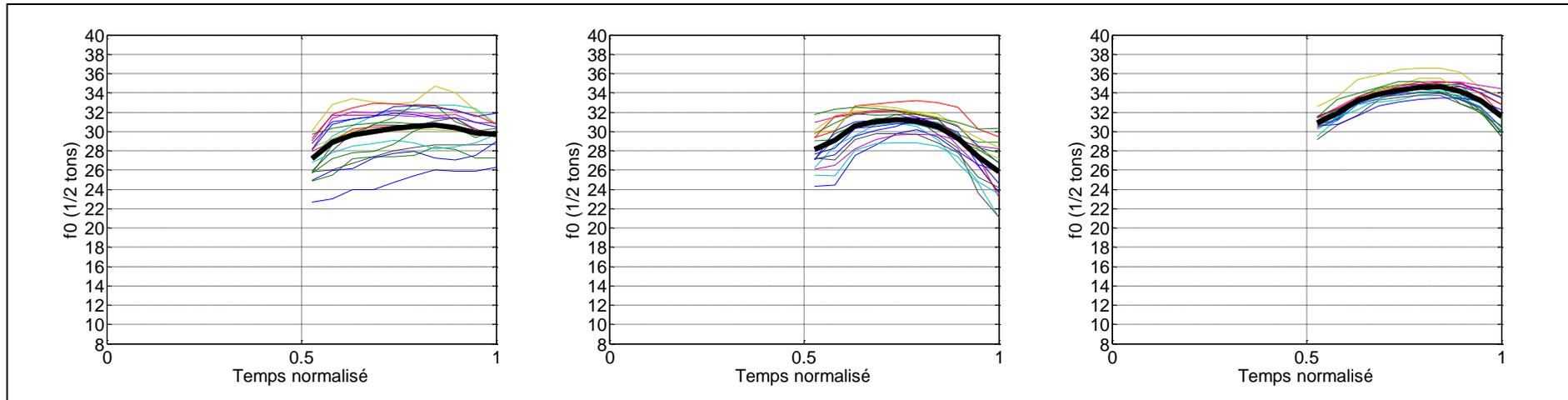


Figure 9-28: Contour de F0 des logatomes CV en voix criée des 3 locuteurs dans le cas des consonnes non voisées. Le trait gras représente la moyenne des courbes.

### 9.4.2 Les contours mélodiques des logatomes CVC

Sur Figure 9-30 sont superposés les contours de F0 de la voix parlée pour le cas des consonnes voisées. La Figure 9-31 représente quant à elle, les contours de ces mêmes logatomes en voix criée. Afin de permettre ces superpositions ces courbes sont normalisées en temps. C'est-à-dire que le contour de la première consonne est ramené entre 0 et 1/3, le contour de la voyelle est ramené entre 1/3 et 2/3 et le contour de la consonne finale est ramené entre 2/3 et 1. Sur ces courbes, la voyelle est donc représentée entre 1/3 et 2/3. La Figure 9-32 et la Figure 9-33 reprennent ce tracé des contours mais pour les logatomes CVC ne contenant que des consonnes non-voisées.

De la même manière que pour les logatomes CV, on observe pour les voix parlées, peu de variations sur la F0 si ce n'est une forte augmentation finale pour L1 et L2. Les voix criées présentent quant à elles des lobes fortement proéminant sur les voyelles pour l'ensemble des logatomes. Les contours de F0 des voyelles, dans le cas des logatomes contenant des consonnes non voisées, ne semblent pas être affectés par ces dernières. En effet, les contours de la voyelle sont similaires dans le cas où les consonnes sont voisées ou non. Concernant les contours de F0 des consonnes, ils débutent et se terminent à des valeurs bien inférieures à celles de la voyelle. On note d'ailleurs une symétrie centrale sur les contours des F0.

Pour mieux d'écrire les évolutions de la F0 nous considérons ici 5 points caractéristiques de ces contours :

1. **F0<sub>c1-init</sub>** : la valeur de F0 en début de consonne C1,
2. **F0<sub>v1-init</sub>** : la valeur de F0 en début de voyelle V1,
3. **F0<sub>v1-max</sub>** : la valeur maximale de F0 de la voyelle V1,
4. **F0<sub>v1-finale</sub>** : la valeur de F0 en fin de voyelle V1.
5. **F0<sub>c2-finale</sub>** : la valeur finale de la consonne C2

Le Tableau 9-18, reprend ces 5 points caractéristiques et liste les valeurs moyennes pour l'ensemble des logatomes CVC. Les tableaux détaillés sont donnés en Annexe E.

A nouveau, on constate une valeur initiale des consonnes nasales (**F0<sub>c1-init</sub>**) plus forte que les autres consonnes. Les valeurs initiales (**F0<sub>v1-init</sub>**) des voyelles sont plus fortes pour l'ensemble des consonnes non voisées. Les valeurs maximales des voyelles (**F0<sub>v1-max</sub>**) sont stables pour l'ensemble des consonnes. Enfin, les valeurs finales des voyelles (**F0<sub>v1-finale</sub>**) sont plus élevées pour les nasales et les liquides. Comme déjà mentionné, la F0 montre une symétrie autour de la voyelle. Ainsi les valeurs finales des consonnes (**F0<sub>c2-finale</sub>**) sont proches des valeurs initiales (à 1 demi-ton près) sauf pour les valeurs finales des consonnes occlusives voisées.

Tableau 9-18 : Valeurs caractéristiques moyennes des contours de F0 des logatomes CVC pour les 3 locuteurs classés par famille de consonne. Les valeurs sont exprimées en demi-ton

	FO <sub>c1</sub> -init	FO <sub>v1</sub> -init	FO <sub>v1</sub> -max	FO <sub>v1</sub> -finale	FO <sub>c2</sub> -finale
<b>Fricatives voisées</b>	17,5	25,4	32,3	27,1	17,0
<b>Fricatives non-voisées</b>	/	28,8	32,8	27,8	/
<b>Occlusives voisées</b>	17,5	24,6	32,2	28,2	20,3
<b>Occlusives non-voisées</b>	/	28,3	32,4	27,1	/
<b>Nasales</b>	19,1	27,6	32,9	31,2	18,8
<b>Liquides</b>	17,1	27,1	32,9	29,0	18,7
<b>MOYENNE</b>	<b>17,8</b>	<b>27,0</b>	<b>32,6</b>	<b>28,4</b>	<b>18,7</b>

Considérons à présent la position du maximum de F0 sur la voyelle. Sur Figure 9-29 on constate à nouveau, pour les voix parlées, que les locuteurs L1 et L2 possèdent un maximum aux alentours des 90% de la durée de la voyelle dû à leurs intonations et que le locuteur 3, quant à lui, présente des maxima autour de 45 %. En voix criée, ces valeurs tendent à s'homogénéiser. En effet pour les trois locuteurs la position du maximum de F0 de la voyelle se situe entre 60 % et 70 %. Ainsi, le pic de F0 de la voyelle apparaît plus tôt en voix criée pour les locuteurs 1 et 2.

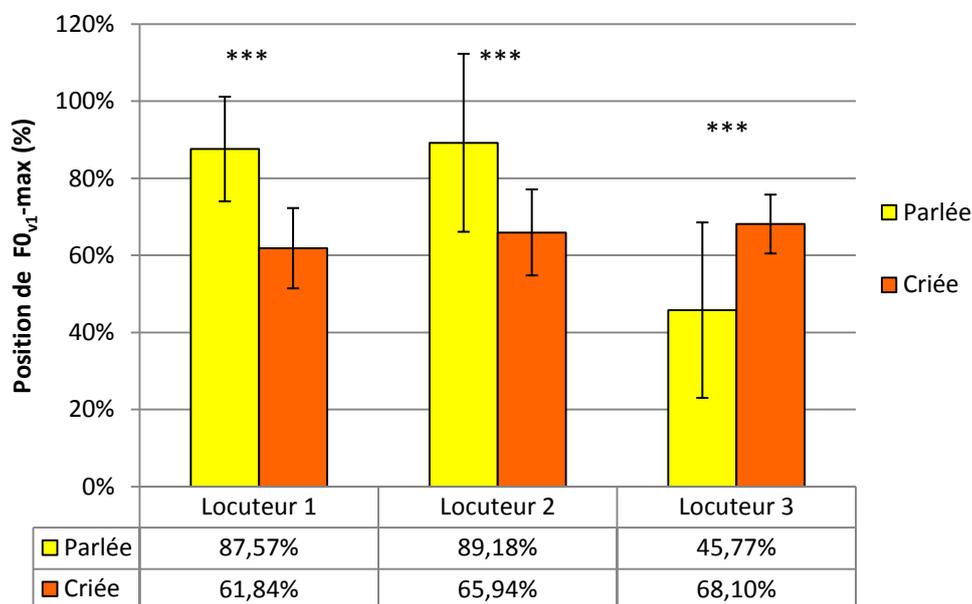


Figure 9-29 : Position du maximum de F0 sur la voyelle des logatomes CVC pour les 3 locuteurs (FO<sub>v1</sub>-max). La position est exprimée en % de la durée de la voyelle.

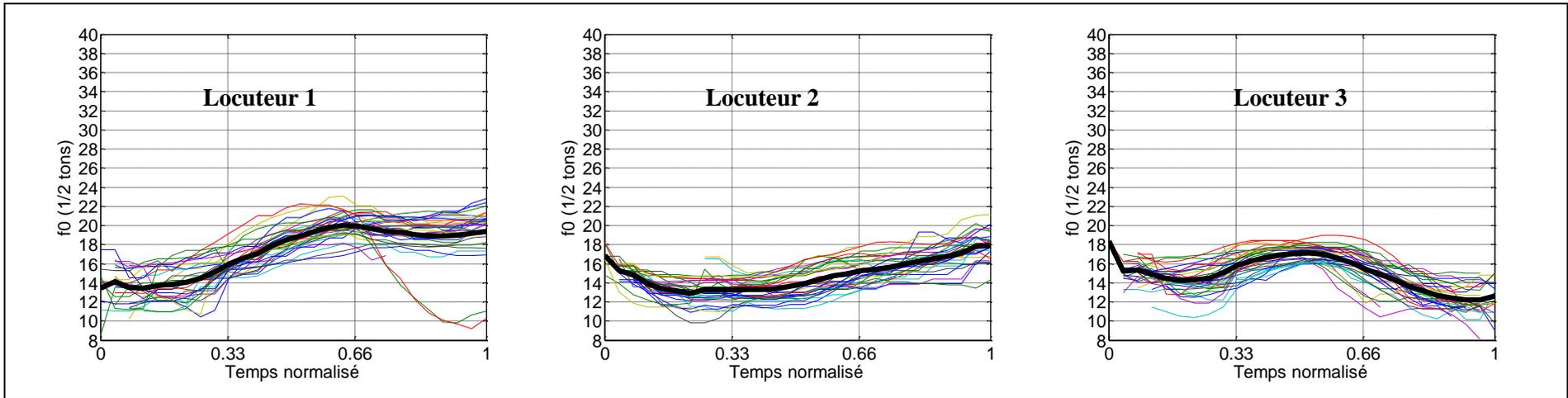


Figure 9-30 : Contour de F0 des logatomes CVC en voix modale des 3 locuteurs dans le cas des consonnes voisées. Le trait gras représente la moyenne des courbes.

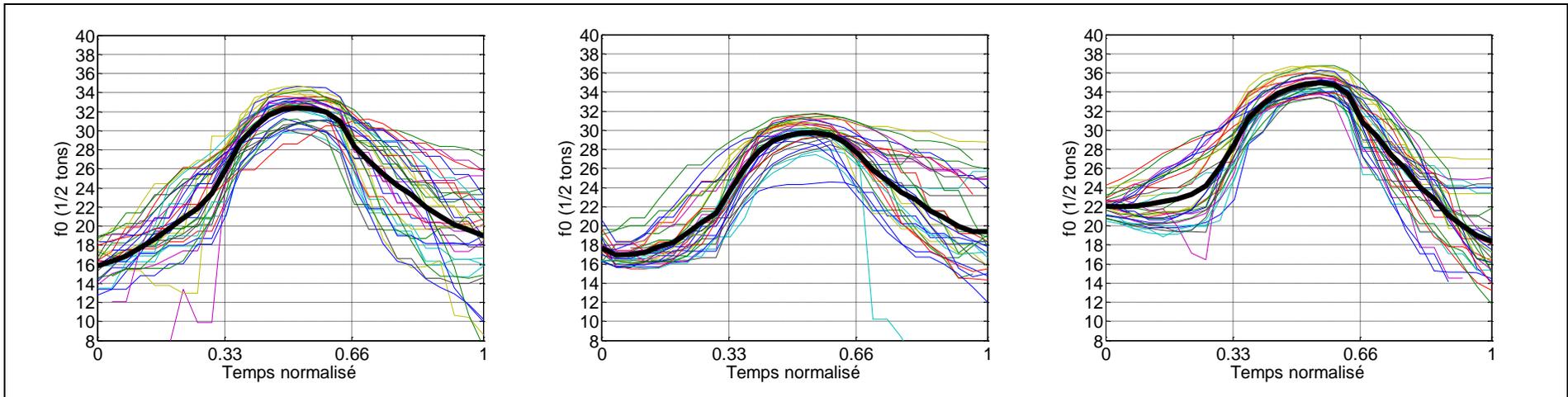


Figure 9-31 : Contour de F0 des logatomes CVC en voix criée des 3 locuteurs dans le cas des consonnes voisées. Le trait gras représente la moyenne des courbes.

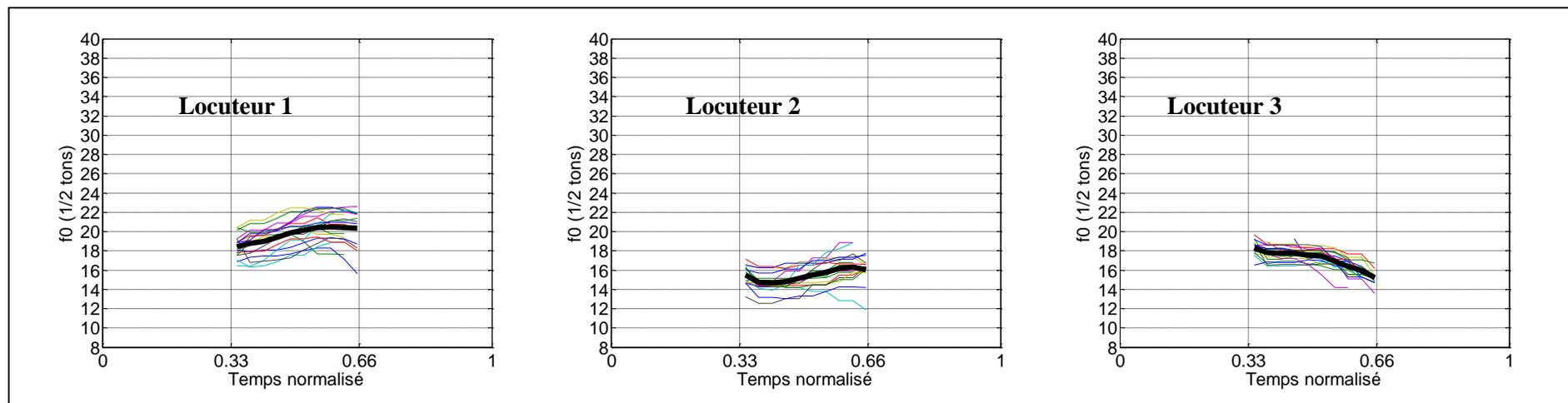


Figure 9-32: Contour de F0 des logatomes CV en voix modale des 3 locuteurs dans le cas des consonnes non voisées. Le trait gras représente la moyenne des courbes.

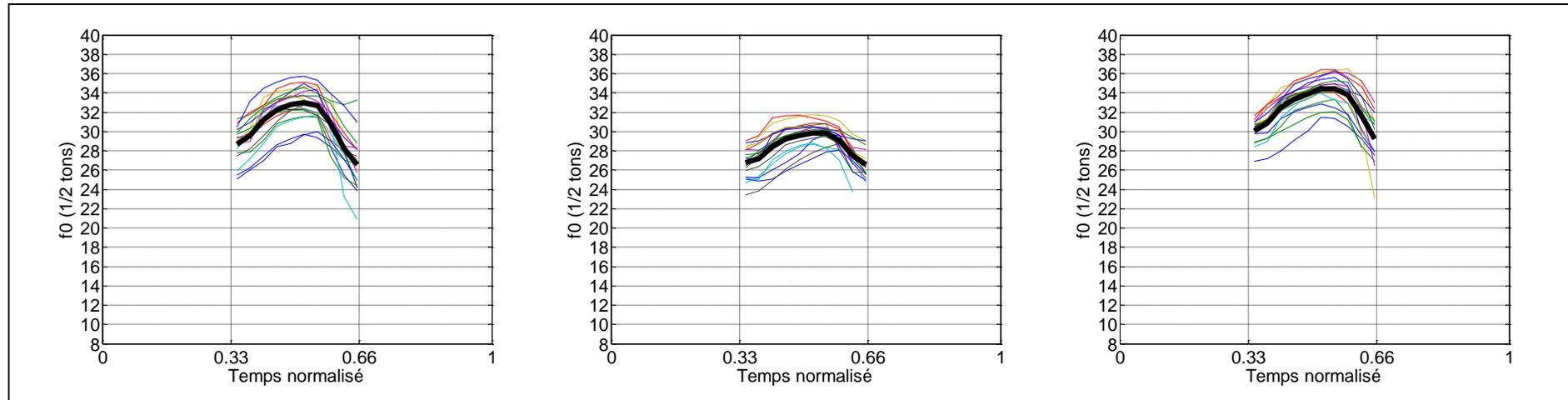


Figure 9-33: Contour de F0 des logatomes CV en voix criée des 3 locuteurs dans le cas des consonnes non voisées. Le trait gras représente la moyenne des courbes.

### 9.4.3 Les contours mélodiques des logatomes VCV

De la même manière que pour les deux cas précédents on retrouve sur la Figure 9-36 et la Figure 9-37 les contours normalisés en temps pour les logatomes contenant des consonnes voisées. Dans ce cas, la voyelle se situe entre 0 et 1/3 et la deuxième entre 2/3 et 1. La Figure 9-38 et la Figure 9-39 sont les équivalents des courbes précédentes pour les logatomes contenant des consonnes non-voisées.

Le cas des logatomes VCV est moins évident. En effet, on constate à nouveau le trait intonatif des locuteurs 1 et 2 en voix parlée tandis que L3 effectue une intonation neutre. Étrangement, on constate que L1 possède également des lobes sur les voyelles en voix parlée. Toutefois nous ne nous expliquons pas ce phénomène. Les voix criées présentent, quant à elle, également des lobes sur chacune des voyelles mais ces derniers sont moins marqués que dans les cas précédemment étudiés (notamment pour L2). On constate notamment sur ces courbes que la F0 de la consonne intervocalique est très diffuse allant de fort creux jusqu'à une transition plate entre les deux voyelles adjacentes.

Pour mieux d'écrire les évolutions de la F0 nous considérons ici 7 points caractéristiques de ces contours :

1. **F0<sub>V1-init</sub>** : la valeur de F0 en début de voyelle V1,
2. **F0<sub>V1-max</sub>** : la valeur maximale de F0 de la voyelle V1,
3. **F0<sub>V1-finale</sub>** : la valeur de F0 en fin de voyelle V1.
4. **F0<sub>C1-min</sub>**: la valeur minimale de la consonne C1
5. **F0<sub>V2-init</sub>** : la valeur de F0 en début de voyelle V2,
6. **F0<sub>V2-max</sub>** : la valeur maximale de F0 de la voyelle V2,
7. **F0<sub>V2-finale</sub>** : la valeur de F0 en fin de voyelle V2.

Les valeurs initiales des premières voyelles (**F0<sub>V1-init</sub>**), les valeurs maximales des premières voyelles (**F0<sub>V1-max</sub>**) ainsi que les trois descripteurs des deuxièmes voyelles (**F0<sub>V2-init</sub>**, **F0<sub>V2-max</sub>** et **F0<sub>V2-finale</sub>**) sont relativement identiques pour les différentes classes de consonnes (cf. Tableau 9-19). Concernant les valeurs finales des premières voyelles (**F0<sub>V1-finale</sub>**) celles-ci sont plus hautes pour toutes les consonnes voisées. D'autre part, les valeurs minimales de la F0 des consonnes intervocaliques (**F0<sub>C1-min</sub>**) expliquent le fait que les creux sur les courbes de moyennes ne soient pas aussi proéminent et homogènes que ce que nous pensions. En effet, les valeurs de **F0<sub>C1-min</sub>** pour les nasales et les liquides sont quasiment aussi hautes que les valeurs extrêmes des voyelles adjacentes. Seuls des creux sont observés pour les occlusives et les fricatives voisées. Ainsi, il semble que les consonnes intervocaliques influent beaucoup sur la forme du contour globale de F0 des logatomes

VCV. D'autre part, fait très étonnant, les valeurs caractéristiques de la F0 des secondes voyelles sont plus fortes que celle des premières voyelles. Ce qui se traduit par une déclinaison positive de la pente de F0 qui traditionnellement décrit une intonation interrogative ou encore une intonation dite d'implication (Delattre, 1966).

*Tableau 9-19 : Valeurs caractéristiques moyennes des contours de F0 des logatomes VCV pour les 3 locuteurs classées par famille de consonne. Les valeurs sont exprimées en demi-ton*

	<b>F0<sub>V1</sub>-init</b>	<b>F0<sub>V1</sub>-max</b>	<b>F0<sub>V1</sub>-finale</b>	<b>F0<sub>C1</sub>-min</b>	<b>F0<sub>V2</sub>-init</b>	<b>F0<sub>V2</sub>-max</b>	<b>F0<sub>V2</sub>-finale</b>
<b>Fricatives voisées</b>	22,7	28,5	27,1	24,7	28,3	31,3	28,9
<b>Fricatives non-voisées</b>	23,5	28,4	26,3	/	29,6	31,2	29,2
<b>Occlusives voisées</b>	23,8	29,1	27,5	25,5	28,2	31,4	28,6
<b>Occlusives non-voisées</b>	23,6	28,2	24,8	/	29,6	31,1	28,6
<b>Nasales</b>	23,7	29,1	28,7	28,4	29,9	31,9	29,5
<b>Liquides</b>	24,3	28,8	28,0	27,0	29,3	31,7	29,6
<b>MOYENNE</b>	<b>23,6</b>	<b>28,7</b>	<b>27,0</b>	<b>26,4</b>	<b>29,2</b>	<b>31,4</b>	<b>29,1</b>

Considérons à présent la position du maximum de F0 sur les deux voyelles. Sur la Figure 9-34, nous considérons le cas des premières voyelles des logatomes VCV. Pour les voix parlées, on constate clairement le fait que le locuteur 1 est fait une sorte de focus en voix parlée. Pour les deux autres locuteurs, les valeurs sont assez banales et reflètent une déclinaison de F0 au cours de l'énoncé. Toutefois, dans le cas des voix criées, la position du maximum de F0 tend à être similaire pour les trois locuteurs ; à savoir une valeur aux alentours de 78 %. Ainsi, le pic de F0 de la première voyelle apparaît plus tard en voix criée à l'inverse des logatomes CV et CVC.

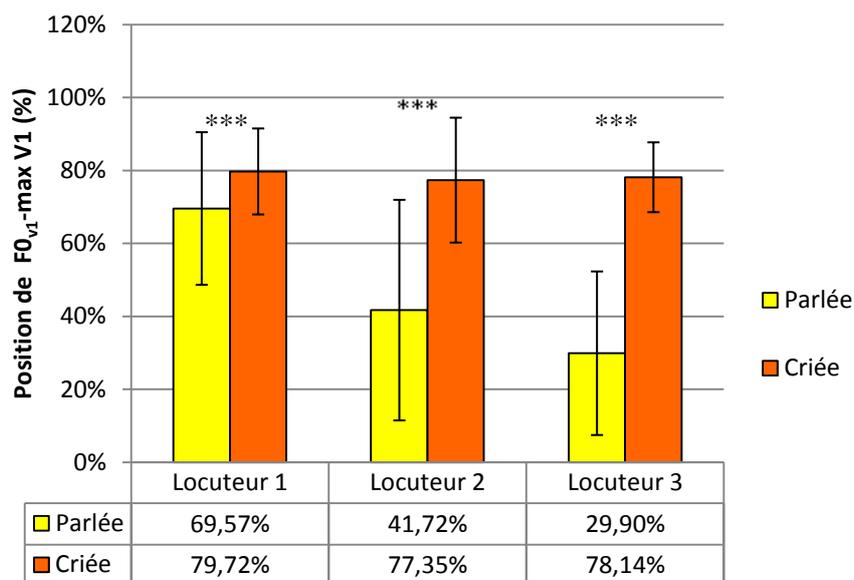


Figure 9-34: Position du maximum de F0 sur la première voyelle des logatomes VCV pour les 3 locuteurs ( $F0_{v1-max}$ ). La position est exprimée en % de la durée de la voyelle.

La Figure 9-35, représente la position du maximum de F0 pour les deuxièmes voyelles. Dans ce cas, la position du maximum aussi bien en voix parlée qu'en voix criée est confuse ; notamment pour les voix criées du locuteur 2. En effet, sur les courbes précédentes il est bien montré que ce dernier n'a pas réalisé de focus, ou tout du moins, des focus moins poussés que les deux autres locuteurs. En revanche, les locuteurs L1 et L3 montrent des maxima, de F0 pour les secondes voyelles aux alentours de 60%. Ainsi, le pic de F0 de la deuxième voyelle apparaît plus tôt en voix criée pour les locuteurs 1 et 2 comme pour les logatomes CV et CVC.

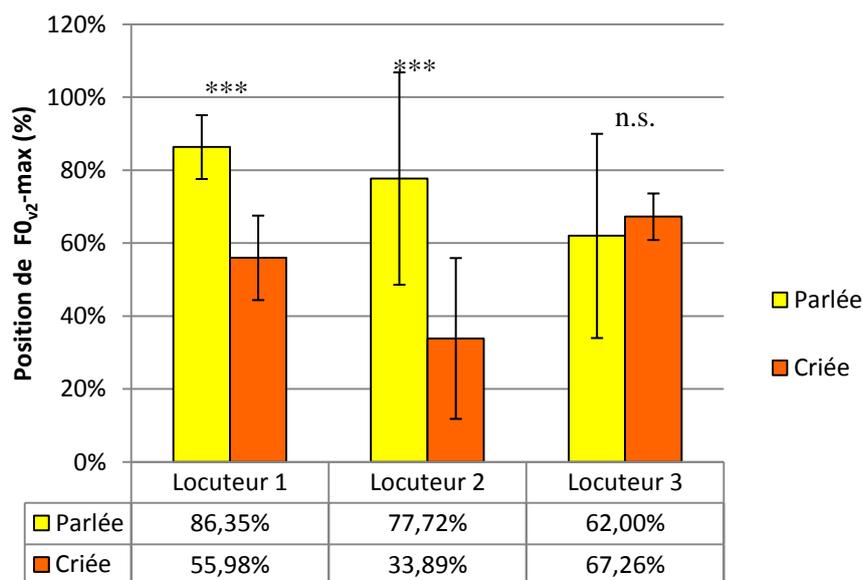


Figure 9-35: Position du maximum de F0 sur la deuxième voyelle des logatomes VCV pour les 3 locuteurs ( $F0_{v2-max}$ ). La position est exprimée en % de la durée de la voyelle.

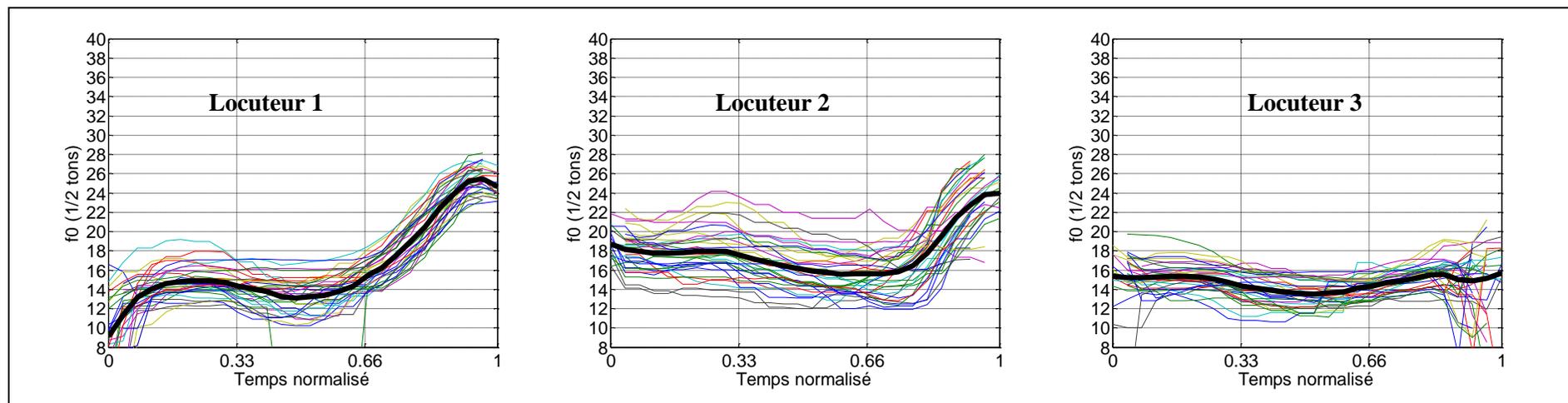


Figure 9-36 : Contour de F0 des logotomes VCV en voix modale des 3 locuteurs dans le cas des consonnes voisées. Le trait gras représente la moyenne des courbes.

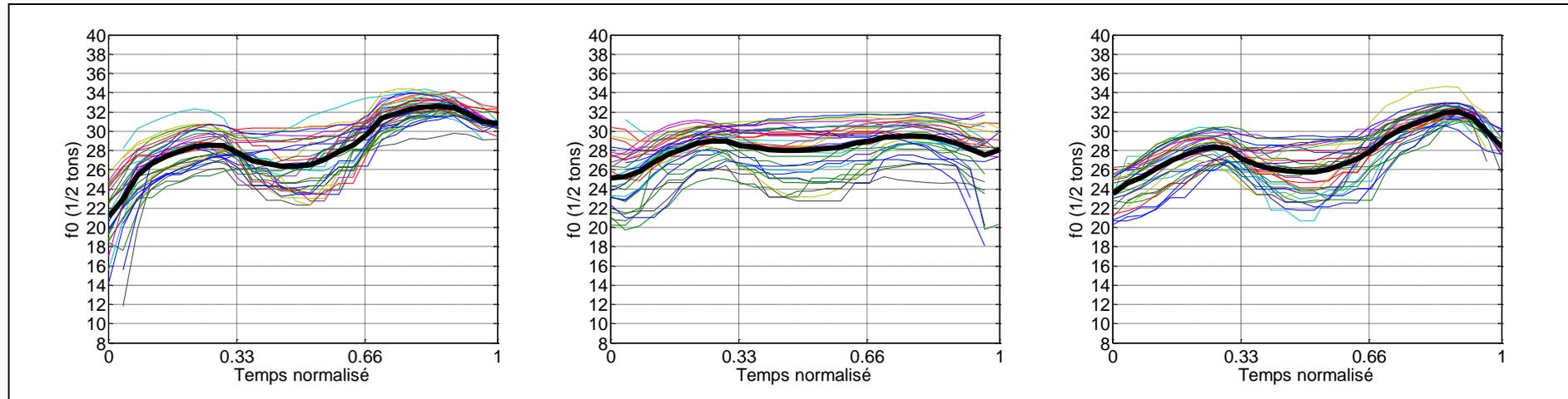


Figure 9-37 : Contour de F0 des logotomes VCV en voix criée des 3 locuteurs dans le cas des consonnes voisées. Le trait gras représente la moyenne des courbes.

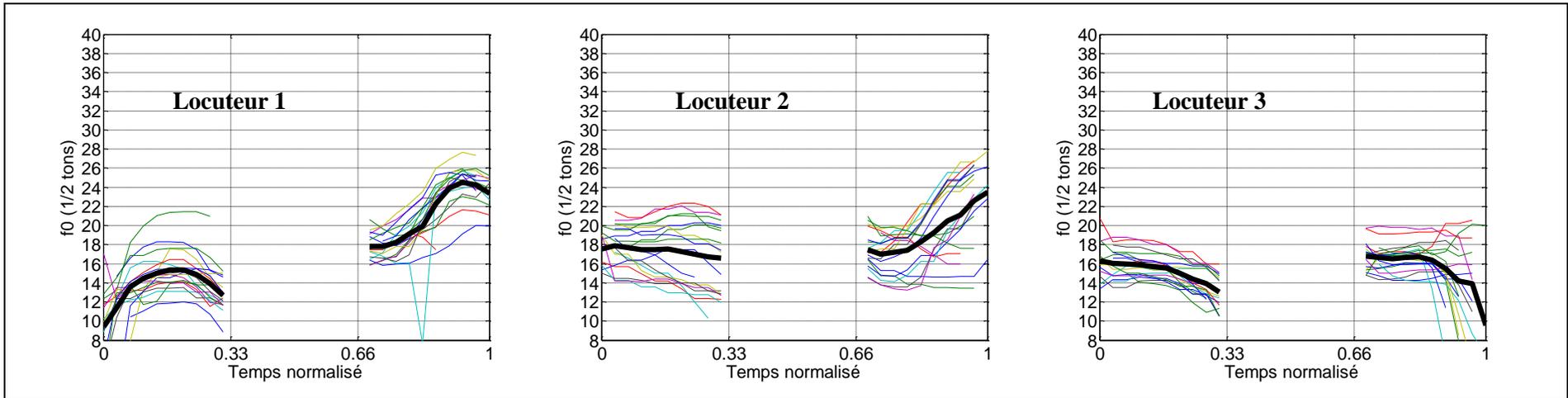


Figure 9-38: Contour de F0 des logatomes VCV en voix modale des 3 locuteurs dans le cas des consonnes non voisées. Le trait gras représente la moyenne des courbes.

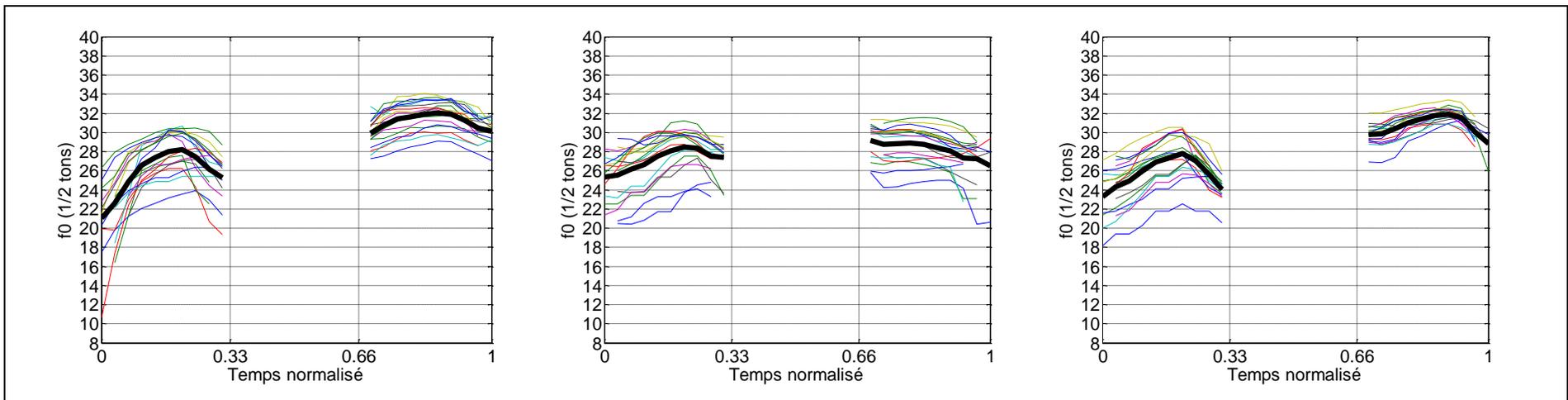


Figure 9-39: Contour de F0 des logatomes VCV en voix crée des 3 locuteurs dans le cas des consonnes non voisées. Le trait gras représente la moyenne des courbes.

#### 9.4.4 Les contours mélodiques des logatomes CVCV

La Figure 9-42 et la Figure 9-43 représentent la superposition des contours de F0 pour les voix parlées et criées des logatomes CVCV contenant des consonnes voisées et la Figure 9-44 et la Figure 9-45 les contours des logatomes CVCV pour des consonnes non voisées. Ici, le temps est normalisé pour que la première consonne C1 soit contenue entre 0 et 1/4, la première voyelle V1 entre 1/4 et 2/4, la deuxième consonne entre 2/4 et 3/4 et enfin la seconde voyelle entre 3/4 et 1. Comme précédemment pour les voix parlées on observe des légers lobes sur les voyelles mais uniquement pour le cas de logatomes avec des consonnes voisées. Pour le cas des CVCV avec consonnes non voisées, on n'observe des lobes uniquement sur la seconde voyelle. Toutefois ces derniers sont assez faibles. On constate à nouveau en voix parlée, des augmentations significatives de F0 en fin d'énoncé pour L1 et L2 mais également, dans des proportions plus faible, pour L3.

Les voix criées en revanche présentent de fort focus sur les voyelles comme dans tous les cas précédents. On constate notamment que les consonnes C1 présentent les mêmes évolutions que celles observées pour les consonnes initiales des logatomes CV et CVC et que les évolutions des deuxièmes consonnes semblent être similaires à celles des logatomes VCV.

Pour mieux d'écrire les évolutions de la F0 nous considérons ici 7 points caractéristiques de ces contours :

1. **F0<sub>C1-init</sub>**: la valeur initiale de la consonne C1
2. **F0<sub>V1-init</sub>** : la valeur de F0 en début de voyelle V1,
3. **F0<sub>V1-max</sub>** : la valeur maximale de F0 de la voyelle V1,
4. **F0<sub>V1-finale</sub>** : la valeur de F0 en fin de voyelle V1.
5. **F0<sub>C2-min</sub>**: la valeur minimale de la consonne C2
6. **F0<sub>V2-init</sub>** : la valeur de F0 en début de voyelle V2,
7. **F0<sub>V2-max</sub>** : la valeur maximale de F0 de la voyelle V2,
8. **F0<sub>V2-finale</sub>** : la valeur de F0 en fin de voyelle V2.

Tableau 9-20 : Valeurs caractéristiques moyennes des contours de F0 des logatomes CVCV pour les 3 locuteurs classés par famille de consonne. Les valeurs sont exprimées en demi-ton

	FO <sub>c1</sub> -init	FO <sub>v1</sub> -init	FO <sub>v1</sub> -max	FO <sub>v1</sub> -finale	FO <sub>c2</sub> -min	FO <sub>v2</sub> -init	FO <sub>v2</sub> -max	FO <sub>v2</sub> -finale
<b>Fricatives voisées</b>	17,8	25,2	30,0	27,7	26,1	29,1	32,3	29,4
<b>Fricatives non-voisées</b>	/	28,2	30,7	27,2	/	30,4	32,4	29,4
<b>Occlusives voisées</b>	18,3	23,9	29,9	28,0	26,5	28,8	32,3	29,4
<b>Occlusives non-voisées</b>	/	27,3	29,9	26,7	/	30,1	32,2	29,3
<b>Nasales</b>	20,1	26,7	30,2	29,9	29,6	30,6	32,4	29,8
<b>Liquides</b>	19,0	25,4	30,0	28,8	28,1	29,5	32,5	29,6
<b>MOYENNE</b>	18,8	26,1	30,1	28,0	27,6	29,7	32,3	29,5

Les valeurs du Tableau 9-20 ne sont pas sans rappeler les différentes observations faites sur les logatomes précédents. En effet, les valeurs initiales des consonnes C1 ( $F0_{C1-init}$ ) sont plus fortes pour les consonnes nasales et liquides. Les valeurs minimales des consonnes intervocaliques C2 ( $F0_{C2-min}$ ) sont plus fortes pour les liquides et les nasales. Les valeurs maximales des voyelles V1 ( $F0_{V1-max}$ ) et V2 ( $F0_{V2-max}$ ) sont relativement constantes mais sont plus élevées pour V2 que pour V1. Les valeurs initiales de V1 ( $F0_{V1-init}$ ) sont dépendantes du caractère voisé de la consonne précédente et sont plus fortes quand la consonne est non voisée. Les valeurs finales des secondes voyelles V2 ( $F0_{V2-finale}$ ) sont similaires pour toutes les classes de consonnes. Les valeurs finales de V1 ( $F0_{V1-finale}$ ) sont moins fortes dans le cas des consonnes non-voisées et les valeurs initiales de V2 ( $F0_{V2-init}$ ) sont quant à elles relativement constantes (seul point différent du cas précédent : VCV).

Qu'en est-il des positions des maxima de la F0 ? La Figure 9-40 montre des similitudes par rapport aux logatomes CVC, à savoir que la position du maximum en voix criée s'homogénéise pour les trois locuteurs autour d'une valeur d'environ 70 %. On constate également, dans ce cas, que la position du maximum en voix parlée est très similaire pour les trois locuteurs. En effet, nous pensons que la position de ces voyelles n'est pas sujet aux différences d'intonation faites par les locuteurs L1 et L2. En revanche le maximum des voyelles finales est quant à lui dépendant de l'intonation. Ainsi, le pic de F0 de la première voyelle apparaît plus tard en voix criée comme pour les logatomes VCV.

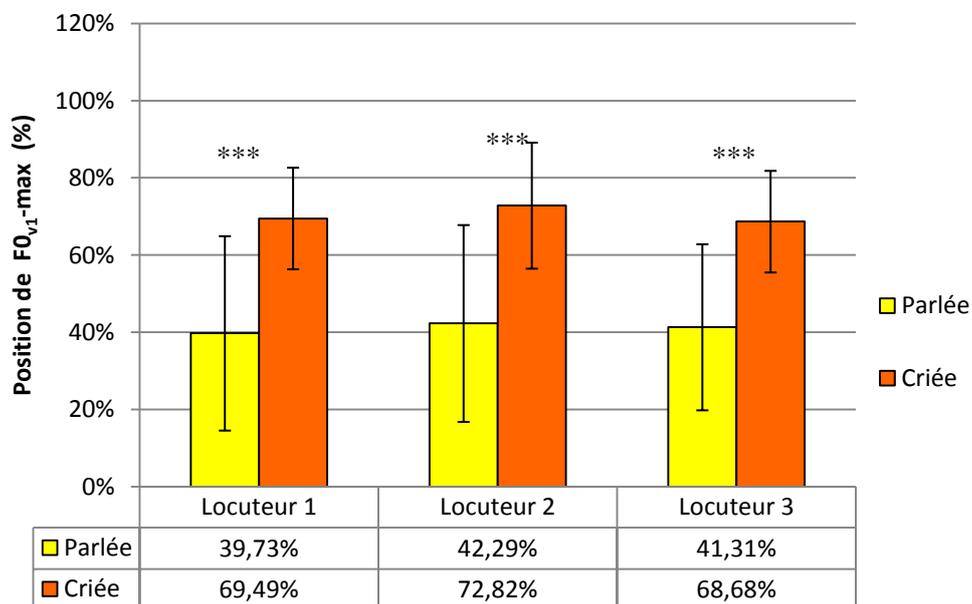


Figure 9-40 : Position du maximum de F0 sur la première voyelle des logatomes VCV pour les 3 locuteurs ( $F0_{V1-max}$ ). La position est exprimée en % de la durée de la voyelle.

En effet, la Figure 9-41 montre des différences, aussi bien en voix parlée qu'en voix criée, pour les trois locuteurs. Pour les 3 locuteurs les positions des maxima en voix parlée sont proches de la fin de la voyelle. En voix criée, on retrouve ici le fait que L2 ne réalise pas autant de focus que les deux autres locuteurs. Néanmoins, les valeurs de L1 et L3 sont ici similaires et environ égale à 63 %. Ainsi,

le pic de F0 de la deuxième voyelle apparaît plus tôt en voix criée comme pour les logatomes CV, VCV, et CVCV.

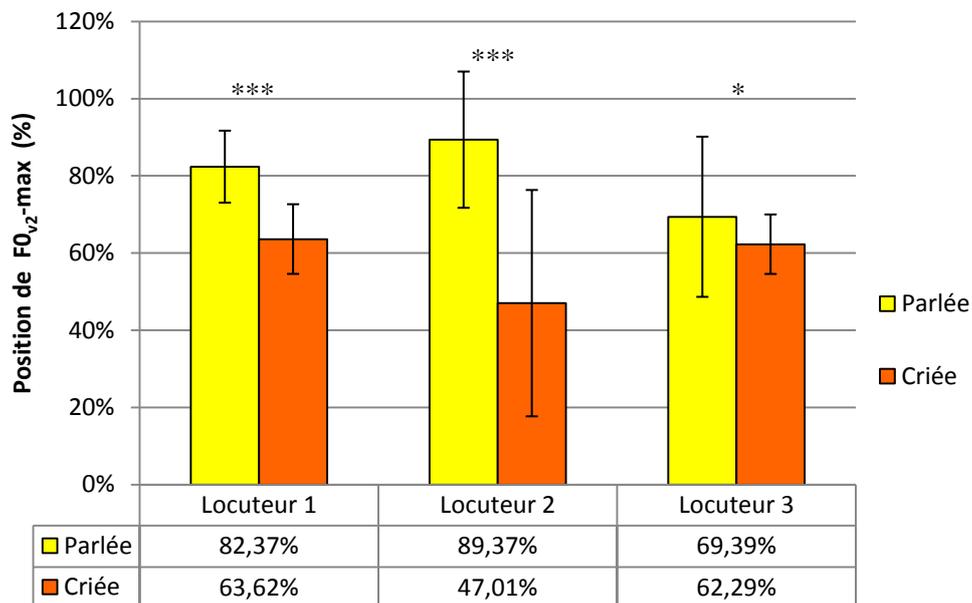


Figure 9-41: Position du maximum de F0 sur la deuxième voyelle des logatomes VCV pour les 3 locuteurs ( $F0_{v2-max}$ ). La position est exprimée en % de la durée de la voyelle

#### 9.4.5 Résumé

En résumé, on constate que sur l'ensemble des logatomes en voix criée du corpus DB4 un focus est réalisé sur les voyelles. D'autre part, les contours de F0 sont similaires entre les 3 locuteurs pour un type de logatome donné. Un fait marquant est que le contour des voyelles semble être peu affecté par le caractère voisé ou non de la consonne qui la précède ou qui la succède. Ces focus, présent sur les voyelles, présentent des grandes dynamiques. En effet, entre les valeurs initiales/finales de la voyelle et les F0 maximales, des écarts de l'ordre 5 demi-tons sont régulièrement observés.

Les F0 des consonnes en positions initiales et finales sont bien moins intenses que celles des voyelles. De plus la F0 débute (ou finie pour les consonnes finales) à des valeurs similaires pour l'ensemble des 3 locuteurs (env. 18 demi-tons). D'autre part, les consonnes intervocaliques ne présentent pas des valeurs de F0 aussi basses. En effet, la forme des contours de ces consonnes présentent une forme concave, plus ou moins prononcée, en fonction du type de consonne. Il a été démontré que les consonnes liquides et nasales ne présentent pas des creux aussi proéminents (0 à 1 demi-ton par rapport aux valeurs initiales des voyelles adjacentes) que les consonnes fricatives ou occlusives (2-3 demi-tons).

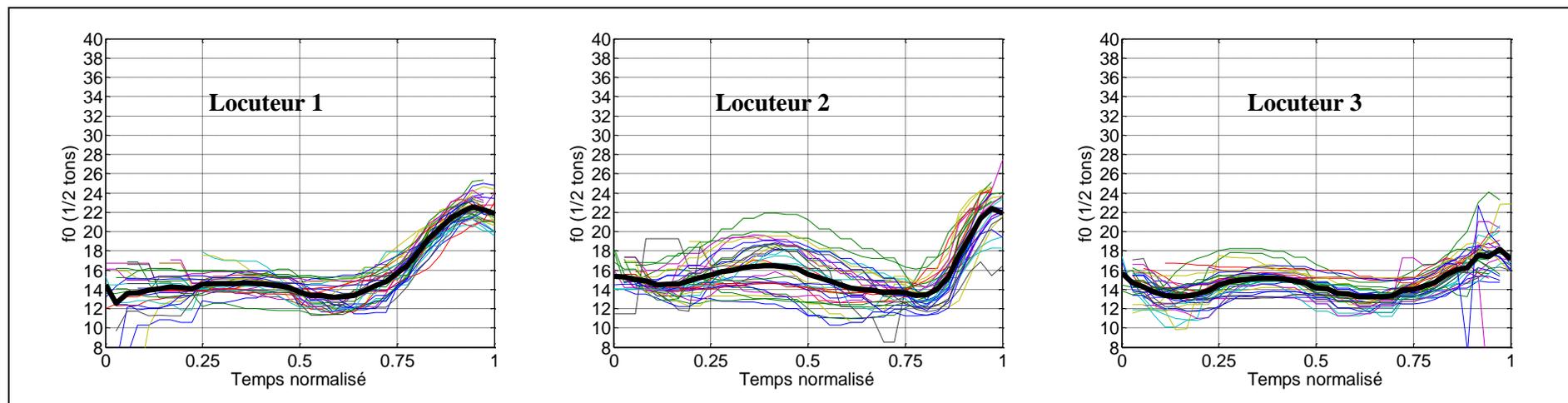


Figure 9-42 : Contour de F0 des logatomes CVCV en voix modale des 3 locuteurs dans le cas des consonnes voisées

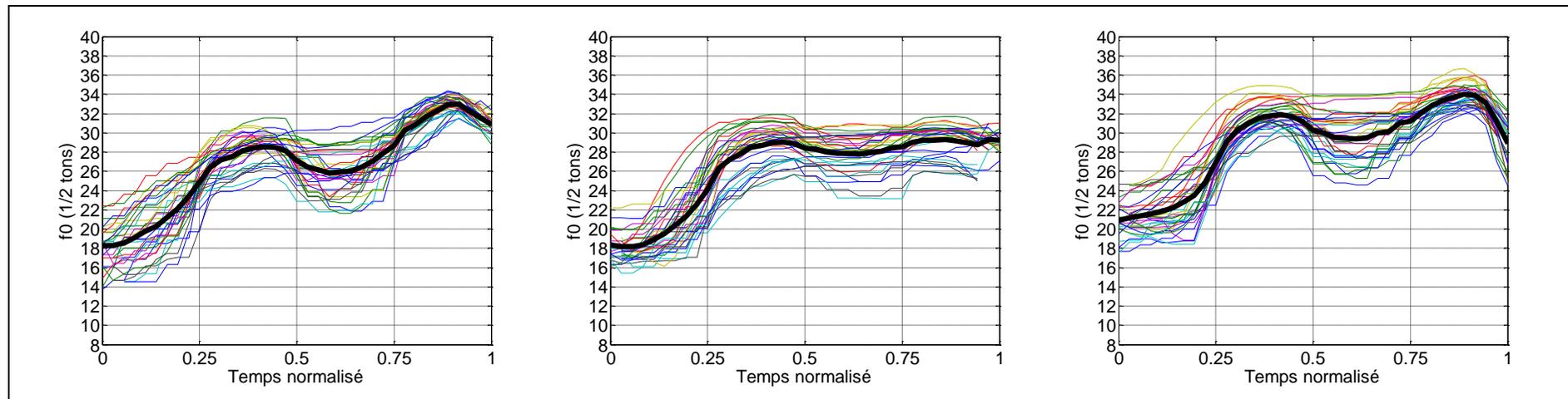


Figure 9-43 : Contour de F0 des logatomes CVCV en voix criée des 3 locuteurs dans le cas des consonnes voisées

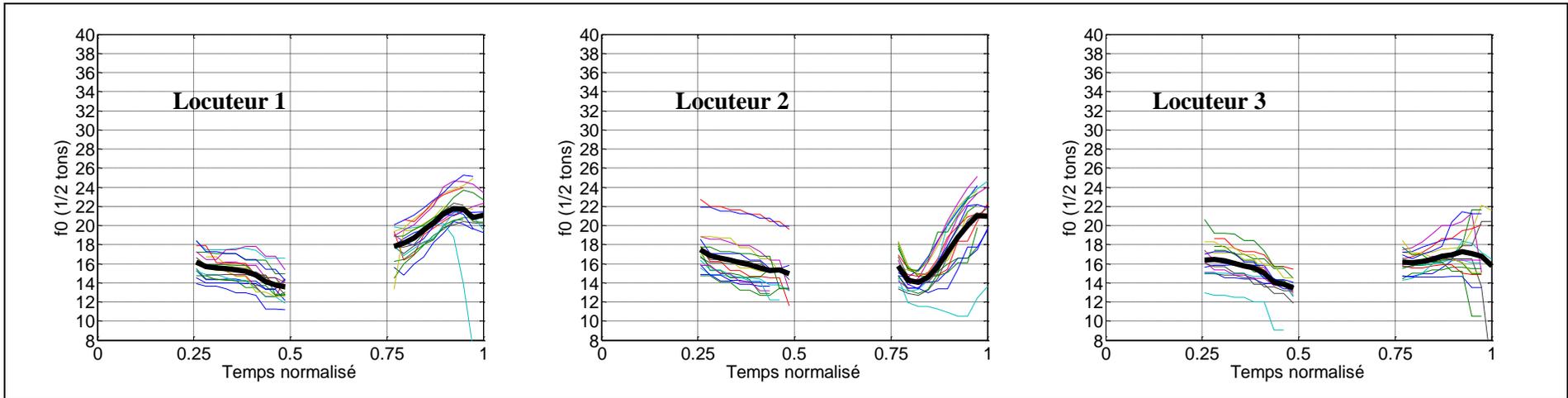


Figure 9-44: Contour de F0 des logotomes CVCV en voix modale des 3 locuteurs dans le cas des consonnes non voisées

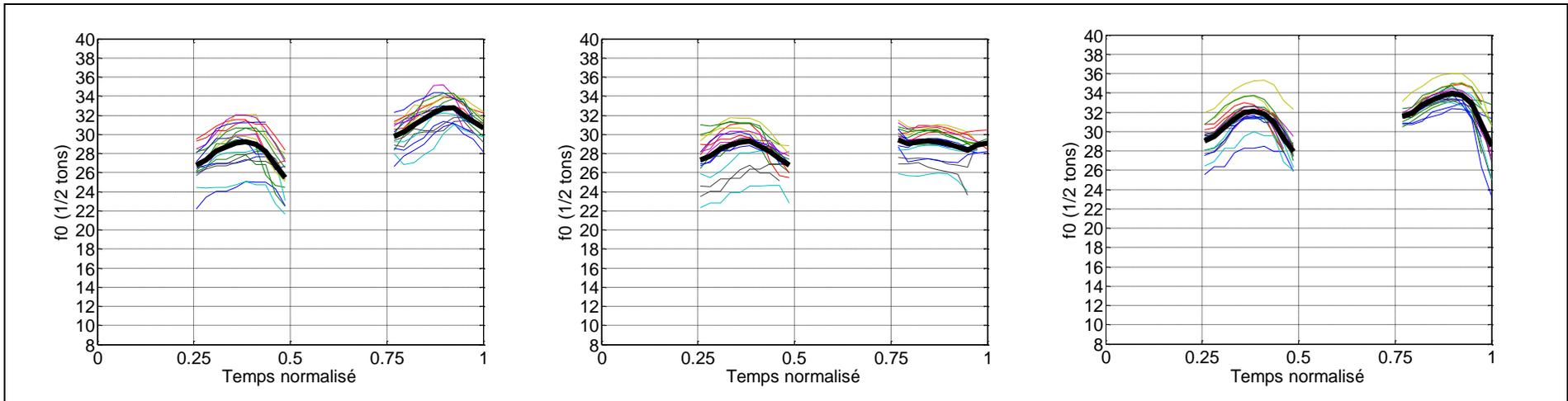


Figure 9-45: Contour de F0 des logotomes CVCV en voix crée des 3 locuteurs dans le cas des consonnes non voisées

## 9.5 Interprétations hypothétiques

Plusieurs éléments ressortent de cette étude et notamment des variations systématiquement plus fortes de l'intensité des voyelles, de la durée des voyelles ainsi que des contours de F0 similaires pour un logatome donné. Nous cherchons alors à comprendre les raisons de ces variations. S'agit-il de phénomènes volontaires ou involontaires? Toutefois, pour pouvoir interpréter les phénomènes observés, il nous faut la lumière des théories. Nous nous basons ainsi sur les différentes théories proposées permettant d'expliquer ces phénomènes. Trois types d'hypothèses peuvent être mentionnées, qui ne sont pas forcément indépendantes l'une de l'autre et peuvent très bien être combinées.

### 9.5.1 Hypothèse 1

Lors de la communication à distance, le signal arrivant aux oreilles de l'auditeur est dégradé par les phénomènes de propagation. Le signal résultant est alors moins intense, de composition spectrale légèrement différente et éventuellement plongé dans un bruit ambiant plus ou moins fort. Ceci détériore indéniablement l'intelligibilité de la voix. L'accès lexical ainsi perturbé, le locuteur emploie donc des stratégies particulières visant à compenser cette perte. L'une d'entre elles, se traduit dans notre situation, par **une complète réorganisation prosodique du message** ; tant au niveau mélodique, intonatif ou encore de la durée. A partir des contours mélodiques des quatre types de structures phonétiques étudiées, nous avons remarqué une hausse de F0 de manière générale. D'un point de vue linguistique, l'accentuation des voyelles correspondrait à un focus (i.e. une mise en relief de la partie de l'énoncé) qui consiste à pointer sur un élément précis du message. Le fait que l'augmentation de F0 est accompagnée d'une augmentation de l'intensité et de la durée des voyelles, corrobore cette théorie. **Le but de ce focus est alors d'insister sur une partie de l'énoncé. Il s'agit donc d'un contrôle volontaire fait par le locuteur dans le but de faire émerger l'information importante dans le flux de la parole.** Dans notre cas ceci signifie que les locuteurs emploient une stratégie visant à faire émerger les voyelles de l'énoncé. Ainsi plus d'effort est fourni sur les voyelles qui comme on le sait sont porteuse de plus d'énergie que les consonnes. Le fait de porter l'accent sur chaque voyelle nous amène alors à parler de focus métalinguistiques ((Jackendoff, 2002) d'après (Aubergé et Rilliard, 2006)). Par opposition au focus linguistique, visant à segmenter un flux de parole, le focus métalinguistique a pour rôle d'accentuer certaines parties du message. Dû à cet hypothétique focus sur les voyelles, l'intensité de celles-ci est plus forte. Ainsi, le rapport d'intensité consonne-voyelle diminue en voix criée. Or l'intelligibilité d'un énoncé est plus grand pour un rapport d'intensité C/V élevé (Kennedy, 1998). Dans le cas de la voix criée, celui-ci diminue. La voix criée doit alors

théoriquement perdre en intelligibilité. D'un point de vue de la conservation de l'intelligibilité la théorie du focus métalinguistique visant à faire émerger les voyelles et de permettre une parfaite compréhension de celle-ci (les consonnes pouvant être mal comprises) prend alors tout son sens.

### 9.5.2 Hypothèse 2

L'hypothèse précédente peut également se rapprocher des théories de Diehl (Diehl et Kluender, 1989) basés sur les travaux de (Traunmüller, 1981) qui démontre que la perception de l'ouverture d'une voyelle (i.e. de sa hauteur) est liée à la distance spectrale entre F0 et F1. La théorie de Diehl vise à dire que la F0 intrinsèque de chaque voyelle (différent en fonction de la position de l'articulation) n'est pas liée à des phénomènes physique mais qu'il s'agit de réactions intentionnelles de la part du locuteur visant à mieux délimiter les différentes catégories de voyelle (i.e. la hauteur ou l'ouverture). En effet, plus la voyelle est haute (i.e fermée) plus la F0 est élevée (Whalen et Levitt, 1995). Nous observons également d'après nos données (cf. Annexe E) l'effet de la voyelle sur la valeur maximale de la F0. Ainsi, comme la position des formants est augmentée en voix criée, liée à des ajustements du conduit vocal, il serait alors nécessaire d'augmenter la F0 dans le but de conserver la distance F0-F1 permettant l'identification de la hauteur de la voyelle. En effet, prenons l'exemple de la voyelle /a/ ayant la position de son premier formant aux alentours de 600 Hz en voix parlée. En voix criée, les différentes études rapportent régulièrement une position du premier formant d'un /a/ en voix criée à 800 Hz. Ainsi, dans le but de conserver la distance F0-F1, une augmentation de 200 Hz de la F0 sur la voyelle est nécessaire. Cet exemple reflète bien les différentes données que nous exposons ici. Il s'agit là d'une hypothèse qui considère le déplacement des formants comme une conséquence directe de l'ajustement volontaire du conduit vocal. Toutefois, elle n'explique pas les faits observés sur l'intensité ou sur les variations de la durée.

On peut aussi, évoquer la possibilité d'une troisième hypothèse basée sur les caractéristiques intrinsèques des phonèmes et leurs micro-prosodies qui contrairement aux deux précédentes hypothèses ne constitue pas en soit une action volontaire de la part des locuteurs.

### 9.5.3 Hypothèse 3

D'après les travaux de (Di Cristo, 1982) notamment, qui reprend également les travaux antérieurs aux siens, des phénomènes micro-prosodiques ont été identifiés pour la F0 des consonnes voisées notamment. En voix parlée, il est constaté des chutes de F0 pour les occlusives et les fricatives de l'ordre de 10 à 20 Hz avec des nuances légères en fonction de la position de la consonne dans le mot. En position initiale les occlusives sont croissantes avec un point de départ 10-20 Hz plus bas que la voyelle qui suit. Les fricatives en positions initiales arborent une signature concave avec un minimum

à 15-20 Hz en dessous de la voyelle et les liquides débutent 15-20 Hz en dessous de la valeur de la voyelle. Les nasales quant à elles, ne subissent pas de chute de F0. En position intervocalique les consonnes fricatives possèdent une signature concave avec un point bas inférieur de 15-20 Hz au reste du contour de F0, inférieur de 5-10 Hz pour les occlusives, les liquides voir les nasales (quand celles-ci ne s'intègrent pas parfaitement au continuum de F0 et ne possède ainsi aucune chute de F0). En position finale l'ensemble des consonnes subissent une chute de 15-20 Hz. (valeur issue de (Di Cristo, 1982) d'après une étude de (Léon et Martin, 1969)). Bien que les valeurs mentionnées ici soient discutables le fait d'observer des chutes de F0 en fonction de la classe des consonnes est indéniable. « *La plupart des recherches s'accordent, en effet, à admettre que ces variations découlent de la relation aérodynamique qui lie le flux d'air laryngien aux modes d'articulation du conduit vocal* » (Di Cristo 1982,p. 81). En effet, l'obstruction faite au niveau de la bouche génère une pression intra buccale élevée. Ainsi, l'équilibre entre la pression supra-glottique et subglottique, qui constitue la base de la mise la vibration des cordes vocales, est affecté. Afin de ne pas bloquer la vibration des cordes vocales, plusieurs ajustements sont possibles afin de maintenir un déséquilibre de pression entre l'amont et l'aval de la glotte. Pour cela, plusieurs procédés peuvent être mis en œuvre tels que ceux listés dans (Di Cristo, 1982) :

- Un abaissement du larynx
- Un élargissement du pharynx
- Une ouverture vélo pharyngal provoquant une fuite d'air vers le conduit nasale régulant ainsi la pression intra-orale
- Nous rajoutons également le fait qu'un abaissement de la mâchoire peut également en faire partie. En effet, cette action à pour conséquence d'accroître le volume de la cavité buccale et ainsi de diminuer la pression intra-orale. Ce type d'ajustement peut être mis en évidence en tentant de produire un /b/ le plus longtemps possible tout en maintenant la bouche fermée.

Ces adaptations visent à maintenir la vibration des cordes vocales malgré l'obstruction au niveau de la bouche, mais ne semblent pas être suffisant pour atteindre des valeurs de F0 aussi élevées que celles des voyelles. Le fait que les nasales et les liquides offrent moins de résistance à l'air, et qu'ils aient une F0 plus élevée que les fricatives ou les occlusives vient étayer cette hypothèse. Nos analyses vont également dans ce sens. En effet, une consonne en position finale ou initiale engendre une chute brutale de la F0 tandis qu'une consonne en position intervocalique engendre une forme de F0 concave qui est plus ou moins prononcée, voire non prononcée dans le cas des consonnes nasales et liquides. Notons toutefois que les ordres de grandeur sont dans le cas des voix criées bien plus élevée que ceux observés dans la littérature lors d'une phonation modale. En effet, alors qu'une chute de 10-20 Hz est attribuée aux consonnes dans notre étude nous observons des variations moyenne de 100 Hz de la base

de la consonne (initiale ou finale) jusqu'au début de la voyelle et près de 170 Hz jusqu'au maximum de la F0 de la voyelle. Les consonnes occlusives et fricatives en positions intervocaliques subissent quant à elles un abaissement de F0 de 30 Hz du minimum de la consonne jusqu'au début de la voyelle et de 70 Hz du minimum de la consonne jusqu'au maximum de la voyelle. Les nasales et liquides quant à elles possèdent une chute de F0 de l'ordre de 15 Hz jusqu'au début/fin de la voyelle et 40 Hz jusqu'au maximum de la voyelle.

Nous n'entrons ici pas plus dans les détails dans le sens ou pour répondre aux causes et rôles des phonèmes observés, nous estimons qu'une spécialisation en prosodie ou encore en aérodynamique du conduit vocal est primordiale et nous ne nous prêtons pas ces capacités là. De plus, nous n'avons pas la prétention de vouloir expliquer ces phénomènes mais le désir d'apporter un angle de réflexion sur ce sujet.

## 9.6 Conclusions du chapitre 9

---

Dans ce chapitre nous détaillons une étude nous permettant de mettre en évidence des phénomènes prosodiques révélateurs de la communication à distance et pertinents pour la perception de l'effort vocal (Fux et al., 2012a). Par ces observations nous souhaitons mettre en évidence une stratégie globale entre les locuteurs, permettant ainsi d'élaborer des règles de modification prosodiques dans le but de transformer une voix modale en voix criée. A partir des analyses réalisées dans ce chapitre nous pouvons tirer plusieurs conclusions.

D'une part, la longueur des voyelles est systématiquement allongée en voix criée. La diminution de la durée des consonnes observées par certains auteurs (Traunmüller et Eriksson, 2000) n'est pas systématique dans nos analyses. L'intensité varie également. Nous avons pu voir que l'intensité des voyelles en voix criée par rapport à la voix modale, augmente plus que celle des consonnes. Ce constat permet d'expliquer l'augmentation de la dynamique de l'intensité que nous avons pu observer dans le chapitre précédent. D'autre part, la F0 présente également des augmentations plus fortes sur les voyelles que sur les consonnes. De plus, en superposant les contours de F0 pour un locuteur donné, nous avons pu mettre en évidence des similitudes. En effet, sur l'ensemble des voyelles un focus est présent. Les consonnes voisées en positions initiales ou finales présentent une forte étendue. Ainsi, la F0 augmente plus pour les consonnes intervocaliques que pour les voyelles, qui elles augmentent plus que les consonnes en position initiales ou finales.

Par la suite nous avons tenté d'apporter une dimension théorique à ces variations. Plusieurs hypothèses sont mentionnées dans ce chapitre dans le but de mieux comprendre la cause de ces phénomènes.

Nous n'avons pas pu définir une hypothèse permettant d'expliquer ces derniers. Néanmoins, nous pensons que la théorie qui se base sur l'aérodynamique du conduit vocal prédomine. Néanmoins, une stratégie volontaire de la part des locuteurs accentuant ces phénomènes n'est pas à exclure. En effet, des différences de variations autour des phénomènes que nous avons mentionnés ont été observées entre les locuteurs. Ainsi ces différences interlocuteurs traduiraient des stratégies de communication différentes d'un locuteur à l'autre.

Ce chapitre nous permet ainsi de mieux caractériser la prosodie des voix criées. Toutefois, seules quatre structures phonétiques (CV, CVC, VCV et CVCV), qui sont des logatomes ont été étudiés. En effet, l'analyse prosodique sur des phrases complètes est apparue bien trop complexe pour une première étude prosodique approfondie. D'autre part, seul 3 locuteurs et un niveau d'effort vocal ont été étudiés. De ce fait, bien que une stratégie globale peut être issue de cette étude le faible nombre de participant nous oblige à émettre certaines réserves sur nos conclusions. Plus de locuteurs auraient été préférables. Cependant, l'orientation tardive de l'étude vers la prosodie de l'effort vocal nous a obligé à contraindre notre étude à ces 3 locuteurs. Nous avons néanmoins, constitué une base de données de 1224 logatomes. (51 logatomes \* 4 structures \* 2 niveaux d'effort \* 3 locuteurs). Ainsi les analyses sur ces logatomes sont révélatrices pour un locuteur donné, mais les différences interlocuteurs sont difficilement interprétables dans notre situation.

Rappelons également que nos enregistrements sont basés sur des logatomes lus par les sujets. Ceci a tendance à accentuer certaines stratégies afin de maximiser l'intelligibilité des logatomes. Ce type d'enregistrement peut également engendrer un effet de liste. L'effet de liste, dans notre étude est toutefois un avantage permettant de conserver une homogénéité des contours prosodiques.

Des artefacts ont toutefois été observés qui se manifestent par des intonations interrogatives pour les locuteurs 1 et 2 notamment. Toutefois, malgré la gêne que cela peut occasionner, ces artefacts nous confortent dans le fait que l'exercice lié à l'enregistrement à bien était compris. En effet, la consigne était, lors de ces enregistrements, de communiquer une liste de mots dans le but de se faire comprendre. Ainsi ces sujets, en réalisant des intonations interrogatives, ont parfaitement assimilés se points et de ce fait inconsciemment pose la question : est ce que tu m'as bien compris ?

# **PARTIE III**

# **TRANSFORMATION**



# Transformation de la voix

---

*“Building high-quality voice transformation systems require to take into account phenomena that are usually ignored or overlooked in other speech research areas”*

— Yannis Stylianou

Jusqu’à présent, notre travail s’est consacré à l’analyse et la détermination des paramètres pertinents de la voix permettant à un auditeur d’apprécier la distance d’un locuteur. A travers ces analyses, nous avons alors pu mettre en évidence le rôle important de la prosodie dans la voix criée et dans la perception de la distance d’un locuteur. En effet, la prosodie est plus importante perceptivement que les paramètres spectraux ou que les variations liées à la propagation de l’onde sonore. C’est pourquoi notre étude s’est fortement concentrée sur l’étude de la prosodie des voix criées, liées à la communication à distance. Pour la voix chuchotée, le paramètre pertinent est l’absence de vibrations des cordes vocales. Toutefois, plusieurs ajustements sont nécessaires, dans le but d’améliorer la qualité des transformations de voix.

Cette partie constitue cependant une frontière avec les travaux détaillés jusqu’à présent. Notre objectif est ici de mettre en application les faits marquants que nous avons pu identifier, aussi bien pour les transformations de voix modale en voix chuchotée, que pour les transformations de voix modale en voix criée. Ce chapitre se consacre donc à la description de ces transformations et se décompose en 3 parties. Dans un premier temps, nous détaillons les méthodes d’analyse-synthèse que nous avons choisies pour réaliser ces transformations de voix.

Les deux parties suivantes sont dédiées, pour la première à la transformation de voix modale en voix chuchotée, et la deuxième à la transformation de voix modale en voix criée. Chacune de ces parties se compose tout d’abord d’une section ayant pour but d’établir les règles de transformation de la voix. La section suivante détaille les différentes étapes des algorithmes de transformation, permettant d’appliquer les règles qui auront été définies, que nous avons mises en place.

A ce stade de l’avancement, nous estimons judicieux de rappeler les objectifs que nous nous sommes fixés. Le but de ce travail est de transformer une voix modale en une voix criée ou chuchotée dans le but d’apporter une notion de distance d’un locuteur. Il s’agit toutefois, et notamment pour la transformation en voix criée, du stade expérimental de la démarche. En effet, les règles de

transformation n'existant pas et les paramètres pertinents pour la perception de la distance étant mal identifiés, ce type de transformations suscite encore l'intérêt des chercheurs afin d'établir des règles précises de transformation. D'autre part, la modification de la nature de la voix est encore aujourd'hui un domaine complexe dans lequel des résultats aboutissants à des voix naturelles sont utopiques. Les modifications drastiques devant être appliquées à la voix modale dans le but de la transformer en voix créée, engendrent également des problèmes de synthèse qui se traduisent également par une qualité de voix peu naturelle.

Ainsi, ayant conscience de la difficulté de la tâche à laquelle nous nous attelons, nous n'avons pas ici la prétention de créer des transformations de voix inaudibles permettant d'atteindre des qualités de voix naturelle. Notre objectif est de mettre en avant l'importance de certains paramètres, notamment dans la transformation de voix modale en voix créée, qui permettent de caractériser la distance. C'est pourquoi nous ne tentons pas dans ce chapitre de recréer l'ensemble des modifications qui ont pu être observées au cours de nos analyses ainsi que dans la littérature. Nous nous orientons principalement sur les modifications prosodiques pour les voix créées et sur le dévoisement pour les voix chuchotées. Néanmoins, des modifications de paramètres spectraux sont également appliquées.

## 10.1 Choix de la méthode d'analyse-synthèse

---

Les évolutions de ces dernières années dans le domaine de l'analyse-synthèse de la parole, offrent un large choix de méthodes d'approches et de complexités différentes (voir (Huang et al., 2009; Stylianou, 2009)). La majorité de ces méthodes sont basées sur la théorie acoustique de la production de la parole, et sont généralement appelées méthodes d'analyse/synthèse. Il existe également des méthodes, qui ne se basent sur aucun a priori sur le signal pour le modifier. Ce type de méthode est alors appelée méthode de modification. Ces méthodes possèdent l'avantage de ne pas modéliser le signal et, de ce fait, n'introduisent pas d'erreurs de modélisation lors de la transformation.

Parmi les méthodes d'analyse-synthèse (aussi appelées méthodes paramétriques), nous pouvons citer les plus répandues, comme les modèles sinusoïdaux ou harmoniques (*STC : Sinusoidal Transform Code* (McAulay et Quatieri, 1986), *HNM : Harmonic plus Noise Model* (Stylianou, 1996), *DSM : Deterministic plus stochastic model* (Drugman et al., 2009)), *STRAIGHT : Speech Transformation and Representation using Adaptive Interpolation of weigGHT spectrum* (Kawahara, 1999) ou encore les modèles basés sur *ARX (Auto-Regressive eXogenous)* (Agiomyrgiannakis et Rosec, 2009; Vincent et al., 2007). Ces dernières offrent l'avantage de permettre d'accéder à un nombre conséquent de paramètres, et d'atteindre de bonne qualité de synthèse et de naturel. Ces méthodes se basent toutefois

sur la modélisation acoustique de la production de la parole et ainsi, sur des modèles théoriques plus ou moins révélateurs des faits réels. Ce type de synthèse est toutefois complexe à mettre en œuvre.

Parmi les méthodes qui ne se basent sur aucun a priori sur la production de la parole, on peut mentionner les méthodes FD-PSOLA (*Frequency-Domain Pitch-Synchronous Overlap Add*) (Charpentier et Stella, 1986), TD-PSOLA (*Time-Domain Pitch-Synchronous Overlap Add*) (Moulines et Laroche, 1995) ou WSOLA (*Waveform Similarity Overlap Add*) (Verhelst et Roelands, 1993). Ces méthodes peuvent être appliquées sur n'importe quel signal. Pour notre étude, nous avons fait le choix d'utiliser ce type d'approche. Ce choix est effectué pour une raison particulière : le temps réel. En effet, nos travaux se destinent également à évaluer l'influence de certains paramètres pour la perception de la distance d'un locuteur. Ainsi, malgré le désir de réaliser des synthèses de bonne qualité, le choix d'une méthode d'analyse-synthèse permettant une application rapide tout en conservant une qualité de voix acceptable, est pour nous inévitable. En effet, l'utilisation de synthétiseur trop complexe nous aurait fait perdre un temps considérable que nous avons préféré dédier aux analyses, à l'élaboration des règles ainsi qu'à l'évaluation des transformations.

Étant donné que la transformation se fait essentiellement à l'aide de modifications de la prosodie, nous avons orienté notre choix de synthétiseur permettant de réaliser ces modifications de façon aisée. Afin de réaliser des modifications de F0 et de durée sur un signal, la méthode la plus employée est la méthode dite : d'addition-recouvrement de fenêtres temporelles synchrones avec le pitch (TD-PSOLA : *Time-Domain Pitch-Synchronous Overlap Add*). Soulignons que cette méthode ne permet pas d'accéder à des paramètres tels que les formants que nous désirons également modifier dans nos algorithmes. Une méthode dérivée de TD-PSOLA permet pourtant d'associer les capacités de modifications de TD-PSOLA avec celle d'une représentation source-filtre. Cette technique est appelée LP-PSOLA (*Linear Prediction Pitch-Synchronous Overlap Add*). Rappelons également que les deux méthodes précédentes souffrent de limites concernant les plages de modifications de la F0. Nous décrivons ici ces deux méthodes, ainsi qu'une troisième méthode permettant d'améliorer la qualité de modifications pour de fortes modifications de la F0.

### 10.1.1 Modifications dans le domaine temporel : TD-PSOLA

La technique dite d'addition-recouvrement de fenêtres temporelles synchrones avec le pitch (TD-PSOLA) (Moulines et Charpentier, 1990), est une technique de modifications applicable directement sur le signal de la parole qui utilise le principe de réharmonisation spectrale. Elle est basée sur le principe de l'addition-recouvrement (OLA).

Soit un signal  $s(t)$  périodique de période  $T_0$ , le signal  $s(t)$  peut alors se décomposer en la somme de fenêtres élémentaires  $s_i(t)$ , obtenues par la multiplication de  $s(t)$  et d'une fenêtre de pondération  $w(t)$

centrée sur un marqueur de pitch et s'étalant du marqueur de pitch précédent jusqu'au marqueur de pitch suivant. Soit une fenêtre couvrant l'intervalle:  $[(t_i - t_{i-1}), (t_{i+1} - t_i)]$ .

Les modifications de F0 sont effectuées en modifiant l'écartement relatif entre deux fenêtres élémentaires successives. Les modifications temporelles sont quant à elles effectuées par reproduction ou suppression de fenêtres élémentaires. Cette technique de modifications possède un atout majeur car elle ne fait aucun *a priori* sur le signal. TD-PSOLA ne fait pas appel à une modélisation du signal de la parole. De ce fait, la modification du signal via TD-PSOLA, permet d'éviter toute erreur de modélisation et est applicable à tous types de signaux. C'est en partie la raison pour laquelle cette méthode est largement utilisée et offre des qualités de modifications excellentes ; notons toutefois que cette méthode donne des résultats satisfaisants pour des plages de facteurs de modifications de F0, comprises entre 0.5 et 2 et des modifications temporelles de l'ordre de 0.25 à 2 (Boite, 2000).

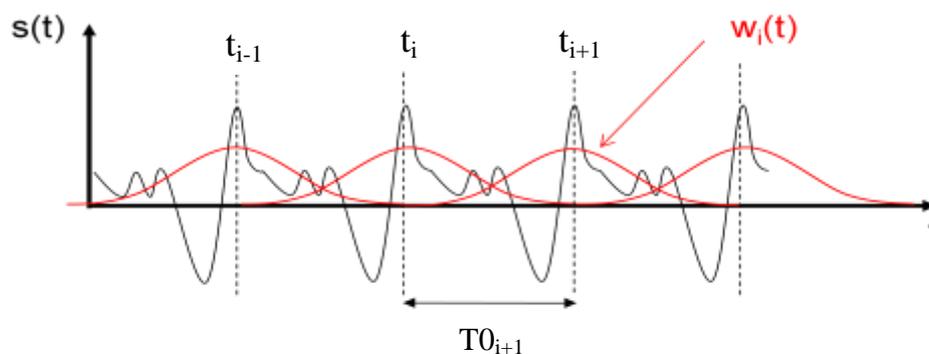


Figure 10-1: Illustration de la décomposition du signal de la parole en fenêtres élémentaires centrées sur les marqueurs de pitch

Néanmoins, cette méthode permet uniquement de réaliser des modifications prosodiques. Une méthode complémentaire a donc fait son apparition afin de permettre d'accéder aux paramètres spectraux, en utilisant une décomposition source filtre au préalable. Cette méthode est appelée LP-PSOLA.

### 10.1.2 Modifications dans le domaine temporel complétées d'une représentation source filtre : LP-PSOLA et RELP-PSOLA

Cette méthode permet d'allier la force de la méthode de modifications prosodiques TD-PSOLA, avec la modélisation de la décomposition source filtre (Moulines et Charpentier, 1990; Moulines et Laroche, 1995). Le principe de cette méthode consiste à décomposer chaque trame  $s_i(t)$  du signal  $s(t)$  suivant le modèle source-filtre. Ainsi, pour chaque trame, une estimation du filtre du conduit vocal est effectuée. Des modifications de ce filtre peuvent ainsi être effectuées.

Par la suite, un filtrage inverse permet de disposer du résidu de cette modélisation (signal de source + erreur de modélisation). La méthode de modifications prosodiques TD-PSOLA est alors effectuée sur

le résidu de la modélisation et non plus sur le signal en lui-même. La particularité de réaliser la re-synthèse à partir du résidu se nomme RELP-PSOLA (*Residual-excited linear prediction*) (Macchi et al., 1993). Enfin, en prenant soin de conserver la relation entre chaque trame du résidu (qui est déplacé par la méthode TD-PSOLA) et le filtre qui lui est associé, une reconstruction s'effectue en 2 temps. Dans un premier temps, le résidu est filtré par le filtre LP afin de reconstruire la trame modifiée. Dans un second temps, le signal complet est reconstruit en réassemblant l'ensemble des trames par la méthode OLA. Ceci permet d'appliquer des modifications prosodiques.

Cette méthode, comme la précédente souffre également de limites concernant la modification de F0. C'est pourquoi, une amélioration de cette méthode a fait son apparition dans le domaine de la modification de la parole. Il s'agit d'une méthode qui consiste à minimiser les plages de recouvrement des fenêtres successives lors de modifications drastiques de F0.

### 10.1.3 Modifications dans le domaine temporel complétées par d'une représentation source filtre et avec un ré-échantillonnage du résidu

La transformation de voix parlée en voix criée implique des modifications intenses de la F0. En effet, il n'est pas rare d'observer des modifications de l'ordre de 3 fois la valeur de la F0 d'origine. Il est bien connu que ce type de modifications engendre des artefacts lors de la reconstruction d'un signal par TD-PSOLA. En effet, pour ce type de transformations, les marqueurs de pitch de synthèse sont très proches les uns des autres. Ainsi, une fenêtre de synthèse s'étalera sur plusieurs fenêtres précédentes et suivantes (cf. Figure 10-2). L'addition superposition est alors fortement biaisée. Pour pallier ce problème, une première solution consiste à ré-échantillonner le résidu contenu dans une fenêtre d'analyses afin qu'il coïncide avec la longueur de la fenêtre de synthèse. Cette méthode proposée par Moulines et Charpentier (1990) permet ainsi de conserver la forme du résidu, même après de modifications poussées (i.e. *shape invariant method*). Cette méthode évoquée dans (Edgington et Lowry, 1996; Moulines et Charpentier, 1990), souffre toutefois d'un inconvénient qui est une diminution de la bande passante du signal.

En effet, le ré-échantillonnage du résidu diminue considérablement le nombre de points dans chaque fenêtre OLA. Prenons ici un exemple d'une fenêtre d'analyse d'un signal échantillonné à 16 kHz, ayant une F0 de 100 Hz. La fenêtre qui s'étale sur  $2 \cdot T_0$  correspond alors à 320 échantillons du signal. (en d'autres termes une bande passante de 8000 Hz). Dans le cas de l'augmentation de la F0 par un facteur 3, après ré-échantillonnage de la trame, ne contient plus que 107 échantillons. Le ré-échantillonnage tient compte approximativement d'un échantillon sur 3 de la fenêtre d'analyse. Ainsi, cette dernière trame ne correspond qu'à une bande passante de 2700 Hz (8000 Hz divisé par 3). Il s'agit ici d'un cas extrême, qui a pour but d'illustrer la conséquence du ré-échantillonnage sur la plage de fréquence d'une trame.

Dans une telle situation, il serait judicieux d'utiliser une décomposition de type « harmoniques plus bruit » (HNM) (Stylianou et al., 1995), permettant ainsi de modéliser le bruit afin de le reconstruire indépendamment du ré-échantillonnage et de ce fait de ne pas réduire la bande passante du signal par cette méthode. Toutefois, il s'agit là d'une amélioration pouvant augmenter la qualité de synthèses mais qui sort du cadre de notre étude.

Malgré la réduction de la bande passante dont nous avons conscience, nous choisissons la méthode basique qui consiste à ré-échantillonner le résidu en utilisant la méthode RELP-PSOLA, afin d'appliquer la modification de la F0.

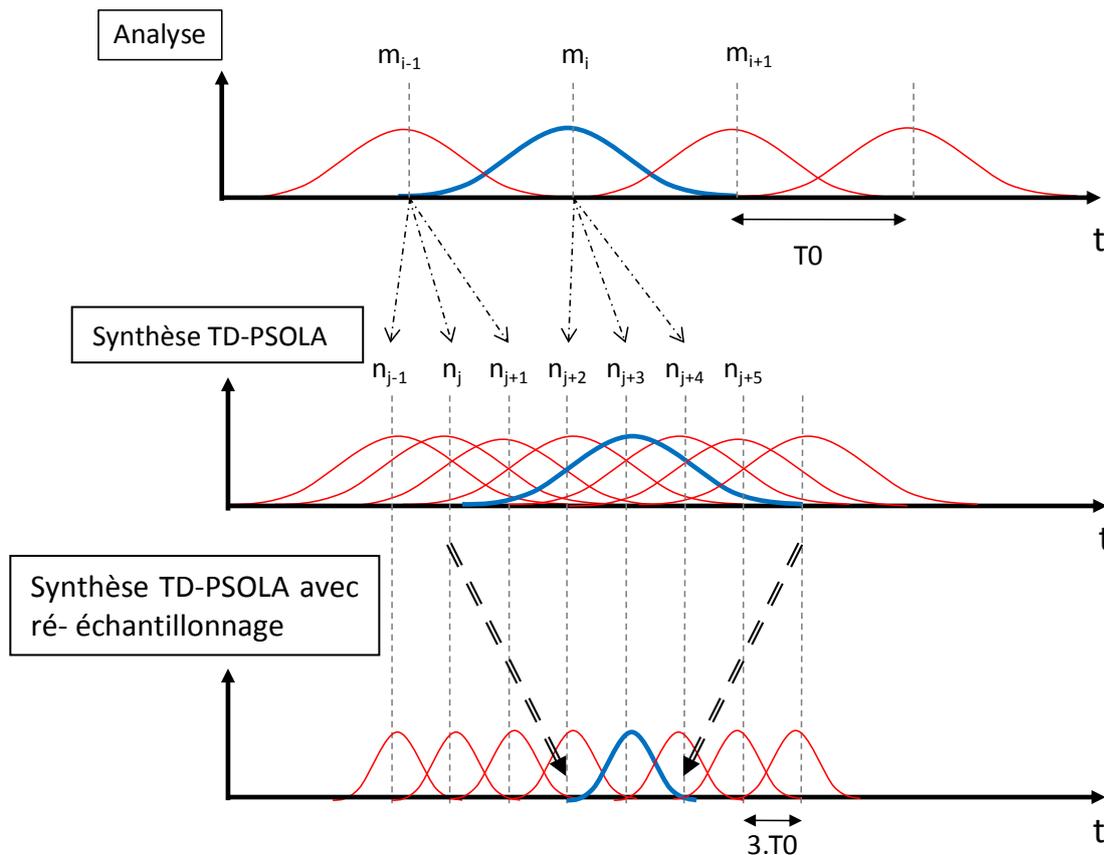


Figure 10-2: Principe de la synthèse TD-PSOLA avec ré-échantillonnage du résidu (en bas), comparé à la synthèse TD-PSOLA (au milieu) pour un exemple d'une augmentation de la F0 d'un facteur de 3.

## 10.2 Description des procédures d'analyse

Dans le but de réaliser une transformation de voix, différentes analyses sont nécessaires afin de disposer des paramètres à modifier. Il est ainsi primordial de choisir les méthodes d'analyses que nous utilisons, aussi bien pour l'estimation de la F0 que pour les filtres du conduit vocal. Cette section

expose ainsi les différents choix effectués et les raisons de ces choix pour mener à bien nos transformations.

Rappelons que nous avons choisi d'utiliser l'algorithme RELP-PSOLA avec ré-échantillonnage du résidu décrit dans le paragraphe précédent. La mise en œuvre de l'algorithme RELP-PSOLA nécessite avant tout de réaliser un marquage du pitch synchrone avec la F0. Pour ce faire, une estimation précise de la F0 est nécessaire. Nous détaillons ainsi, dans un premier temps, la solution retenue pour l'estimation de la F0 avant de nous pencher sur le marquage du pitch. Dans un second temps, nous avons pu constater que la position des formants varie entre la voix modale et la voix criée, mais également entre la voix modale et la voix chuchotée. Ainsi, afin de réaliser des transformations en accord avec les différentes observations, il est nécessaire de disposer de méthodes permettant de modifier la position de ces formants. La méthode retenue est ici présentée, ainsi qu'un bref état de l'art concernant la modification des formants.

### 10.2.1 Marquage du pitch

Afin d'appliquer l'algorithme RELP-PSOLA et ainsi d'appliquer les règles de modification de voix qui sont présentées par la suite, il est primordial de réaliser un marquage du pitch. Nous avons choisi pour réaliser ce marquage d'identifier les instants de fermetures des cordes vocales (*CGI, glottal closure instant*) sur le signal de la parole. Notons par ailleurs, que le marquage peut être effectué différemment en utilisant les maxima du signal ou encore les instants de corrélation des formes d'ondes. La première étape consiste à déterminer la valeur du pitch.

#### 10.2.1.1 Estimation de la F0

Nous avons choisi d'utiliser l'algorithme de Boersma (Boersma, 1993) utilisé dans PRAAT<sup>1</sup>. Pour réduire les éventuelles mauvaises estimations de F0 souvent attribuées à cette méthode, nous réalisons les estimations de F0 sur le signal EGG plutôt que sur le signal de la parole. Ce dernier étant plus harmonique que la parole, les algorithmes d'estimation ont généralement plus de facilités à déterminer la F0 sur ce type de signal. L'algorithme de Boersma est conçu sur la base de l'auto-corrélation d'une fenêtre, permettant ainsi de déterminer une périodicité dans celui-ci. De plus, il ne fait aucune considération qui restreint l'utilisation de cette méthode exclusivement au signal de la parole. De ce fait, il est parfaitement adapté à un signal autre que la parole, tel que le signal EGG.

L'algorithme de Boersma se base sur le principe de l'auto-corrélation d'une fenêtre du signal relativement large, dans le but d'en déterminer la corrélation maximale qui correspond à une période du signal de la parole. Certaines limites de détection correspondant à la F0 minimum et la F0

---

<sup>1</sup> Paul Boersma & David Weenink (2009), Praat: doing phonetics by computer [Computer program], from <http://www.praat.org/>

maximum espérée sont requises, ainsi que la suppression de l'influence de la fenêtre d'analyse pour mener à bien cette détection. D'autre part, l'algorithme de Boersma inclut également une optimisation visant à assurer une linéarité entre les différentes périodes estimées (estimation indépendante d'une fenêtre à l'autre). Cette optimisation consiste pour chacune des fenêtres d'analyse à déterminer plusieurs candidats de la valeur de F0, permettant ainsi de choisir parmi eux le chemin le plus probable pour obtenir une continuité du contour de F0. Pour réaliser cette optimisation, cet algorithme inclue des coûts tels que le coût de saut d'octaves, le coût de passage de zones voisées à une zone non voisée, etc... De plus, cet algorithme possède l'avantage de réaliser également une détermination des zones voisées et non voisées de façon plus précise que les autres méthodes, qui dans la majorité des cas nécessitent une analyse préalable permettant d'identifier les zones.

Rappelons qu'il existe de nombreuses méthodes permettant d'estimer le contour de F0 d'un signal. Pour plus de précisions on pourrait se référer à (Signal, 2009), qui liste de manière détaillée les différentes méthodes existantes. On y retrouve les méthodes temporelles, basées sur l'auto-corrélation (algorithme de Boersma utilisé dans PRAAT), sur l'inter-corrélation (fonction *get\_f0* inclus dans la librairie ESPS<sup>1</sup> et basée sur l'algorithme RAPT (*Robust Alogrithm for Pitch tracking*) (Talkin, 1995)), ou encore la méthode AMDF (*Average Magnitude Difference Function*), base de l'algorithme YIN (De Cheveigné et Kawahara, 2002) qui sont les plus répandues, ainsi que les méthodes spectrales telles que la méthode basée sur le cepstre, l'histogramme de Schroeder, le peigne de Martin, les méthodes SHS (*SubHarmonic Summation*) et bien d'autres.

### 10.2.1.2 Instant de fermeture ou marque de F0

L'instant de fermeture glottique correspond à l'instant d'excitation maximale du conduit vocal, qui correspond à l'instant de fermeture des cordes vocales. Ces instants de fermeture sont indispensables pour définir les fenêtres OLA. L'identification des instants de fermeture dans le signal de la parole est facilitée par l'utilisation d'enregistrements électroglottographiques. Grâce à ces signaux, la détection des instants de fermeture est facilitée. En effet, la dérivée du signal EGG (DEGG), présente des pics positifs fortement émergents, correspondant aux instants de fermeture (Henrich et al., 2004). Le signal EGG est toutefois décalé dans le temps par rapport au signal acoustique de la parole. En effet, le signal EGG est mesuré au niveau des cordes vocales tandis que le signal de la parole est enregistré plusieurs centimètres devant la bouche du locuteur. Nous utilisons dans cette étude cette dernière solution car nous disposons des enregistrements EGG pour DB2, DB3 et DB4.

Considérons le cas concret suivant. Le conduit vocal d'un homme mesure environ 17 cm de long. Dans les enregistrements de DB4, le microphone se situe à 15 cm de la bouche du locuteur. Ainsi, le signal d'excitation issu des cordes vocales parcourt 32 cm jusqu'à arriver au microphone. La célérité

---

<sup>1</sup> Librairie ESPS (Entropic Signal Processing System) disponible sur <http://www.speech.kth.se/software/>

du son étant d'environ 340 m/s, le signal EGG est alors en avance sur le signal de la parole de 0,94 ms (soit 15 échantillons à 16 kHz).

Dans la majorité des cas, le signal EGG peut ne pas être disponible. Dans ce cas l'estimation des instants de fermeture peut être réalisée de plusieurs manières (cf. (Degottex, 2010; Sturmel, 2011)). Dans une telle situation, l'estimation des instants de fermeture se fait en considérant la source glottique comme étant une suite d'impulsions, où chaque impulsion se situe au niveau de l'excitation maximale du conduit vocal. Nous citons ici les principes les plus répandus de la détection des instants de fermeture.

- Par détection des maxima de la norme de Frobenius (Kamp et Willems, 1994)
- Par détection des singularités du signal de la parole : détection des maxima les plus probables sur le signal de la parole (Laprie et Colotte, 1998), par détection du maximum de l'erreur de prédiction linéaire (Atal et al., 1971) par utilisation de transformations en ondelettes (Tuan et D'Alessandro, 1999)
- Par l'utilisation d'un modèle glottique (Degottex, 2010)
- Par le biais des délais de groupe : DYPSA (Kounoudes et al., 2002)
- Par le produit multi-échelle (Bouزيد et Ellouze, 2005)

Pour ce faire, nous considérons tout d'abord le signal DEGG dans son intégralité. Par la suite, pour toutes les zones voisées, l'algorithme de marquage détecte en premier lieu le maximum sur l'ensemble de la zone. Ce maximum est alors considéré comme étant le point de départ du marquage et constitue le premier instant de fermeture détecté. Par la suite, en accord avec la valeur de  $F_0$  en ce point, l'instant suivant se situe à une distance de  $T_0$  en avant de ce maximum. Nous répétons ainsi l'opération en considérant le dernier instant défini jusqu'à la fin de chaque partie voisée. Cette méthode, dû à la précision de la mesure ainsi qu'à la fréquence d'échantillonnage, peut souffrir d'une dérive progressive. C'est pourquoi, nous procédons à un affinement de la position de l'instant de fermeture. Nous retenons comme instant de fermeture, le point ayant le plus d'énergie dans une plage d'un échantillon, autour de l'instant défini par la considération de  $T_0$ . D'autre part, par définition dans les zones non voisées, il n'existe pas d'instant de fermeture. Ainsi, afin de bénéficier malgré tous d'instant de références pour l'utilisation de l'algorithme LP-PSOLA, nous plaçons dans les parties non voisées du signal, un marqueur de pitch toutes les 10 ms.

### 10.2.2 Mesure et modification des formants

La mesure des formants se fait via l'estimation du filtre du conduit vocal par la méthode du prédicteur linéaire (Markel et Gray, 1976). L'estimation de ce filtre se fait grâce à la méthode de résolution de

Levinson-Durbin (Durbin, 1960). Cet algorithme permet d'estimer les coefficients  $a_i$  d'un filtre autorégressif (AR) à partir du signal de la parole. A partir de ces coefficients, il est aisé de déterminer la position des formants ainsi que leurs largeurs de bande. En effet, la première étape consiste à déterminer les racines  $z_i$  du polynôme représentant le filtre (i.e. la position des pôles). Les racines sont alors décrites par  $z_i = r_i e^{\pm j\theta_i}$ , où  $r_i$  représente le radius du pôle et  $\theta_i$  sa pulsation. La détermination des fréquences des formants ( $F_i$ ) et de leurs largeurs de bande ( $Bw_i$ ) peut alors être exprimée de la manière suivante :

$$F_i = \frac{F_s}{2\pi} \theta_i, (Hz) \quad (\text{Eq. 10-1})$$

$$Bw_i = -\frac{F_s}{\pi} \ln(r_i), (Hz) \quad (\text{Eq. 10-2})$$

La modification de la position des formants se fait alors de la manière suivante. Après estimation du filtre LP, les formants  $F_i$  sont calculés. Les valeurs de ces formants sont ensuite ajustées à la valeur désirée avant de reconstruire successivement la valeur des pôles, le polynôme caractéristique du filtre et les coefficients du filtre AR.

Cette méthode permet de déplacer simplement la position des formants. Toutefois, déplacer un formant ou modifier sa largeur de bande peut engendrer des modifications importantes de la réponse impulsionnelle du filtre AR. En effet, par interaction entre les pôles, les formants peuvent être fortement diminués en amplitude, ou à l'inverse, augmentés par rapport aux caractéristiques d'origines. La déclinaison spectrale du filtre peut également être affectée. C'est pourquoi, certains auteurs se sont intéressés à cette problématique de la modification des formants, pour proposer des méthodes de modifications tenant compte de ces effets (Mizuno et Abe, 1996; Morris et Clements, 2002). Le but de ces études est de réaliser le déplacement d'un ou plusieurs formants tout en conservant le reste de la réponse impulsionnelle identique. En d'autres termes, cette méthode permet de prévoir et corriger les interactions entre les pôles.

Toutefois, en première approche, nous nous contentons dans notre étude de modifier la position des formants par la méthode la plus basique décrite ci-dessus, tout en ayant conscience des biais que cela peut impliquer et en ayant conscience des problèmes de stabilité que cela peut engendrer. Concernant ce dernier point, nous avons complété l'algorithme par une vérification de la position des formants afin de s'assurer que les valeurs obtenues se situent dans la plage de fréquence appropriée (i.e.  $F_i \in [200, 1000]$ ). Dans le cas contraire l'algorithme de modification signale une erreur.

---

## 10.3 Transformation de voix modale vers une voix chuchotée

---

Les procédures précédemment décrites nous permettent de concrétiser les transformations de voix proposées. Cependant, l'essentiel reste à faire : quantifier les modifications qui sont appliquées aux voix modales. Nous nous concentrons ici sur la modification du signal de la parole sans tenir compte des effets liés à la propagation de l'onde sonore.

Comme nous avons pu le voir, la transformation de voix modale en voix chuchotée passe avant tout par le dévoisement du signal de la parole. Toutefois, il est également nécessaire d'appliquer des modifications spectrales et prosodiques pour obtenir des voix transformées plus naturelles. Nous proposons ici différentes règles pour réaliser cette transformation. C'est pourquoi, avant de rentrer dans le vif du sujet qui concerne l'algorithme de transformation en lui-même une section est consacrée à l'élaboration des règles de modification propre à la transformation.

### 10.3.1 Élaboration des règles de transformation

Dans le but d'améliorer la qualité des transformations en voix chuchotée classique nous nous attachons ici à définir des règles établies sur la base de la littérature. Ces règles ont pour but de recréer les faits les plus pertinents des voix chuchotées. Nous nous basons ainsi sur les valeurs et réflexions issues de la littérature. Nous proposons ici des règles de modification qui se concentrent sur le dévoisement, par l'utilisation d'une source de bruit qui se rapproche plus de la réalité que le bruit blanc traditionnellement utilisé. Par la suite nous proposons une règle de transformation de la position des formants, de la pente spectrale et enfin une modification de l'intensité des voyelles uniquement.

La principale modification que nous appliquons dans cet algorithme consiste à créer une source sonore proche de celle générée par la glotte lors du passage de l'air issu des poumons. D'autre part nous déplaçons les formants liés aux conditions aux limites différentes (Barney et al., 2007; Matsuda et Kasuya, 1999; Swerdlin et al., 2010) en voix chuchotée (i.e. glotte ouverte) ainsi qu'une modification relative d'intensité entre les voyelles et les consonnes (Ito et al., 2005; Jovičić et Šarić, 2008). L'ensemble de ces règles de modification sont présentées ci-dessous.

#### 10.3.1.1 *Génération de la source de bruit de constriction*

Même si un bruit blanc peut être suffisant pour générer une voix chuchotée, nous choisissons toutefois d'utiliser comme source glottique un bruit ayant les caractéristiques d'un bruit de constriction. D'un point de vue théorique, assimiler le bruit de source de la voix chuchotée à un bruit de constriction

semble être plus réaliste qu'un bruit blanc. En effet, rappelons que la source glottique lors d'un chuchotement est créée par les turbulences générées par la constriction de la glotte. De plus, l'utilisation d'un bruit blanc peut engendrer certaines problématiques au niveau de la reconstruction du signal. En effet, l'analyse LPC sur de courtes fenêtres d'analyse a pour conséquence une mauvaise estimation des bases fréquences (BF). Ceci engendre généralement une estimation des BF trop forte. Dans le cas d'une reconstruction du signal par une source voisée, ceci n'engendre pas de conséquences dramatiques étant donné la faible énergie, voir l'absence d'énergie en deçà de la fréquence de la fondamentale (env. une centaine de Hz). Toutefois dans le cas de l'utilisation d'un bruit blanc cette zone est excitée ce qui résulte en des voix modifiées contenant trop de basses fréquences et ainsi une voix résultante trop grave par rapport à une voix chuchotée naturelle. En ce sens, un simple filtrage d'un bruit blanc par un passe haut peut alors convenir. Toutefois, plutôt que d'utiliser arbitrairement un filtre passe-haut, nous préférons nous rattacher à la théorie et proposer un filtre passe-bande représentatif de la source glottique en voix chuchotée.

La source de bruit de turbulence est décrite théoriquement comme ayant un spectre relativement plat sur une bande de fréquence de 2-3 octaves autour d'une fréquence centrale comprise entre 500 et 3000 Hz (Stevens, 1971). La fréquence centrale est définie comme étant  $0,2 U/A^{2/3}$  où  $U$  est le débit d'air et  $A$  est la section glottique. L'air glottique lors de la production de voyelle chuchotée peut varier de 0,1 à 0,5 cm<sup>2</sup> et le débit peut être au maximum de 1500 cm<sup>3</sup>/s (Stevens, 1971). Ainsi en considérant des valeurs intermédiaire la fréquence centrale peut être calculée et est alors de 1200 Hz. Il a été démontré que ce type de spectre est une bonne approximation de la source d'une voix chuchotée (Hillman et al., 1983). Nous réalisons ce filtre en utilisant un filtre passe-bande de type *Butterworth*. Le choix d'un filtre de *Butterworth* est arbitraire. Nous avons principalement choisi d'utiliser un filtre de *Butterworth* car il permet de reconstruire facilement le spectre mentionné sans avoir à utiliser un grand nombre de coefficients comme il l'aurait fallu avec un filtre à réponse impulsionnelle finie (FIR). Dans l'optique d'une application temps réel il est plus judicieux de s'orienter vers des filtres à réponse impulsionnelle infinie (IIR) qui pour un filtre donné génère moins de retard de phase entre l'entrée et la sortie du système. De plus ce type de filtre est réalisable analogiquement. Il offre également l'avantage d'être facilement paramétrable afin d'en modifier ces fréquences de coupure ou sa largeur de bande.

Ainsi en tenant compte de ces caractéristiques du bruit de constriction, et sur la base des valeurs mentionnées dans la littérature pour ce type de bruit, nous proposons un filtre numérique permettant de créer une telle signature fréquentielle à partir d'un bruit blanc. Nous choisissons ainsi un filtre passe bande ayant une bande passante d'une largeur de 3 octaves ( $F_{c1}=424$  Hz et  $F_{c2}=3394$  Hz). La fonction de transfert de ce filtre est donnée ci-dessous (cf. Eq.10-4) et est représenté sur la Figure 10-3 ( $F_s=16$  kHz).

$$G(z) = \frac{0,3975 - 0,3975 \cdot z^{-2}}{1 - 1,0566 \cdot z^{-1} + 0,2051 \cdot z^{-2}} \quad (\text{Eq. 10-4})$$

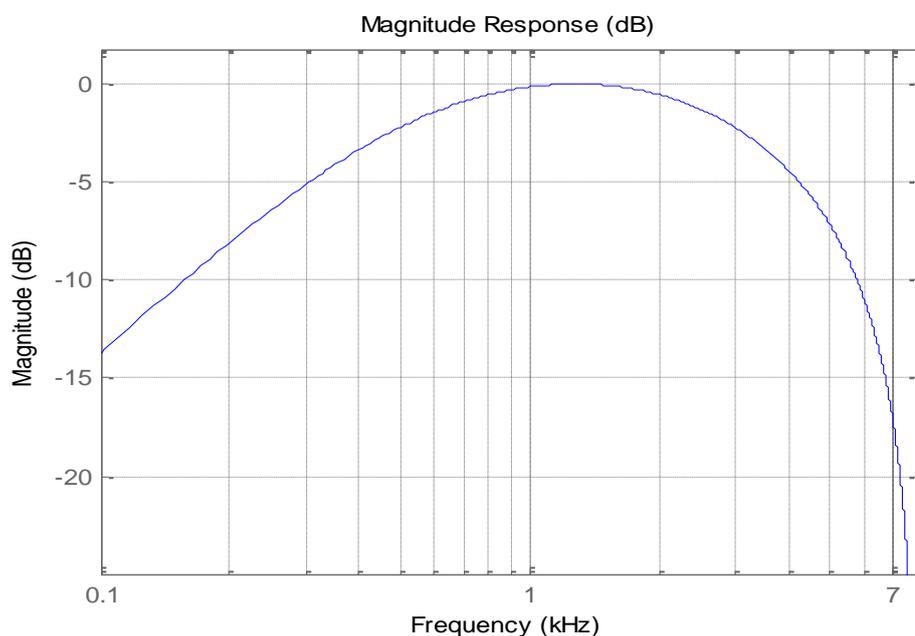


Figure 10-3: Spectre théorique de la source de bruit de constriction

### 10.3.1.2 Modifications des formants

L'absence de vibrations des cordes vocales, lors de la production de voix chuchotée, laisse place à une glotte ouverte. Cette modification des conditions aux limites du conduit vocal engendre un déplacement des formants vers les hautes fréquences (cf. CHAPITRE 5). Ce déplacement relativement important, semble être un élément pertinent dans la transformation en voix chuchotée. Bien que les études, visant à transformer une voix chuchotée en voix parlée, tiennent compte de ces variations (Morris, 2003), nous n'avons pas trouvé d'études de transformations de voix modale en voix chuchotée qui fassent de même. Nous pensons toutefois que tenir compte de ces déplacements des formants a pour conséquence d'améliorer le naturel ainsi que l'intelligibilité des transformations. C'est pourquoi, nous nous basons ici sur une étude visant à reconstruire une voix modale à partir d'une voix chuchotée, pour en déduire la règle de transformation des formants à appliquer dans notre cas.

Dans l'objectif inverse de notre étude, c'est-à-dire la transformation de voix chuchotée en voix modale, Morris (2003) propose une modélisation de la modification des formants. La modélisation du déplacement des formants est effectuée via une fonction linéaire par partie qui définit le déplacement des formants inférieurs à 2500 Hz, entre la voix modale et la voix criée. La Figure 10-4 représente le déplacement des formants entre une voix chuchotée et une voix parlée pour des locuteurs masculins. Les données sont issues de (Kallail et Emanuel, 1984b). La figure représente toutefois un déplacement en valeurs absolues. Il est bien connu que les formants en voix chuchotée

sont déplacés vers les hautes fréquences par rapport à la voix parlée. Ainsi les déplacements mentionnés sur cette figure, correspondent à des déplacements vers les basses fréquences. Sur cette figure, en traits pleins, est également représenté le modèle proposé par Morris (Morris, 2003).

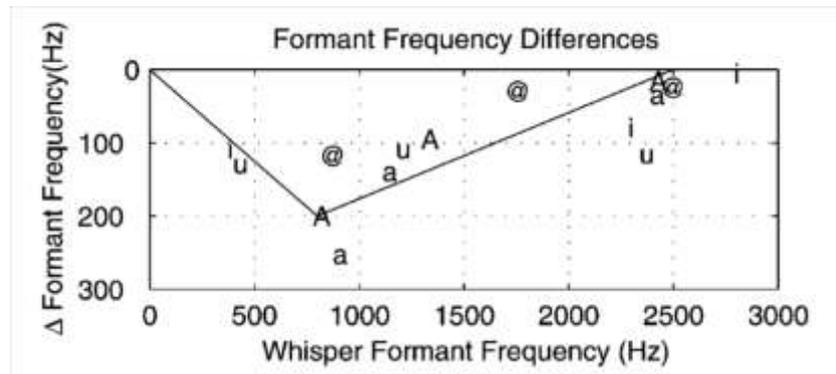


Figure 10-4 : Position des formants en voix chuchotée en fonction de la différence (en valeurs absolue) entre la fréquence du formant en voix chuchotée et en voix modale. Le trait plein représente l'une des modélisations proposée par Morris (d'après (Morris, 2003))

Sur cette Figure 10-4, le maximum de déplacement intervient pour un formant en voix chuchotée, situé à 600 Hz. Ainsi, une voyelle chuchotée ayant un formant situé à 600 Hz, verra la position de ce formant déplacée de 200 Hz en voix parlée. La position de ce formant en voix parlée serait alors de 400 Hz. Après avoir affiné son modèle Morris (2003) estime qu'une meilleure modélisation consiste à choisir le point d'inflexion à 700 Hz pour un déplacement équivalent de 300 Hz.

En comparant les résultats de plusieurs études, nous arrivons à une meilleure concordance avec le point d'inflexion placé à 787 Hz pour un déplacement de 227 Hz. La Figure 10-5 représente l'ensemble des données pour les formants allant de F1 à F4 en voix parlée, en fonction de leurs positions en voix chuchotée. Les points sur le graphique sont issus des études suivantes : (Ito et al., 2005; Jovičić, 1998; Matsuda et Kasuya, 1999; Petrushin et al., 2010; Sharifzadeh et al., 2010; Swerdlin et al., 2010). La courbe rouge correspond à notre estimation du modèle linéaire par parties proposé par Morris que nous affinions par minimisation de l'erreur quadratique. La droite en pointillés correspond à une droite de coefficient directeur unitaire.

Ainsi, en inversant le modèle de Morris (2003), et en utilisant le point d'inflexion estimé grâce à plusieurs relevés indépendants, nous proposons les règles de transformation des formants en voix parlée vers les formants en voix chuchotée. Ce qui se résume par le système d'équation ci-dessous, représentant la loi de modification illustrée sur la Figure 10-6.

$$F'_x = \begin{cases} 1,41 \cdot F_x & , & 0 < F_x \leq 560 \\ 0,88 \cdot F_x + 292 & , & 560 < F_x \leq 2500 \\ F_x & , & F_x > 2500 \end{cases} \quad (\text{Eq. 10-5})$$

où  $F_x$  est la fréquence centrale du formant considéré en voix parlée et  $F'_x$  la fréquence centrale estimée du formant en voix chuchotée.

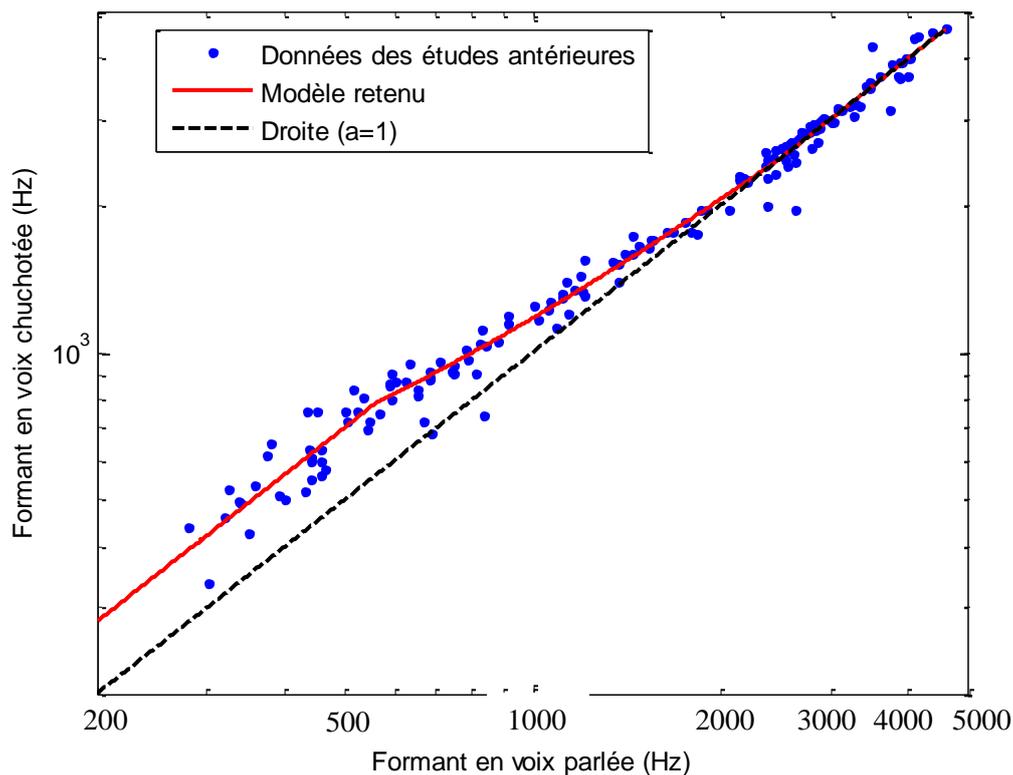


Figure 10-5 : Affinement du modèle de Morris pour la modification des fréquences des formants. Les valeurs des formants (points bleus) sont issues des études suivantes : (Ito et al., 2005; Jovičić, 1998; Matsuda et Kasuya, 1999; Petrushin et al., 2010; Sharifzadeh et al., 2010; Swerdlin et al., 2010)

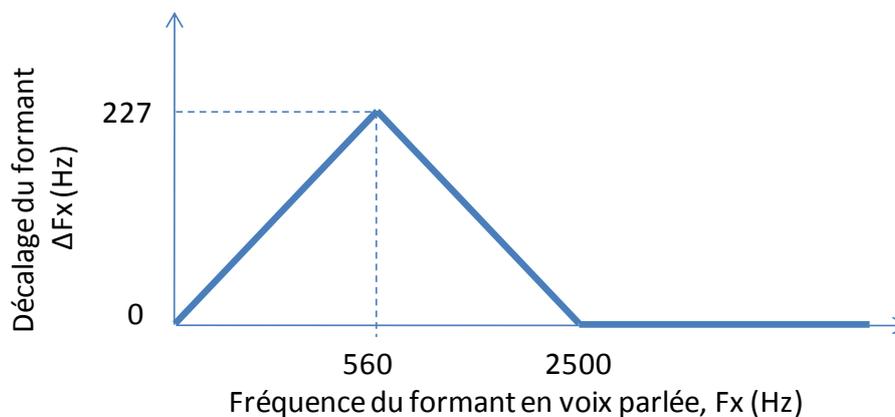


Figure 10-6 : Schématisation de la règle de déplacement des formants pour la transformation en voix chuchotée

Même si l'augmentation de la largeur de bandes en voix chuchotée, par rapport à la voix parlée, est un fait avéré (Jovičić, 1998), nous avons fait le choix de ne pas modifier ces dernières. La première raison est que nous pensons qu'une augmentation des largeurs de bande des formants peut sensiblement dégrader l'intelligibilité. Ainsi nous préférons dans ce cas privilégier la compréhension du message plutôt que le naturel des transformations. D'autre part, la modélisation du conduit vocal par un filtre AR laisse place à des erreurs de modélisation, et en particulier en ce qui concerne les largeurs de bande. Ces erreurs de modélisation se reflète en particulier par une sous-estimation considérable des

largeurs de bande (Paliwal, 1984). Ainsi, une augmentation de ces largeurs pourrait engendrer des largeurs bien différentes que celles observées en voix chuchotée, et ainsi éventuellement engendrer une perte d'intelligibilité. Enfin, des tests informels ont montrés que cette modification n'améliore pas considérablement la qualité des modifications et de plus, crée de manière régulière des erreurs lors de la synthèse. Pour toutes ces raisons, nous ne considérons pas, dans nos règles de transformation, l'augmentation des largeurs de bande.

#### 10.3.1.3 *Pente spectrale*

L'excitation du conduit vocal par un bruit de turbulence (spectre relativement plat) plutôt que par l'onde de débit glottique (d'une pente théorique de  $-12$  dB/octave) engendre une modification de la pente spectrale de la voix chuchotée. En considérant le modèle source filtre basique, l'utilisation de l'excitation uniquement par bruit blanc, engendre une pente spectrale théorique positive de  $+6$  dB/octave). En effet, la source de bruit ( $0$  dB/octave) modulée par le filtre du conduit vocal ( $+0$  dB/octave) puis rayonnée par les lèvres ( $+6$  dB/octave) dispose d'une pente spectrale qui correspond plus à celle d'une consonne. Or, le spectre d'une voyelle chuchotée dispose généralement d'une pente spectrale de  $+3$  dB/octave (Zhang et Hansen, 2007). En effet, l'utilisation d'un bruit blanc dans le modèle source filtre, est destiné à recréer les sons non voisés, et donc certaines des consonnes. On s'aperçoit que la synthèse, en utilisant un bruit blanc via le modèle source filtre, n'est pas adaptée à créer des voix chuchotées.

Ainsi pour pallier cette problématique théorique, nous choisissons de ne pas appliquer le rayonnement aux lèvres pour les trames que nous « dévoisons ». Ceci revient à modifier la pente spectrale de la source de bruit. Toutefois, nous choisissons la première solution dans un souci de simplicité. Remarquons alors que la voix transformée disposera d'une pente spectrale de  $0$  dB/octave et non pas de  $3$  dB/octave comme mentionné dans la littérature. Néanmoins, les valeurs concernant la pente spectrale des voix chuchotées étant rares, la majorité des études mentionne un spectre relativement plat. C'est pourquoi, en première approche, nous retenons cette solution.

#### 10.3.1.4 *Modification d'intensité*

La voix chuchotée présente des différences d'intensité entre type de phonème. En effet, rappelons que les phonèmes non voisés ne varient, au maximum, que de  $3,5$  dB par rapport à une voix parlée et que les phonèmes voisés, quant à eux diminuent de  $20$  à  $25$  dB (Ito et al., 2005; Jovičić et Šarić, 2008). La différence d'intensité entre phonèmes voisés et non voisés étant conséquente, nous prenons également cette caractéristique en compte pour la transformation de voix modale en voix chuchotée. Ainsi, nous proposons d'appliquer à toutes trames voisées, un gain de  $1/10$  (i.e.  $-20$  dB) pour être en cohérence avec les relevés effectués par les études antérieures.

### 10.3.2 Réalisation de l'algorithme de transformation de voix modale en voix chuchotée

Après avoir détaillé les règles de transformation, nous présentons dans ce paragraphe l'algorithme pour produire une voix chuchotée à partir d'une voix modale. L'algorithme de transformation de voix modale en voix chuchotée se décompose en plusieurs étapes. La Figure 10-7 représente cet algorithme sous forme de bloc fonctionnel.

#### Étape 1.

La première étape consiste à estimer la fréquence fondamentale via l'algorithme de détection du pitch réalisé par Boersma pour, par la suite, permettre d'estimer les instants de fermeture glottique.

#### Étape 2.

Le signal  $s(t)$  échantillonné à 16 kHz est ensuite **décomposé** en trame  $s_i(t)$ , centré sur un marqueur de pitch ( $t_i$ ) et allant de  $t_{i-1}$  à  $t_{i+1}$ . Ceci correspond à un recouvrement proche de 50%. Par la suite, pour chacune de ces trames les étapes suivantes sont réalisées.

#### Étape 3.

Pour chaque trame, il est nécessaire de définir si celle-ci contient une partie du signal voisé ou non-voisé. La **détection du voisement** peut être réalisée à partir de méthodes spécifiques à la détection de voisement (cf. (Kondoz, 2004)). Toutefois, dans notre situation, nous utilisons l'algorithme de détection du pitch réalisé par Boersma (cf. étape 1) qui intègre de telles méthodes. Ainsi, pour déterminer le caractère voisé ou non d'une trame nous nous référons aux résultats obtenus à l'étape 1. Les trames qui ne présentent pas un caractère voisé sur l'intégralité de leur longueur sont également considérées comme non-voisées ; ceci, afin d'éliminer la totalité des périodicités qui auront pu être identifiées.

#### Étape 4.

Les trames considérées comme non voisées ne sont aucunement modifiées ; l'ensemble des trames non-voisées est directement redirigé sur la sortie, sans avoir subi une quelconque modification. Nous avons choisi cette solution pour deux raisons en particulier. D'une part, plusieurs auteurs montrent que la voix chuchotée n'engendre que peu, ou pas, de modifications lors de la phonation des sons non-voisés (Jovičić et Šarić, 2008). De plus, la modélisation via un prédicteur linéaire engendre certaines difficultés quant à la modélisation des zones non-voisées et notamment les fricatives. Ainsi, en ne considérant pas ces trames, nous espérons réduire les éventuels artefacts liés à la mauvaise modélisation et ainsi, améliorer la qualité de la transformation.

En revanche, pour les trames voisées, une modification de la fréquence des formants, de l'intensité ainsi qu'un dévoisement est effectuée.

- a- Dans un premier temps, l'intensité de la trame est calculée
- b- Par la suite la trame considérée est préaccentuée par un filtre de **préaccentuation** ( $V(z)=1-0,95z^{-1}$ ). Les coefficients du prédicteur sont ensuite estimés.
- c- L'**estimation des coefficients LP** est réalisée à partir du signal  $s_i(t)$  pondéré par une fenêtre de *Hamming*, en utilisant une méthode de prédiction linéaire (Markel et Gray, 1976), ainsi que l'algorithme de résolution de Durbin-Levinson (Durbin, 1960). L'ordre  $p$  du filtre AR est fixé à 18. (i.e.  $p = F_s/1000 + 2$ )
- d- A partir des coefficients LP, une **modification des formants** est appliquée au filtre  $H(z)$ . Les fréquences des formants sont modifiées en accord avec la loi de transformation précédemment décrite. Toutefois, les largeurs de bande de ces derniers sont quant à elles préservées. Les modifications du filtre engendrent ainsi le filtre de synthèse  $H'(z)$ .
- e- Afin de réaliser le dévoisement des trames  $s_i(t)$  considérées comme voisées, un **bruit blanc** est généré. Ce bruit blanc est, par la suite, **filtré** par la réponse en fréquence correspondant à un bruit de constriction que nous avons décrit précédemment. L'intensité du bruit ainsi filtré est **ajustée et la modification de l'intensité** des zones voisées est appliquée à cet instant (i.e. une diminution de 20 dB). A noter qu'aucune modification d'intensité n'est appliquée sur les zones non-voisées.

### Étape 5.

Enfin, la **synthèse** du signal est effectuée en filtrant le bruit précédemment créé par le filtre de synthèse  $H'(z)$ . Il est à noter que nous n'appliquons à aucun moment un filtre lié au rayonnement aux lèvres. En effet, celui-ci engendre des voix transformées trop stridentes. Nous pensons alors que le filtre utilisé pour créer le bruit de source n'est pas parfaitement adapté ou que sa modélisation contient déjà dans une certaine mesure, des phénomènes liés au rayonnement.

### Étape 6.

La dernière étape consiste à **reconstruire le signal** suivant la technique OLA à partir des trames modifiées, ou les trames originelles (dans le cas où celles-ci sont non voisées) en utilisant une pondération par fenêtre de *Hanning* (i.e. la fenêtre de synthèse).

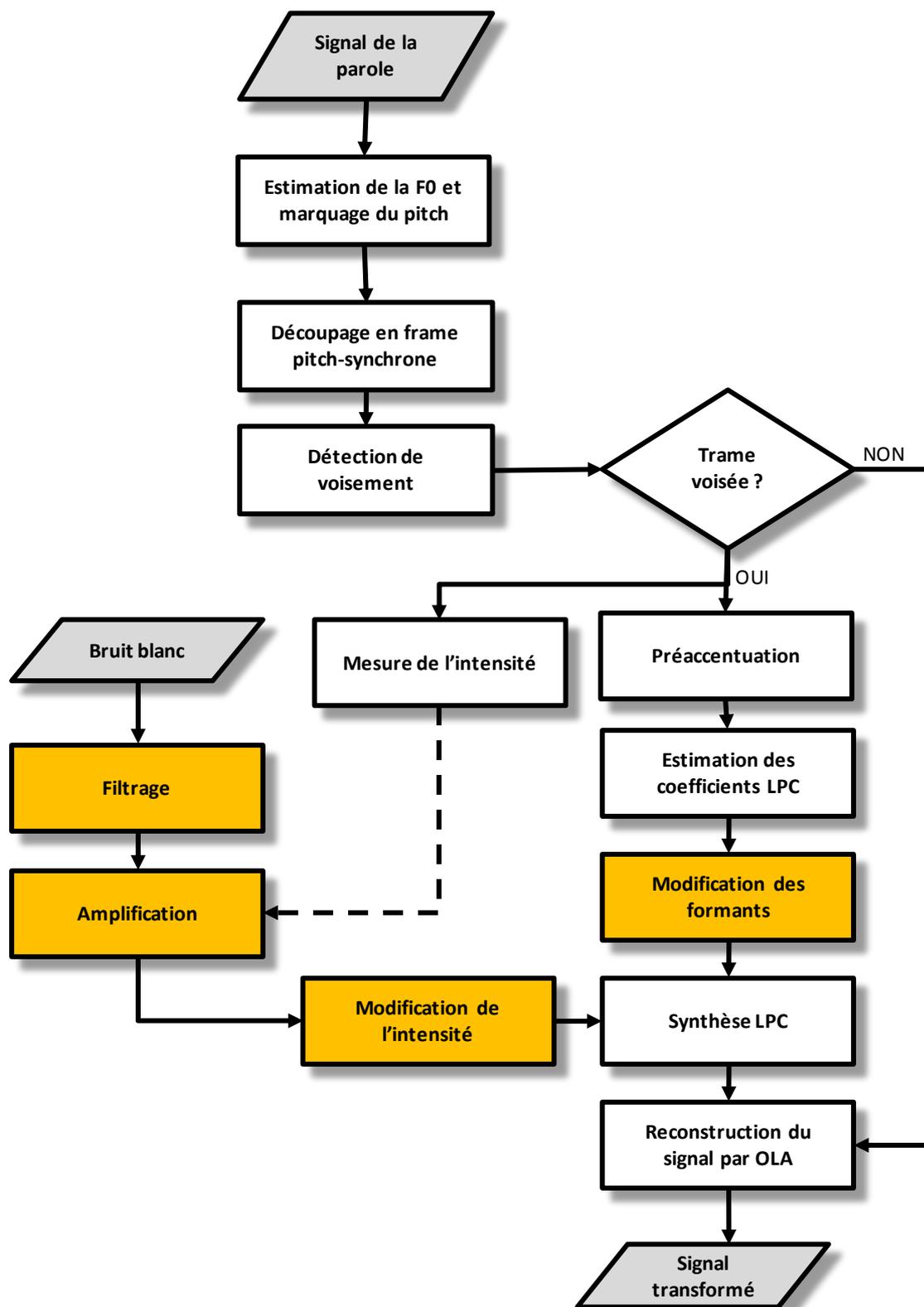


Figure 10-7 : Description de l’algorithme de transformation de voix modale en voix chuchotée

### 10.3.3 Remarques

Notons que le choix de trames à modifier ne s'effectue que par détection du voisement. Ainsi les modifications de la position des formants sont également appliquées sur les zones correspondant aux consonnes voisées. Dans l'absolu, étant donné que le déplacement des formants est lié à l'ouverture de la glotte, changeant ainsi les conditions aux limites du système, il est logique d'appliquer des modifications des formants également sur les consonnes.

Dans l'ensemble, l'algorithme proposé ci-dessus donne de bons résultats, comme nous le verrons dans le chapitre suivant. Une attention particulière doit toutefois être donnée à l'estimation de la F0 afin qu'une trame contenant du voisement ne soit pas considérée comme non-voisée. En effet cette situation provoquerait alors une reconstruction du signal à partir d'une trame contenant un voisement. De plus, une trame voisée est généralement plus énergétique qu'une trame non-voisée. Ainsi comme nous appliquons une réduction d'intensité des trames voisées de 20 dB, une mauvaise détection du voisement, aurait pour conséquence que l'intensité de ces trames ne serait pas réduite. Lors de la reconstruction du signal, celle-ci est bien trop énergétique par rapport au reste du signal modifié, ce qui a pour conséquences de créer des artefacts lors de la modification.

D'autre part, par la diminution de l'intensité des voyelles, le rapport signal sur bruit diminue. Le bruit ambiant des enregistrements ainsi que les différents bruits effectués par l'auditeur deviennent alors plus audibles. Toutefois, les bruits de bouche (bruits qui selon nous jouent un rôle dans la perception de la distance du locuteur) ressortent également dans certaines mesures, ce qui nous semble être un avantage.

Après de nombreux essais sur des voix masculines et féminines, il ressort que pour obtenir un résultat de bonne qualité, les règles de modification de la position des formants doivent être légèrement ajustées d'un locuteur à l'autre. Dans certains cas qui ne sont pas clairement identifiés, modifier la largeur de bande du filtre du bruit de source en la ramenant à 2 octaves permet, dans certaines mesures, d'améliorer la qualité et le naturel des transformations.

## 10.4 Transformation de voix modale en voix criée

---

Cette section se consacre à la description de l'algorithme de transformation de voix modale en voix criée. Comparée avec la voix chuchotée, cette transformation, qui modifie radicalement la nature de la voix, est particulièrement difficile. En effet, il est nécessaire d'appliquer des modifications sur plus de paramètres et de façon plus forte que pour les transformations en voix chuchotée. Toutefois, les variations des paramètres ainsi que leurs interdépendances sont encore mal connues. Nous avons au

début de ce travail de thèse, l'intention de transformer des phrases parlées en phrases criées. Néanmoins, les différentes analyses et notamment les relevés de la prosodie de l'effort vocal, nous ont obligés à réduire nos objectifs. En effet, nous avons pu observer entre une phrase parlée et une phrase criée, des phénomènes de réorganisation prosodique, d'où nous ne pouvons extraire facilement des lois de transformation. Ainsi, un modèle de modification prosodique dépendant fortement de la structure phonétique à modifier est nécessaire. Cette dépendance à la structure phonétique limite ainsi nos transformations aux structures de base que nous avons étudiées. Forcé de constater que la transformation de phrase est bien trop complexe à ce stade de nos recherches, nous avons alors fait le choix de réaliser des transformations uniquement pour les logatomes du corpus DB4. Ainsi, la méthode que nous proposons n'est applicable que sur les structures de base que nous avons étudiées précisément ; à savoir les structures CV, CVC, VCV et CVCV.

Nous décrivons tout d'abord dans cette section, les règles de transformation que nous avons établies pour mener à bien ces transformations, et notamment la génération du nouveau contour mélodique en fonction de la structure à transformer. Par la suite, les différentes étapes de l'algorithme sont présentées. Nous nous concentrons ici sur la modification du signal de la parole sans tenir compte des effets liés à la propagation de l'onde sonore.

#### 10.4.1 Élaboration des règles de transformation

Nous débutons cette partie destinée à établir les règles de transformation par la règle de transformation concernant F0, qui traite de la construction du contour mélodique, utilisée pour réaliser la transformation en fonction de la structure phonétique à transformer. Nous évoquons par la suite, la transformation de l'intensité et notamment la modification du rapport d'intensité entre consonnes et voyelles, ainsi que l'évolution temporelle de celle-ci. Les valeurs concernant les variations de durée appliquées ainsi que les règles relatives à la modification de la position du premier formant sont également présentées. En effet, la position du premier formant dépend beaucoup de l'effort vocal et doit être modifiée afin d'apporter, une qualité plus naturelle à la transformation. Enfin nous décrivons le filtre employé dans le but de modifier la pente spectrale.

##### 10.4.1.1 *Modification du contour mélodique (F0)*

La modification des contours de F0 est réalisée suivant les règles issues de l'analyse de la prosodie que nous avons effectuée au CHAPITRE 9. Nous créons ainsi un pattern de F0 en accord avec le phonème considéré et en fonction de sa place dans l'énoncé. Nous aurions pu choisir d'utiliser un modèle d'intonation/modèle prosodique pour générer ces courbes permettant de créer des contours de F0 plus naturels. Toutefois, pour une première approche nous considérons ce choix trop complexe. De plus, ces modèles (type Fujisaki (Fujisaki et Sudo, 1971) ou TILT (Taylor, 2000) ) ne sont pas

forcément adaptés à générer des contours mélodiques aussi variables que ceux observés pour des voix criées (tout du moins ils n'ont pas été validés sur ce point à notre connaissance). Une autre alternative bien plus évoluée consiste à se baser sur des annotations via la représentation ToBI (Silverman et al., 1992), afin de générer les contours mélodiques. A nouveau cette approche peut constituer une perspective, mais dans ce travail cette piste n'est pas étudiée. Nous avons ainsi fait le choix de décrire les contours de F0 en utilisant plusieurs points caractéristiques. Nous récréons le contour de chaque phonème séparément, puis nous les concaténons afin de créer le contour global de F0.

Dans cette section nous décrivons alors les différentes zones du pattern global en commençant par le pattern de F0 des voyelles. Nous expliquons ensuite les patterns des consonnes voisées en position initiale ou finale et en position intervocalique, avant de présenter le pattern global créé à partir de ces patterns isolés.

#### a- Pattern de F0 pour les voyelles

La première constatation fut un **focus** systématiquement présent sur les voyelles d'un mot. Ce focus a été analysé avec précision afin d'en extraire une forme générale. Sur la Figure 10-8 nous avons tracé l'ensemble des contours réalisés sur la voyelle des mots CV pour les 3 locuteurs, ainsi que la superposition de l'ensemble des ces contours (i.e. des trois locuteurs confondus). La moyenne de ces contours normalisés a ensuite été calculée pour en déduire un pattern mélodique normalisé. A partir de cette moyenne, nous considérons un pattern plus général, tracé en trait plein noir. Le pattern des voyelles ( $\Lambda_v$ ) est alors décrit suivant 8 points mentionnés dans le Tableau 10-1. Ce tableau décrit les points caractéristiques sur une échelle d'amplitude normalisée (de 0 à 1) et de temps normalisé  $tp$  (de 0 à 1).

#### Pattern de F0 pour les consonnes voisées en position initiale

Nous avons également constaté qu'une consonne voisée placée en début ou en fin d'énoncé n'était pas modifiée de la même manière que les consonnes en position intervocalique. En effet, une consonne initiale voisée subit une augmentation progressive de la F0, à partir d'une valeur relativement faible, jusqu'à atteindre le début du focus de la voyelle qui la succède. Une consonne en position finale a tendance à suivre le schéma inverse. Elle débutera à la fin du focus de la voyelle pour atteindre progressivement une valeur basse. Nous proposons alors de définir le pattern mélodique d'une consonne initiale de la manière suivante :

$$\Lambda_{c_i} = e^{\lambda \cdot tp - 1} / e(\lambda) \quad (\text{Eq. 10-6})$$

Avec  $tp=[0 ; 1/7 ; 2/7 ; 3/7 ; 4/7 ; 5/7 ; 6/7 ; 1]$  et  $\lambda$  un facteur de forme permettant d'ajuster la raideur de la courbe de F0. Dans notre cas  $\lambda$  est fixé 5 ce qui correspond approximativement à la forme observée sur nos relevés.

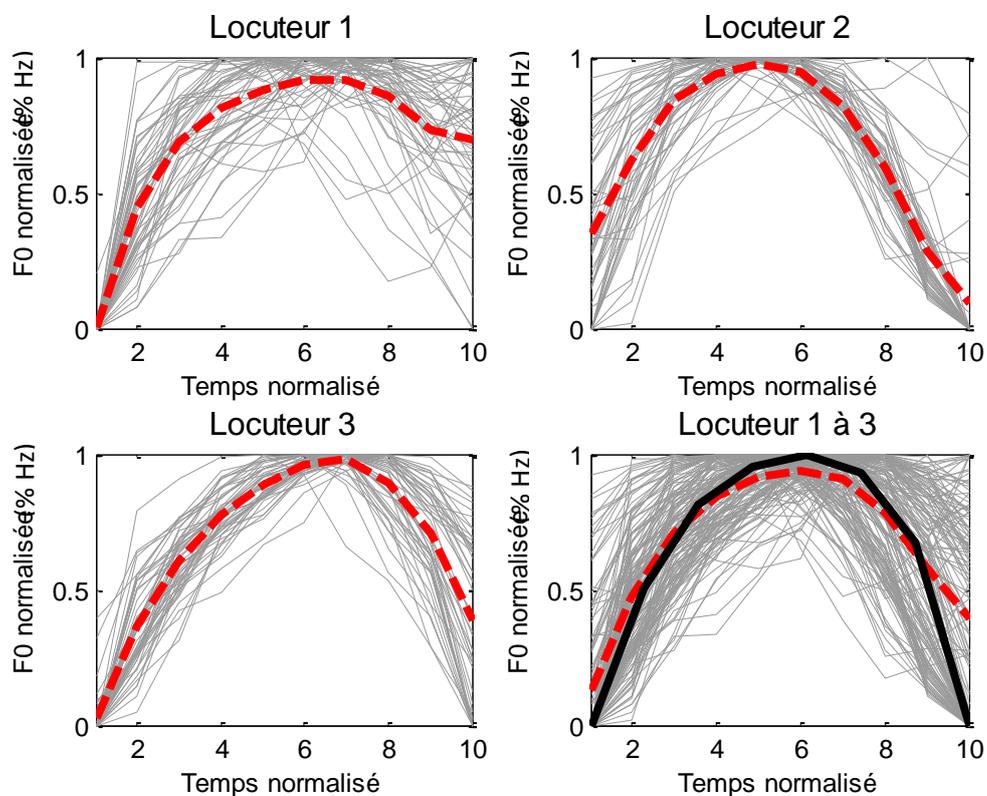


Figure 10-8: Représentation des contours de F0 normalisés en voix criée pour les voyelles des trois locuteurs. Les traits en pointillés correspondent à la moyenne. Le trait plein correspond au pattern de F0 retenu.

Tableau 10-1: Valeurs du pattern de F0 retenu pour les voyelles

<b>tp (%)</b>	0	1/7	2/7	3/7	4/7	5/7	6/7	1
<b><math>\Delta v(tp)</math></b>	0	0.50	0.81	0.95	1	0.89	0.52	0

**b- Pattern de F0 pour les consonnes voisées en position intervocalique**

Concernant les consonnes intervocaliques, leur micro-prosodie est créée à partir d'une fonction sinus décrit entre  $\pi$  et  $2\pi$  (cf. Eq. 10-7).

$$\Delta c_{inter} = \sin (\pi . tp + \pi), \quad \text{pour } 0 \leq tp \leq 1 \quad (\text{Eq. 10-7})$$

Toutefois, il est fréquent d'observer dans nos analyses, que la valeur de F0 finale de la première voyelle n'est pas identique à la valeur initiale de la seconde voyelle. Ainsi nous choisissons de décomposer le pattern des consonnes intervocaliques en deux fonctions identiques : l'une décrivant la zone de descente de F0 ( $0 < t_p \leq 4/7$ ) et l'autre, la montée de F0 ( $4/7 < t_p \leq 1$ ). Ainsi les valeurs initiales et finales peuvent être contrôlées séparément.

#### a- Pattern de F0 globales

Après avoir créé les différentes zones constituant le contour global d'un mot, chacune de ces zones est ensuite dilatée en temps et en amplitude afin de correspondre au pattern global à créer. Le pattern des voyelles se décrit alors par sa valeur initiale (et finale) ainsi que par sa valeur maximale. Le pattern des consonnes initiales (ou finales) est décrit par sa valeur de départ et sa valeur d'arrivée. Par contre le pattern des consonnes intervocaliques est contrôlé par sa valeur de début et fin ainsi que sa valeur minimale.

Notons ainsi que pour un mot CVC, la valeur finale de la première consonne est égale à la valeur initiale de la voyelle suivante. De la même manière, la valeur finale de la voyelle (identique à sa valeur initiale) est égale à la valeur initiale de la seconde consonne qui la suit. Ainsi, chaque voyelle est décrite par deux valeurs caractéristiques (initiale/finale et maximale) et les consonnes par une valeur caractéristique (valeur initiale/finale pour les consonnes initiales/finales ou valeur minimale pour les consonnes intervocaliques). Ainsi un mot CVC sera décrit par 4 points caractéristiques et un mot CVCV sera décrit par 6 points caractéristiques.

Sur la Figure 10-9, nous donnons différents exemples de pattern créés pour des mots CV, CVC, VCV ainsi que CVCV. Sur cette figure, les traits en pointillés représentent les zones correspondant aux consonnes ; consonnes qui peuvent être voisées ou non. Si la consonne du mot considéré n'est pas voisée, cela ne change en rien le pattern des voyelles. Seules les zones en pointillés sont éliminées dans le pattern global. Les ronds représentent les points caractéristiques entrés dans l'algorithme pour créer les contours (cf. Tableau 10-2). Notons que nous avons mentionné 6 points pour le cas d'un pattern mélodique pour les mots CVCV et que l'on observe 8 points sur la figure. Toutefois, la valeur initiale et finale de la voyelle étant identique dans notre modèle, cela ne constitue en soit qu'un seul point caractéristique.

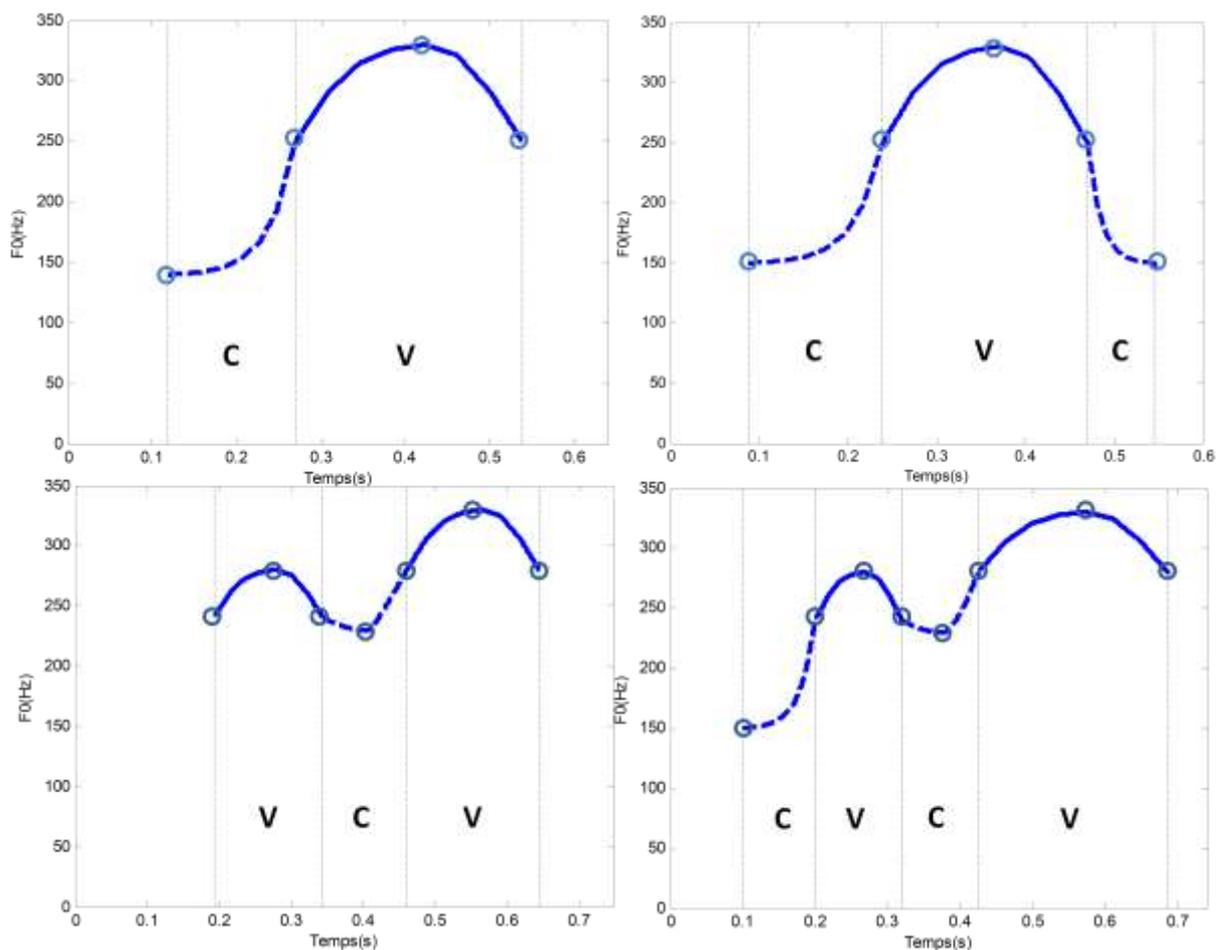


Figure 10-9: Exemples de pattern mélodique générés par l'algorithme pour des mots CV, CVC, VCV, et CVCV. Les zones en pointillée représentent les consonnes qui peuvent être voisées ou non.

Tableau 10-2: Récapitulatif des points caractéristiques pour l'élaboration du pattern de F0 pour des 4 structures de base

	CV	CVC	VCV	CVCV
Valeurs initiale de F0 de la consonne initiale	$F0_{c1-init}$	$F0_{c1-init}$		$F0_{c1-init}$
Valeurs initiale de F0 de la voyelle initiale/finale	$F0_{v1-init}$	$F0_{v1-init}$	$F0_{v1-init}$	$F0_{v1-init}$
Valeurs maximale de F0 de la voyelle	$F0_{v1-max}$	$F0_{v1-max}$	$F0_{v1-max}$	$F0_{v1-max}$
Valeurs finale de F0 de la consonne finale		$F0_{c2-final}$		
Valeurs minimale de F0 de la consonne intervocalique			$F0_{c1-min}$	$F0_{c1-min}$
Valeurs initiale de F0 de la voyelle initiale/finale			$F0_{v2-init}$	$F0_{v2-init}$
Valeurs maximale de F0 de la voyelle			$F0_{v2-max}$	$F0_{v2-max}$

#### 10.4.1.2 *Modification de l'intensité des voyelles*

Nous avons pu voir dans la littérature et par le biais de nos analyses que l'intensité des consonnes augmente moins fortement que l'intensité des voyelles en voix criée. Toutefois, nous ne considérons pas l'intensité globale de la voix criée dans nos transformations. Celle-ci pouvant être effectué *a posteriori*. Ainsi, nous avons choisi, afin d'augmenter la dynamique de l'intensité observée, d'augmenter l'intensité des voyelles. Nous avons choisi, au regard des différentes valeurs estimées d'accroître l'intensité des voyelles d'un facteur 2 (i.e. +6 dB).

Nous avons également observé pour une augmentation plus forte sur la deuxième voyelle des mots bisyllabiques. Ainsi la modification d'intensité de la première voyelle (+6 dB) est inférieure à l'augmentation de la deuxième. Dans notre algorithme de transformation nous avons alors choisi d'augmenter la deuxième voyelle d'un facteur de 2,5 (+8 dB).

#### 10.4.1.3 *Modification de la durée des voyelles*

L'ensemble des études s'accordent à dire que la durée des voyelles augmente en voix criée, et que celle des consonnes à une tendance à diminuer. Dans notre cas, nous décidons d'appliquer une augmentation sur la durée des voyelles uniquement. En effet, la diminution de durée correspond à l'élimination de certaines trames du signal. Nous pensons que dans le cas de certaines consonnes, comme pour les occlusives, cette pratique est dangereuse. En effet, les occlusives ayant une durée très courte, l'élimination d'une seule trame, et en particulier celle qui contiendrait l'explosion de la consonne, peut alors complètement éliminer la présence physique de la consonne. On pourrait imaginer de réaliser la distinction entre les différentes familles de consonnes au sein de l'algorithme pour éviter ce type de problème. Toutefois ce type d'algorithme n'est pas évident à mettre en œuvre. D'un point de vue pratique il est plus facile de réaliser de façon automatique un classement consonne-voyelle. C'est en partie pourquoi nous ne faisons aucune modification temporelle sur les consonnes mais reportons l'ensemble des modifications sur les voyelles. Ainsi d'un point de vue relatif une augmentation de la durée des voyelles correspond à une diminution de la durée des consonnes. De plus la tendance à la diminution de la durée des consonnes, contrairement à celle des voyelles, n'est pas partagée par les 3 locuteurs que nous avons considérés dans le corpus DB4. Ainsi, nous choisissons, comme règle de transformation d'augmenter la durée des voyelles de 40%.

Cette modification est appliquée à l'ensemble des voyelles même si certaines variations ont pu être observées et notamment le fait que la dernière voyelle s'allonge plus que la première. Néanmoins aucune tendance de l'ampleur de celle observée dans (Rostolland, 1982a) sur des mots français dissyllabique, concernant l'augmentation de la deuxième voyelle par rapport à la première n'a été observée (i.e. +33% pour les voyelles et +67% pour la voyelles finale).

#### 10.4.1.4 *Modification de la position du premier formant*

Nous avons pu voir dans les chapitres précédents, que la position des formants et notamment la position du premier formant varie pour des voix criées. Malgré le nombre important d'études sur le sujet nous n'avons pu trouver aucune règle décrivant le passage des formants de la voix parlée vers ceux de la voix criée. Le manque de modélisation dans la littérature, par rapport à la voix chuchotée, ainsi que les incohérences quant aux ordres de grandeur des déplacements présents dans la littérature, nous contraint à faire un choix parmi les différentes théories quant au déplacement du premier formant.

La théorie du « *Formant Tuning* » suggère que dans un souci d'efficacité vocale, le locuteur (en particulier dans le cas du chant) place volontairement le premier formant à l'aplomb d'une harmonique de la F0 (Garnier et al., 2010; Joliveau et al., 2004a, 2004b; Sundberg, 1977). Bien que cette théorie prête encore à controverse, il n'est pas exclu que ce type d'ajustement intervienne lors d'un cri dans le but de maximiser l'intensité vocale. Il existe également une théorie qui mentionne que la distance tonotopique entre F0 et F1 est constante avec l'augmentation de l'effort vocal (Traunmüller, 1985). Toutefois, cette théorie n'est pas encore validée pour le cri dans le but d'une communication à distance et qui d'autre part, comme nous l'avons mentionné ne correspond pas à nos relevées (cf. 7.1.2). La théorie du *formant tuning* étant probable et nos analyses des formants permettent également de se rapprocher de ce constat nous avons choisi de retenir cette dernière.

De ce fait nous avons défini une règle de modification du premier formant en accord avec cette théorie. Nous déplaçons ainsi le premier formant sur l'harmonique supérieure la plus proche. Pour tous F0 au sein de la voyelle la position du premier formant est alors égale à

$$F_1(t) = k.F0(t) \quad (\text{Eq. 10-8})$$

où k est un entier caractérisant le numéro de l'harmonique supérieure la plus proche.

Dans cette modification il s'agit de prendre soin de ne pas créer de saut de formant lié à une mauvaise détection de la position d'un formant. Ainsi la règle de transformation est différente pour le /a/ et pour le /i/ et le /u/. Nous définissons pour le /a/ le premier formant comme étant systématiquement placé sur la 3ème harmonique<sup>1</sup> de F0, et le premier formant de /i/ et /u/ comme étant placé sur la fondamentale (i.e. L'harmonique de rang 1).

Notons qu'il a été rapporté que le fait d'aligner le formant avec la F0 est une source d'interaction forte entre la source et le filtre (Titze, 2008). Toutefois le manque de compréhension de ces phénomènes ne

<sup>1</sup> Nous utilisons ici la définition la plus répandue qui considère la fréquence fondamentale comme étant l'harmonique de rang 1. Toutefois il existe aussi la définition qui considère l'harmonique de rang 1 comme étant un multiple entier (différent de 0 et 1) de la fréquence fondamentale.

nous permet pas de tenir compte de ces interactions. Nous n'avons toutefois pas fait le choix de tenir compte des phénomènes d'interaction.

#### 10.4.1.5 Modification de la pente spectrale

La pente spectrale des voyelles est moins importante en voix criée qu'en voix parlée (Liénard et Di Benedetto, 1999). Pour recréer cette caractéristique de la voix criée directement liée à la source glottique (que nous ne modifions pas directement ici) nous avons choisi d'utiliser un filtre ayant une pente moyenne de +3 dB/octave. Ce filtre est décrit par l'équation ci-dessous (Eq.10-9) et sa réponse fréquentielle est représentée sur la Figure 10-10 ( $F_s=16$  kHz). Chaque trame correspondant à une voyelle est alors filtrée par ce filtre permettant ainsi d'augmenter la pente spectrale de 3 dB/octave.

$$G(z) = \frac{67,4309 - 121,443 \cdot z^{-1} + 54,12 \cdot z^{-2}}{1 - 4,2522 \cdot z^{-1} + 3,3432 \cdot z^{-2}} \quad (\text{Eq. 10-9})$$

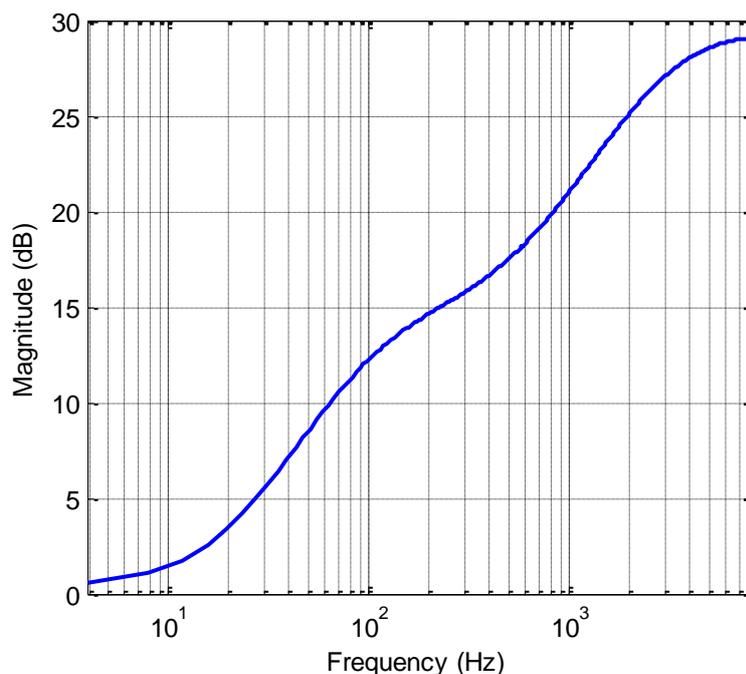


Figure 10-10: Réponse en fréquence du filtre utilisé pour rehausser la pente spectrale

#### 10.4.2 Validité des règles de transformation proposées

Dans une étude préliminaire au travail de caractérisation de la prosodie des logatomes présentée dans le CHAPITRE 9, et de l'élaboration des règles présentées dans ce chapitre, nous avons cherché à valider ces règles de transformation et tout particulièrement la transformation du contour de F0 (cf.(Fux et al., 2012b)).

Dans cette étude nous nous sommes concentrés sur des logatomes monosyllabiques CVC. Ces logatomes sont listés dans le Tableau 10-3. Ils ont été créés à partir de la voyelle /a/ uniquement, ainsi qu'à partir de consonnes (une de chaque nature), en considérant toutes les combinaisons possibles à partir de ces phonèmes : ce qui correspond à un total de 36 mots. Trois locuteurs ont participé à l'enregistrement de ces mots en voix parlée et criée suivant le protocole de simulation pour une communication de 60 m que nous avons développé (cf. CHAPITRE 6). L'analyse de ces enregistrements (108 logatomes criés, 108 logatomes parlés) nous a permis d'observer des contours de F0 similaires à ceux décrits précédemment.

A partir de ces contours nous avons réalisé des transformations de voix parlée en voix criée. Les transformations de logatomes parlés en logatomes criés ont été réalisées en utilisant la technique LP-PSOLA. Une règle de transformation du contour de F0 a été développée, qui est similaire à celle décrite précédemment, en utilisant un patron décrit par plusieurs points reflétant la moyenne des analyses. En outre, des modifications supplémentaires ont été intégrées : une diminution de la pente spectrale, une augmentation de la dynamique de l'intensité, un décalage de la fréquence du premier formant et un allongement de la durée de la voyelle. Ces modifications sont similaires à celles présentées dans ce chapitre mais dans des proportions différentes ou par des méthodes différentes.

Tableau 10-3: Listes des logatomes utilisés pour la validation des règles de transformation de voix parlée en voix criée

C1 \ C2	Occlusives voisées	Occlusives non voisées	Fricatives voisées	Fricatives non voisées	Nasales	Liquides
Occlusives voisées	/ba d/	/ba p/	/ba ʒ/	/ba s/	/ba m/	/ba l/
Occlusives non voisées	/pa d/	/pa p/	/pa ʒ/	/pa s/	/pa m/	/pa l/
Fricatives voisées	/ʒa d/	/ʒa p/	/ʒa ʒ/	/ʒa s/	/ʒa m/	/ʒa l/
Fricatives non voisées	/fa d/	/fa p/	/fa ʒ/	/fa s/	/fa m/	/fa l/
Nasales	/ma d/	/ma p/	/ma ʒ/	/ma s/	/ma m/	/ma l/
Liquides	/la d/	/la p/	/la ʒ/	/la s/	/la m/	/la l/

Dans le but de valider les règles de transformation de la dynamique de F0, deux types de transformations ont été créées. La première transformation correspond à la modification des paramètres de logatomes parlés en logatomes criés, décrits par l'ensemble des règles de transformation que nous venons de mentionner. Cette transformation est appelée « *transformation  $\Delta F_0$*  ». La seconde transformation reprend les mêmes règles que la transformation  $\Delta F_0$ , sauf pour les modifications

concernant F0. En effet, dans cette deuxième transformation la F0 utilisée pour la transformation, correspond à la F0 de la voix parlée multipliée par un facteur permettant d'augmenter le contour de F0. Ce facteur est défini comme étant le facteur multiplicateur du contour de F0 permettant d'obtenir la même valeur moyenne de F0 que la transformation  $\Delta F0$ . Cette transformation est appelée « *transformation xF0* ». Ainsi, entre les deux séries de transformations, et pour un mot donné, la seule différence est le contour de F0. La modification apportée à la voix parlée et la valeur moyenne de F0 sont similaires. Cette seconde transformation a été utilisée pour étudier l'intérêt des règles de transformation de F0 dans la perception de l'effort vocal. Ceci permet également de valider les règles de transformation proposées. Notons que l'ensemble des logatomes enregistrés ont été transformés par les deux méthodes décrites ( $\Delta F0$  et xF0).

Les voix modifiées par les deux transformations ont, par la suite, servi à réaliser un test perceptif. Ce test perceptif se base sur une tâche de classification de voix. Dix-huit sujets ont participé à ce test. Chaque sujet a d'abord entendu la voix parlée naturelle, puis l'une des deux voix transformées. Il devait alors classer la deuxième voix entendue (i.e. l'une des deux voix transformées) comme étant une « voix criée », une « voix simplement aigüe » ou encore « aucune des deux propositions ». La tâche demandée aux sujets correspondait alors à une tâche d'identification de la présence d'effort vocal. Dans le même temps, le sujet devait évaluer la qualité de la transformation à l'aide de l'échelle normalisée MOS (*Mean Opinion Score* : i.e. 5: Excellente; 4: Très bon; 3: Bonne; 2: Mauvais; 1: Très mauvais). Chaque sujet a réalisé 72 classifications (36 pour les transformations xF0 et 36 pour les transformations  $\Delta F0$ ).

Les résultats du test de perception sont tracés dans la Figure 10-11. Les transformations  $\Delta F0$  (n = 648) (barre de gauche) sont perçues comme des voix criées dans 83 % des cas, et les transformations xF0 (n = 648) sont perçus comme aigües dans 78 % des cas. Les transformations xF0 ne sont toutefois perçues comme voix criées que dans 13 % des cas. Notons que les résultats obtenus pour les 3 locuteurs qui composent ce test sont similaires. Ces résultats montrent également une bonne qualité de modification quant aux transformations  $\Delta F0$  (MOS = 3,9).

Les règles de transformation proposées pour créer les transformations  $\Delta F0$ , donnent de bons résultats en termes de qualité de transformation (voir le Tableau 10-4). On remarque également que les transformations  $\Delta F0$  qui ne sont pas classées en voix criée présentent des qualités de transformation moins élevées et possèdent des erreurs relatives élevées, ce qui pourrait alors traduire des transformations « manquées ».

Notons également que les transformations de voix criée  $\Delta F0$  classées comme voix aigües, ne concernent pas toujours les mêmes mots, à l'exception de certains d'entre eux. Cinq mots du locuteur 1 ont été classés en voix aigües par au moins 3 sujets (/b a s/, /f a l/, /b a p/, /m a s/ et /l a p/). Pour le

locuteur 2, seulement 2 mots ont été classés en voix aigüe par au moins 3 sujets (/la m/ et /f a s/) et 3 mots pour le locuteur 3 (/b a p/, /m a ʒ/ et /la ʒ/). Ainsi, aucune tendance particulière parmi les mots « mal classés » n’a pu être observée.

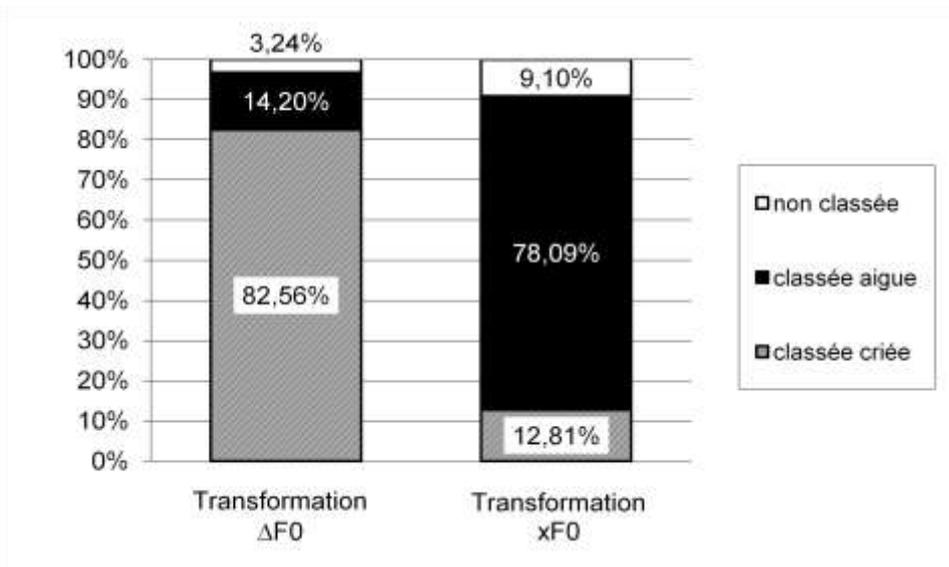


Figure 10-11: Résultats des tests perceptifs consistant à comparer deux types de transformations.

Tableau 10-4: Score d'opinion moyen (MOS)

	Transformation DELTA_F0	Transformation xF0
	MOS (σ)	MOS (σ)
<b>Classée criée</b>	3,91 (0,52)	3,56 (0,71)
<b>Classée aigüe</b>	3,59 (0,66)	3,59 (0,97)
<b>Non classée</b>	2,65 (0,98)	3,65 (0,84)

Cette première étude confirme que le contour de F0 joue un rôle important dans la perception de l’effort vocal. Ce constat est en parfait accord avec les études mentionnant l’importance de la prosodie dans la perception de l’effort vocal (Fux et al., 2010; Tassa et Liénard, 2000) et notamment le rôle prépondérant de la F0 (Brungart et al., 2002). D’autre part, on observe qu’une simple multiplication du contour de F0, comme pour les transformations xF0, produit des voix qui ne reflètent pas complètement l’effort vocal. Bien que cette étude ne concerne que les mots monosyllabiques CVC, les résultats sont très prometteurs et mettent en avant l’importance de notre modélisation du contour de F0.

### 10.4.3 Réalisation des transformations de voix parlée en voix criée

Nous venons de décrire les différentes règles que nous avons élaborées et qui constituent la base de notre algorithme de transformation. L'algorithme de transformation de voix modale en voix criée se décompose en différentes étapes décrites ci-dessous. Une représentation, sous forme de bloc fonctionnel, est donnée en Figure 10-12.

#### Étape 1. Analyse

- a- La première étape de l'algorithme est l'estimation de la F0 en utilisant soit le signal de la parole soit le signal EGG (s'il est disponible). Par la suite le marquage du pitch est réalisé.
- b- A partir des positions des marqueurs définies dans l'étape précédente, le signal  $s(t)$  échantillonné à 16 kHz est ensuite **décomposé** en trames  $s_i(t)$  centré sur un marqueur de pitch ( $t_i$ ) et allant de  $t_{i-1}$  à  $t_{i+1}$ . Ceci correspond à un recouvrement proche de 50%. Par la suite, pour chacune de ses trames, les étapes suivantes sont réalisées.
- c- Avant de réaliser l'estimation des coefficients du prédicteur linéaire, la trame considérée est préaccentuée par un filtre de **préaccentuation** ( $V(z)=1-0,95z^{-1}$ ). Les coefficients du prédicteur sont ensuite estimés.
- d- L'**estimation des coefficients LP** est réalisée à partir du signal  $s_i(t)$  pondérée par une fenêtre de *Hamming* en utilisant une méthode de prédiction linéaire (Markel et Gray, 1976) ainsi que l'algorithme de résolution de Durbin-Levinson (Durbin, 1960). L'ordre  $p$  du filtre AR est fixé à 18. (i.e.  $p = Fs/1000 + 2$ )
- e- Un filtrage inverse de chaque trame permet ensuite de disposer du résidu de chaque trame, qui est alors modifié en fonction des règles de transformation liées à la F0, à l'intensité, à la durée et à la pente spectrale.

#### Étape 2. Élaboration des contours et modifications

- a- A partir des coefficients du filtre LP estimés, et en fonction de la position de la trame dans le mot (i.e. s'il s'agit d'une trame appartenant à une voyelle), la position du premier formant est modifiée en accord avec la valeur de F0.
- b- Le résidu du signal (ayant théoriquement une pente spectrale de -6 dB/octave<sup>1</sup>) est filtré par le filtre de modification spectrale précédemment décrit, afin de rehausser sa pente spectrale de 3 dB/octave (soit un résidu modifié de pente spectrale, équivalent à -3 dB/octave)

<sup>1</sup> La pente de -6 dB/octave théorique du résidu est vraie dans le cas où un filtrage inverse est réalisé directement à partir du signal de la parole sans prendre en considération la radiation aux lèvres qui est encore présente dans celui-ci

- c- Grâce à l'étiquetage de chaque mot effectué manuellement, le pattern global de F0 est créé en utilisant les différents points caractéristiques que nous avons mentionnés précédemment.
- d- Dans un même temps, le pattern d'intensité est créé, ce qui permet d'augmenter l'intensité des voyelles lors de la synthèse,
- e- Enfin, le pattern temporel qui permet de modifier la durée des voyelles est créé.

### **Étape 3 : synthèse**

- a- En utilisant, les modifications de durée et de la F0 définies au préalable, les marqueurs de pitch sont repositionnés.
- b- La dernière étape consiste à reconstruire le signal en tenant compte des filtres LP modifiés, des résidus modifiés ainsi que du pattern de l'intensité et les nouveaux marqueurs de pitch. Chaque trame est donc reconstruite par filtrage du résidu modifié et ré-échantillonnée pour correspondre aux nouvelles marques du pitch par le filtre LP modifié. Ces trames sont alors ajustées en intensité avant d'être reconstruites pour créer le signal final suivant la technique OLA. Notons également que le rayonnement aux lèvres n'est pas appliqué ici, car il est modélisé dans le résidu.

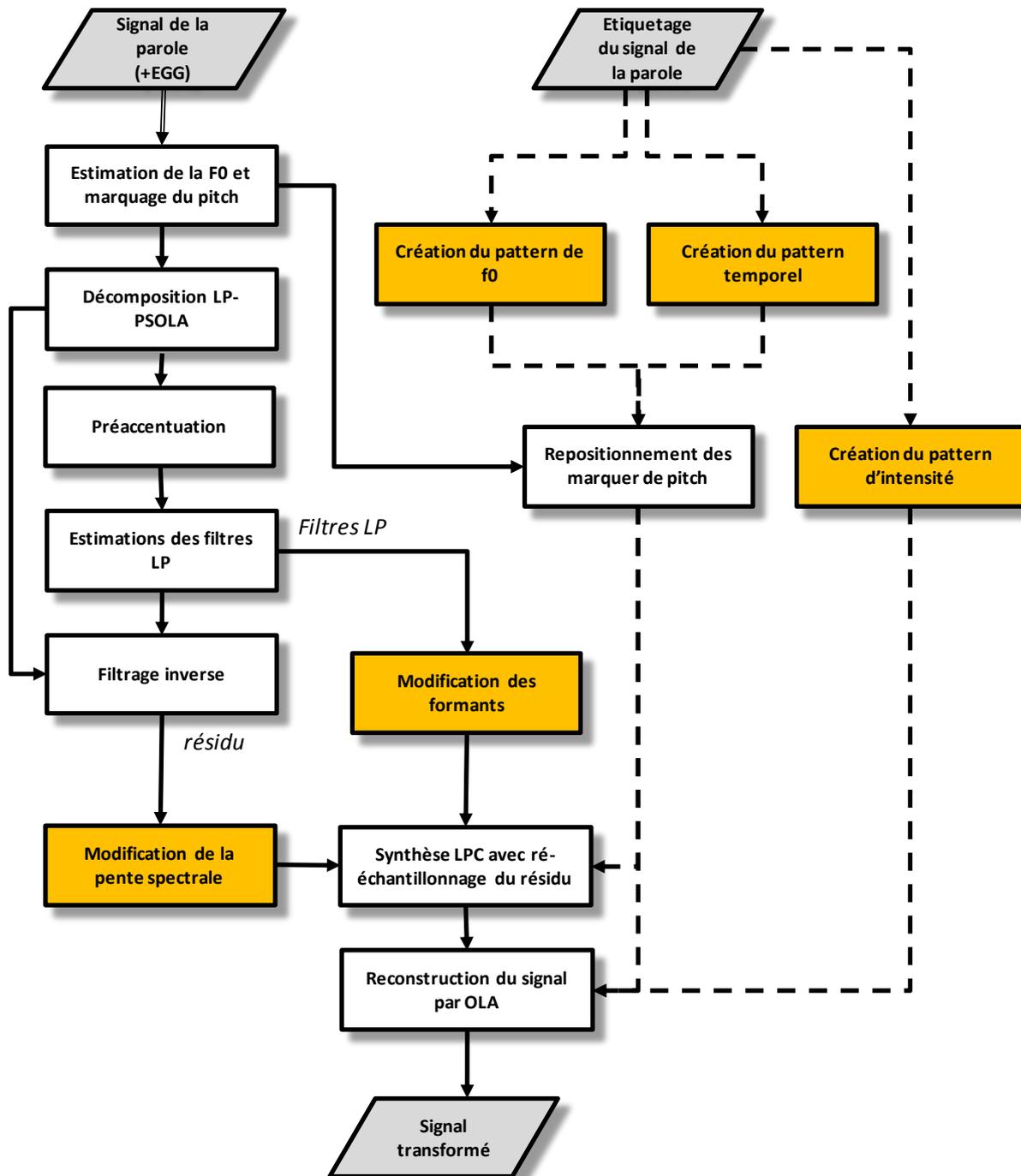


Figure 10-12 : Description de l'algorithme de transformation de voix modale en voix criée

#### 10.4.4 Réflexions quant à la transformation de plusieurs niveaux d'effort vocal :

Nous ne décrivons ici qu'un seul niveau de transformation en voix criée. Toutefois, il est envisageable de créer un continuum de variation des paramètres permettant de nuancer l'effort vocal de la voix transformée. Les valeurs extrêmes de modifications étant celle présentée ici. Toutefois, notre étude ne permet pas de valider ce point. En effet, pour ce faire, une base de données, comprenant plusieurs niveaux d'efforts vocaux, est nécessaire. Or, le corpus sur lequel nous nous sommes basé (DB4) pour

réaliser ces transformations, ne comprend qu'un seul niveau d'effort. Néanmoins, la règle principale de modifications que nous proposons est la modélisation du contour de F0. Cette modification de F0 se traduit notamment par une augmentation de la dynamique de F0. Nous avons vu dans le CHAPITRE 8, que la dynamique de F0 est fonction de la distance de communication. Le même constat peut être fait concernant la dynamique de l'intensité. Ainsi, tout en conservant les règles de construction de F0 que nous établissons, différents niveaux d'effort peuvent être obtenus en modifiant principalement le contour de F0 de telle façon que la valeur moyenne ainsi que la dynamique, augmente progressivement.

## 10.5 Conclusion du chapitre 10

---

Dans ce chapitre, nous avons décrit les différentes règles que nous proposons pour réaliser les transformations de voix parlée en voix chuchotée ou criée. Nous avons alors détaillé deux algorithmes de transformation.

Le premier de ces algorithmes s'attèle à la transformation de voix modale en voix chuchotée, via une modification pitch-synchrone. Cet algorithme a été conçu de telle façon qu'il puisse être implémenté dans un système temps réel (i.e. avec un retard de deux trames d'analyses qui reste acceptable pour la communication). La particularité de ce dernier est la prise en compte des différences d'intensité entre partie voisée et non-voisée, systématiquement observées sur les voix chuchotées. D'autre part, cet algorithme, contrairement aux méthodes de transformation classiques, considère la source de la voix chuchotée comme étant un bruit de constriction plutôt qu'un bruit blanc. D'autre part, un modèle permettant de prédire la position des formants en voix chuchotée, à partir de ceux relevés en voix modale (inspiré de (Morris, 2003)) est utilisé.

Le deuxième algorithme est, quant à lui destiné à transformer une voix modale en voix criée. Cet algorithme n'est aujourd'hui applicable que sur des structures de base de type CV, CVC, VCV et CVCV. En effet, la réorganisation prosodique observée pour ces structures entre une voix modale et une voix criée, nous oblige à construire des modèles basés sur des annotations réalisées au préalable sur les voix à transformer (identification des consonnes et voyelles). Toutefois, nous pensons que le fonctionnement d'un tel algorithme sur des structures de base constitue un grand pas en avant pour la transformation de l'effort vocal. Les phrases complexes ne sont-elles pas formées à partir de structures de base ? Cet algorithme unique en son genre, basé sur des méthodes de modification classiques tient compte de l'aspect prosodique des voix criées. En effet, nous nous intéressons particulièrement à modifier la structure prosodique des voix modales, afin d'apporter une sensation d'effort vocal car ceux-ci semblent être les plus pertinents (Brungart et al., 2002; Fux et al., 2010; Tassa et Liénard,

2000). Ainsi des modèles de pattern de F0 sont créés, des modifications du rapport d'intensité en consonne et voyelles, des modifications de durée des voyelles, un déplacement de formant (qui reste encore peut être à affiner) ainsi qu'une modification de la pente spectrale.

Nos règles de transformation se concentrent principalement sur les paramètres les plus pertinents, permettant de transformer une voix parlée en une voix criée. Ainsi, nous n'avons pas mis en place des règles de transformation pour l'ensemble des variations et paramètres mentionnés dans la littérature. En particulier, l'ensemble des paramètres liés à l'onde de débit glottique ( $O_q$ ,  $\alpha_m$ ,  $Q_a$ ,  $E$ ), ne sont pas directement modifiés ; la variation de la pente spectrale tend toutefois à recréer certaines des variations que la source glottique engendre. Notons que le fait de ne pas considérer ces paramètres ne nie pas pour autant leurs importances. Bien au contraire, l'onde de débit glottique constitue un élément précieux pour la perception de la voix et notamment pour la qualité de voix. Toutefois, nous nous sommes concentrés ici sur les éléments les plus pertinents issus de la littérature (à savoir la prosodie et principalement la variation de F0).

Nous avons toutefois conscience de l'aspect expérimental de cet algorithme. En effet, avec le recul plusieurs perspectives sont envisageables, et notamment l'utilisation de décomposition de partie harmonique et de partie stochastique. Ceci permettrait d'améliorer encore la qualité de la synthèse. L'utilisation d'un modèle de source permettrait de modifier directement les paramètres de source, tels que le quotient ouvert, le coefficient d'asymétrie ou le quotient de la phase de retour qui sont différents en voix criée et qui ont une influence directe sur le spectre de la voix.

Le chapitre suivant se consacre naturellement à l'évaluation de ces deux algorithmes, aussi bien en termes d'intelligibilité des transformations qu'en termes de perception de distance.

# Évaluation des transformations

---

*“However beautiful the strategy, you should occasionally look at the results.”*

— Sir William Churchill

Jusqu’à présent nous avons cherché à déterminer les éléments pertinents des voix chuchotées et criées afin de réaliser des transformations de voix parlée vers une voix chuchotée ou criée permettant de refléter la distance d’un locuteur.

A présent l’étape cruciale consiste à évaluer les transformations que nous avons proposées. Pour ce faire, nous avons choisi de valider nos transformations en 2 étapes successives. La première consiste à vérifier si l’intelligibilité de la voix est conservée après les transformations (par rapport à l’intelligibilité des voix criées et chuchotées naturelles). Il s’agit là de valider un point important lors de toutes modifications apportées à un signal de la parole à savoir l’intelligibilité. En effet, une modification de voix qui engendre une perte non négligeable d’intelligibilité n’a que peu d’intérêt d’un point de vue applicatif. Toutefois, même si l’intelligibilité est conservée, cela ne signifie pas pour autant que les transformations de voix apportent l’effet perceptif escompté ; à savoir une sensation de distance du locuteur. De la même manière, la sensation de distance peut être apportée via la voix malgré une mauvaise intelligibilité. C’est pour ces deux raisons que nous avons réalisé deux tests d’évaluation.

Ainsi dans un premier temps, nous évaluons l’intelligibilité des voix parlée, criée, et chuchotée ainsi que des voix transformées chuchotée et criée. Dans un second temps, nous nous intéressons à la perception de distance évoquée par les transformations en voix criée. Signalons que nous ne testons pas la perception de distance à partir des transformations en voix chuchotée qui sont, de toute évidence, perçues comme très proches, étant donné l’absence de voisement.

## 11.1 Tests d'intelligibilité

Il est primordial de valider l'intelligibilité des voix transformées car une transformation de voix doit avant tout rester intelligible. Nous décrirons tout d'abord le type de voix utilisé au cours de ces tests puis nous décrivons la méthodologie employée pour mener à bien ces tests avant de présenter les résultats obtenus.

### 11.1.1 Caractéristiques des signaux tests

Avant de décrire les tests, nous rappelons les différentes règles de que nous avons mises en place pour réaliser les signaux tests, dans le but de réaliser les tests d'intelligibilité. Nous détaillons et quantifions ainsi les différents paramètres précédemment évoqués tels que l'ajustement du contour de F0, le déplacement des formants, l'ajustement d'intensité et de durée.

#### 11.1.1.1 Règles utilisées pour les transformations en voix chuchotée

Dans le chapitre précédent, nous avons détaillé notre algorithme de transformation de voix modale en voix chuchotée. Nous avons mentionné que l'ajustement de certaines règles en fonction du locuteur semble permettre d'améliorer la qualité des transformations. C'est pourquoi, nous précisons ici les règles retenues. Ainsi la première règle concerne la variation **d'intensité** entre phonèmes voisés et non voisés. Nous avons pour ce test retenu un facteur de 10 entre trames voisées et non voisées (Ito et al., 2005). Ainsi, pour chaque trame reconnue comme voisée l'amplitude de celle-ci est divisée par 10 (soit -20 dB). Concernant la génération du **bruit de source** nous avons retenu le filtre présenté précédemment : à savoir un filtre de *Butterworth*, de fréquences de coupure  $F_{c1}=424$  Hz et  $F_{c2}= 3394$  Hz, qui correspond environ à une largeur de bande de 3 octaves centrée sur 1200 Hz. Pour le déplacement des **formants** nous retenons également la règle décrite au chapitre précédent ; à savoir un déplacement maximum de 227 Hz pour une fréquence de 560 Hz (cf. Figure 10-6)

#### 11.1.1.2 Règles utilisées pour la transformation en voix criée

Nous rappelons ici les différentes règles élaborées pour la transformation en voix criée. Le rapport **d'intensité** entre consonnes et voyelles a été diminué de 2 (soit -6 dB). Soit une augmentation des voyelles de 6 dB. Les bornes retenues pour la génération du **contour de F0** sont les suivantes,  $F0_{Cl-init}= 140$ ,  $F0_{V1-init} = 250$  Hz et  $F0_{V1-max}=330$  Hz  $F0_{V1-init} = 250$  Hz. Ces valeurs correspondent approximativement à la moyenne des observations pour les 3 locuteurs. La position du premier **formant** a été déplacée suivant la règle mentionnée précédemment ; à savoir que le premier formant

est placé sur l'une des harmoniques de F0 (toujours la même sur l'ensemble de la voyelle). La **durée** de la voyelle, quant à elle, a été allongée de 40%.

### 11.1.2 Description du protocole de test d'intelligibilité

Afin d'éviter d'éventuelles facilités à la compréhension par la présence de redondances linguistiques nous avons choisi les mots CV du corpus DB4. En effet, les autres structures phonétiques contenues dans DB4 (CVC, VCV et CVCV) disposent de deux voyelles ou deux consonnes identiques ce qui permet de comprendre l'intégralité du mot sans, pour autant, que chaque phonème soit intelligible.

Les tests d'intelligibilité ont été effectués sur l'ensemble des mots CV du corpus DB4. Nous disposons donc de 153 mots (51 mots \* 3 locuteurs) pour réaliser les tests. Nous testons ainsi l'intelligibilité des voix parlées, des voix criées, des voix chuchotées naturelles (enregistrées dans l'unique but de réaliser des tests perceptifs) ainsi que des transformations chuchotées et criées. Les tests sur les voix naturelles permettront d'obtenir une référence directe en fonction de la nature de la voix (parlée, criée et chuchotée).

Devant le nombre conséquent de mots sur l'ensemble de notre base de données (51 mots CV, 3 locuteurs, 5 types de voix) un test unique sur l'ensemble des échantillons serait difficilement acceptable par les sujets. En effet, un tel test représenterait 765 mots. Ainsi les tests d'intelligibilité ont été divisés pour chaque locuteur en 4 séries de 10 mots et 1 série de 11 mots (cf. Tableau 11-1) ; ce qui pour l'ensemble des échantillons représente 15 séries (5 séries \* 3 locuteurs). Chaque série contient les 10 mots (ou 11 mots) dans les cinq modes de phonation (parlée, criée, chuchotée, transformation chuchotée et criée). Ainsi chacune des séries est constituée de 50 échantillons (ou 55 échantillons) mariant aléatoirement consonnes occlusives, fricatives, nasales et liquides.

Tableau 11-1: Mots contenus dans chaque série

<b>Série 1</b>	<b>Série 2</b>	<b>Série 3</b>	<b>Série 4</b>	<b>Série 5</b>
/b u/	/t i/	/t a/	/p u/	/ʃ u/
/m i/	/v u/	/d i/	/r i/	/v i/
/d a/	/s u/	/z i/	/p i/	/g i/
/z a/	/m u/	/g u/	/b i/	/ʃ a/
/n u/	/s a/	/n a/	/n a/	/f i/
/f a/	/k i/	/l u/	/ʃ i/	/f u/
/n i/	/d u/	/r a/	/m a/	/g a/
/k u/	/t u/	/z u/	/ʒ a/	/r u/
/p a/	/s i/	/l a/	/n u/	/ʒ i/
/k a/	/l i/	/v a/	/ʒ u/	/b a/
/	/	/	/	/n i/

Afin d'obtenir un nombre représentatif de tests, chaque sujet a réalisé 3 tests (un pour chaque locuteur). Les trois séries présentées sont systématiquement différentes pour un sujet donné. Ainsi, chaque sujet a entendu un minimum de 150 mots (155 mots si le sujet a entendu la série n°5).

Les échantillons sonores de chaque série ont été présentés à l'aide d'un ordinateur et d'un casque audio. Afin que les sujets donnent leurs réponses, un programme a été développé permettant ainsi de jouer chaque mot de la série et de permettre aux sujets taper les mots entendus directement sur le clavier de l'ordinateur. Ce programme fait écouter aux sujets les mots de la série dans un ordre aléatoirement. Les mots étaient précédés par la phrase type : « *Écrivez le mot* : ». Après l'écoute d'un mot, les sujets avaient tout le temps nécessaire pour écrire le mot avant de valider pour entendre le mot suivant. Toutefois, ils ne pouvaient écouter qu'une seule fois chaque mot. L'ensemble des mots a été égalisés en intensité. L'unique consigne donnée aux sujets était que l'ensemble des mots étaient de type CV dans des modes de phonation différents. Nous avons également fourni aux sujets une liste de l'ensemble des consonnes ainsi que l'ensemble des voyelles (hors semi-voyelle) pour faciliter l'écriture des mots. En effet, la majorité des sujets qui ont participé à ce test n'avait pas connaissance de l'alphabet phonétique et devaient alors écrire des mots qui n'existent pas en phonétique (ex : /ɲ u/ est écrit « *niou* » ou « *gnou* »).

### 11.1.3 Résultats des tests d'intelligibilité

Cette section présente les résultats du test d'intelligibilité et se décompose en 3 sous-parties. La première est consacrée aux résultats globaux. Elle décrit brièvement le taux d'intelligibilité globale des mots. La deuxième partie est consacrée à l'intelligibilité des voyelles et la troisième partie à celle des consonnes. Dans cette section les différents graphiques sont tracés à partir de l'ensemble des résultats collectés (tous locuteurs et séries confondus). Les détails concernant l'intelligibilité en fonction des locuteurs sont donnés en Annexe A. Les résultats présentés dans cette section correspondent aux résultats obtenus par **25 sujets** ayant passé chacun 3 séries différentes issues de locuteur différents. L'ensemble des résultats collectés correspond donc à 3825 réponses. Ce qui signifie que chaque mot, pour un locuteur donné, et un mode de phonation donné (parlée, criée ou chuchotée), a été entendu 5 fois.

#### 11.1.3.1 *Intelligibilité globale*

Les premiers résultats concernent le taux de compréhension des mots dans leurs globalités. La Figure 11-1 représente ces taux d'intelligibilité total pour les 3 modes de phonation ainsi que pour nos deux types de transformation.

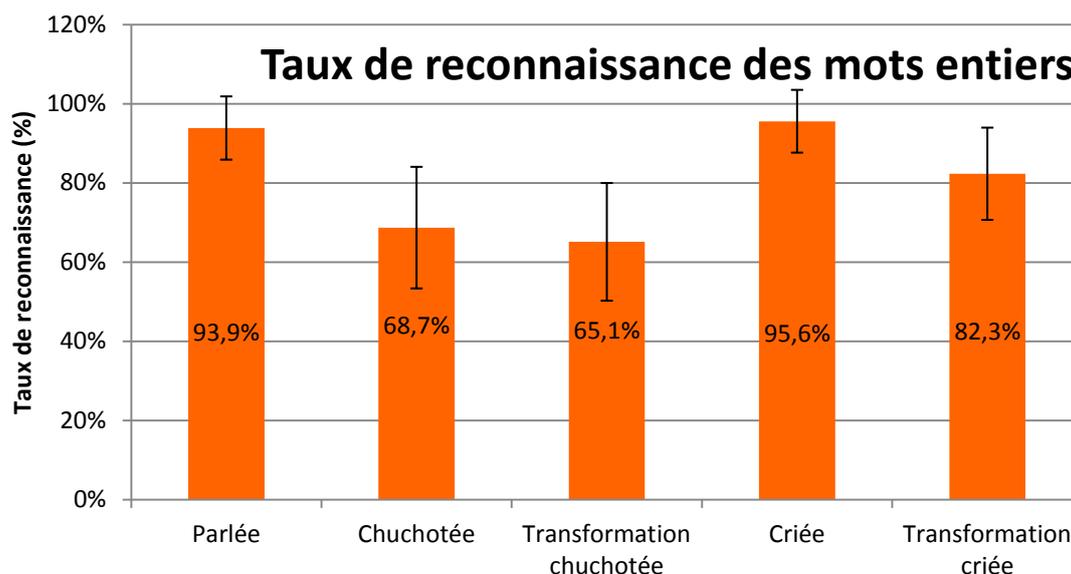


Figure 11-1: Moyennes des taux d'intelligibilité, obtenues par 25 sujets, des mots en fonction des modes de phonation ou des transformations

D'une façon générale, le taux de compréhension de la voix parlée est de 94%. La voix criée quant à elle, a un taux de compréhension légèrement supérieur (96%). Ceci peut s'expliquer par l'effort supplémentaire fait par les locuteurs dans le but d'être intelligible lors des enregistrements. Les voix chuchotées ont un taux de 69%. Ce taux correspond à la moyenne mentionnée dans la littérature (Tartter, 1989). Concernant la transformation en voix chuchotée, celle-ci a un taux de reconnaissance de 65%, ce qui est légèrement inférieur à la voix chuchotée naturelle. La transformation en voix criée quant à elle a un taux de reconnaissance de 82%, qui est inférieur au taux de la voix criée naturelle mais qui reste néanmoins bon.

Ainsi ces résultats nous montrent que nos règles de transformation ainsi que notre algorithme ne dégradent pas davantage l'intelligibilité qu'elle ne l'est pour une voix chuchotée naturelle. D'autre part, l'intelligibilité globale des transformations en voix criée est très satisfaisante. En effet, étant donné la structure des tests d'intelligibilité mise en place, loin d'être évidente (logatomes CV), un taux de 82% est un résultat satisfaisant.

### 11.1.3.2 Intelligibilité des voyelles

Penchons-nous à présent sur l'intelligibilité des voyelles uniquement. La Figure 11-2 donne les valeurs pour les 5 types de voix étudiées. L'intelligibilité globale des voyelles est similaire pour l'ensemble des 3 modes de phonation ainsi que pour les 2 transformations de voix. Malgré de fortes modifications sur la position des formants réalisées par nos transformations, le taux d'intelligibilité des voyelles est remarquablement grand. Notons alors que l'intelligibilité globale des mots semble dépendre

essentiellement de la compréhension des consonnes. De plus aucune tendance particulière sur la confusion des voyelles n'a été observée (voir les matrices de confusion en 0 et en Annexe A).

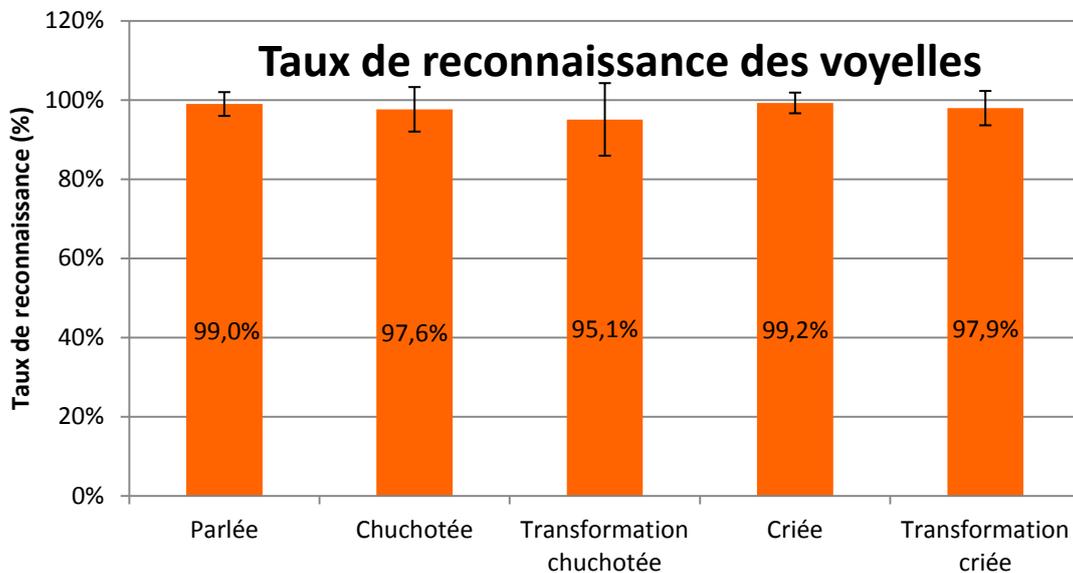


Figure 11-2 : Moyennes des taux d'intelligibilité, obtenues par 25 sujets, des voyelles en fonction des modes de phonation ou des transformations

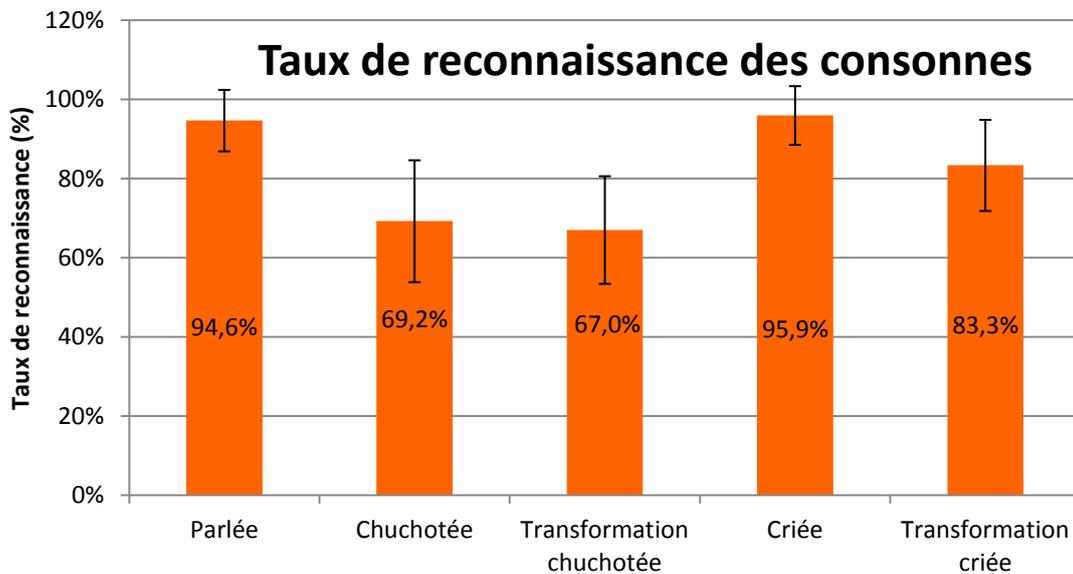


Figure 11-3: Moyennes des taux d'intelligibilité, obtenues par 25 sujets, des consonnes en fonction des modes de phonation ou des transformations

### 11.1.3.3 Intelligibilité des consonnes

Cette section se consacre donc aux taux d'intelligibilité des consonnes contenues dans les mots CV. On remarque sur la Figure 11-3 que les consonnes en voix criée (96%) sont aussi intelligibles qu'en voix parlée (95%). Les consonnes des voix chuchotées naturelles (69%) sont aussi bien reconnues

pour les voix transformées chuchotées (67%). Les consonnes des voix transformées criées sont compréhensibles à un niveau intermédiaire et reste acceptable (83%).

Ce graphique nous confirme que le taux de compréhension globale des logatomes dépend essentiellement de la compréhension des consonnes. Le taux d'intelligibilité des consonnes étant très différent de ceux obtenus pour les voyelles, nous nous intéressons alors à l'intelligibilité de chaque consonne séparément pour chacun des 5 modes de phonation étudiés.

**Intelligibilité des consonnes en voix parlée**

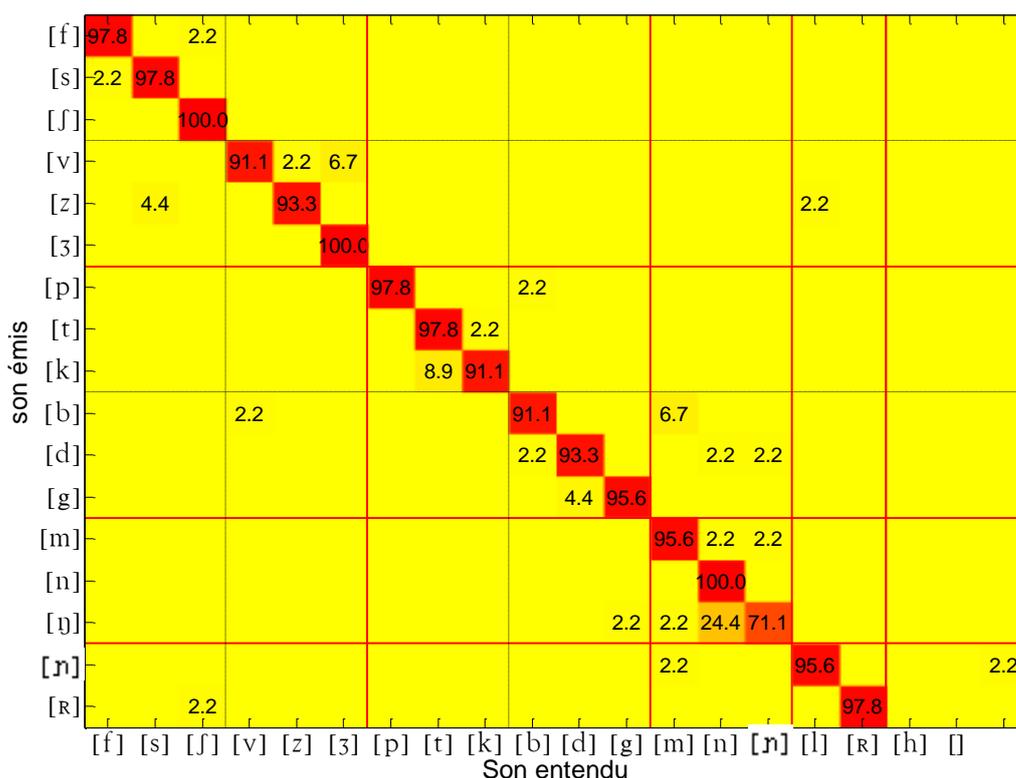


Figure 11-4: Matrice de confusion des consonnes pour la voix parlée. Les nombres correspondent aux pourcentages de réponse.

La Figure 11-4 représente la matrice de confusion pour les voix parlées entre la consonne contenue dans les mots CV et la consonne qui a été reconnue. En abscisse de la matrice de confusion on peut voir les symboles [h] et [ ]. Le [h] correspond au « h » d'exclamation que l'on retrouve dans les interjections comme le « ha ! » par exemple. Le symbole [ ] correspond à des réponses autre que l'une des 17 consonnes et du [h], comme par exemple lorsqu'un sujet a inscrit un mot qui commence pas deux consonnes phonétiques (ex : « pfou »). La dernière colonne sans aucune étiquette, correspond à l'absence de réponse par le sujet. Par souci d'homogénéité, nous conserverons ces 3 colonnes sur l'ensemble des matrices de confusion même si dans certaines situations aucune réponse n'est attribuée à ces labels.

On remarque notamment sur la matrice de confusion pour les voix parlées (cf. Figure 11-4) une légère confusion entre les deux consonnes occlusives /k/ et /t/. On peut également voir des confusions légères notamment sur /b/ et /m/ ou encore /v/ et /z/. Ces confusions ainsi que les autres, que nous ne mentionnons pas, restent toutefois anecdotiques. **En effet, sachant que chaque consonne a été entendue 45 fois, les confusions de 2,2% qu'on observe ne correspondent toutefois qu'à une seule occurrence ( $1/45=2,2\%$ ).** La plus grosse confusion concerne la consonne nasale /ɲ/ qui est confondue dans 24,4% des cas avec la consonne nasale /n/. Cette confusion est toutefois complètement portée par le mot /ɲ i/ qui prête à confusion. En effet, les analyses montrent que sur les 15 présentations du mot /ɲ i/, 11 sont perçus comme /n i/.

Ainsi hormis certaines confusions, qui peuvent dans la majorité des cas être attribuées à un manque de concentration ou à une difficulté inhérente au mot (ex : /ɲ i/), l'intelligibilité des mots CV en voix parlée de l'ensemble du corpus DB4, qui constituent la base de nos transformations, sont excellent.

### Intelligibilité des consonnes en voix criée

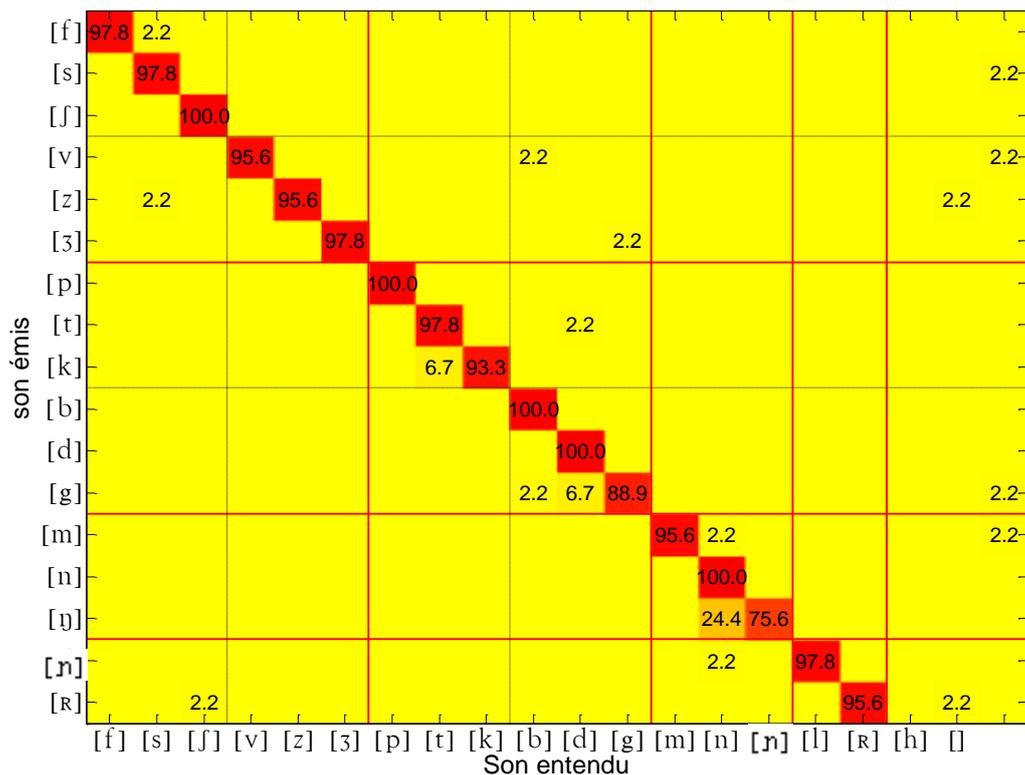


Figure 11-5: Matrice de confusion des consonnes pour la voix criée. Les nombres correspondent aux pourcentages de réponse.

En voix criée, la confusion entre le /k/ et le /t/ persiste mais est moins importante (cf. Figure 11-5). On note à nouveau la confusion entre les deux consonnes nasales /ɲ/ et /n/. Ces confusions concernent, à nouveau, uniquement le mot /ɲ i/. Hormis les quelques confusions qui, malgré tout sont anecdotiques, l'intelligibilité des voix criées reste très bonne.

**Intelligibilité des consonnes en voix chuchotée**

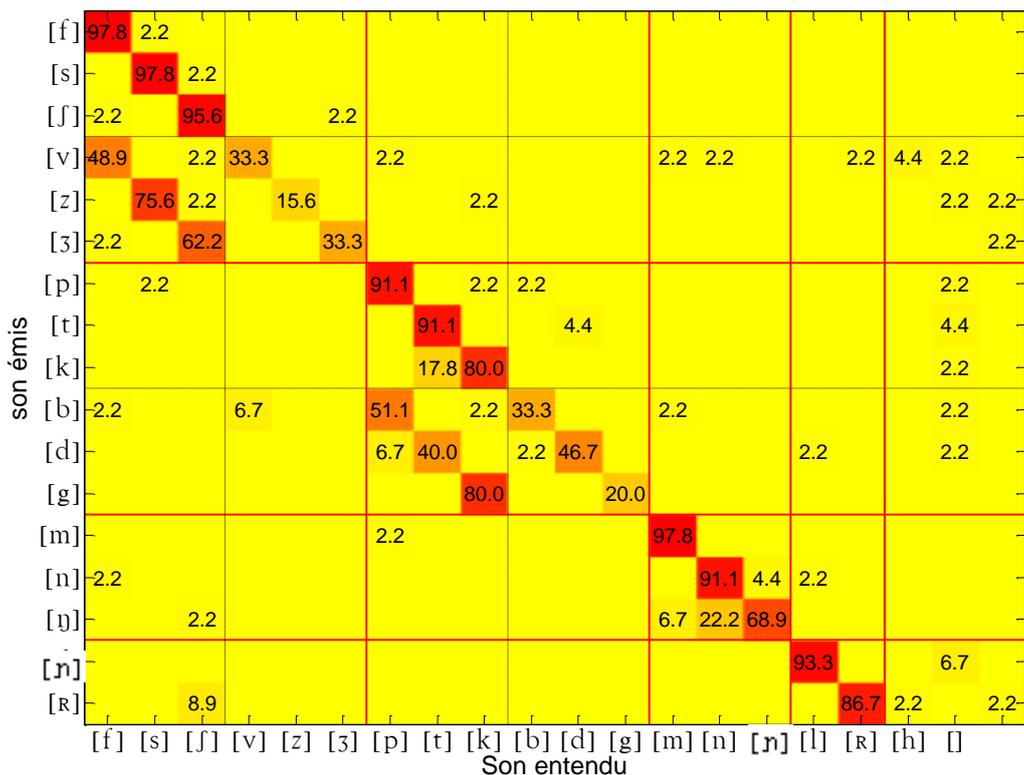


Figure 11-6: Matrice de confusion des consonnes pour la voix chuchotée. Les nombres correspondent aux pourcentages de réponse.

La matrice de confusion des consonnes en voix chuchotée est représentée sur la Figure 11-6. Les résultats présentés sur cette matrice sont différents de ceux observés sur les matrices précédentes (plus de confusions). En premier lieu, pour les fricatives voisées on observe une confusion fréquente avec leurs équivalents en non-voisé. Ce qui en soi n'est pas étonnant. En combinant les taux d'intelligibilité des fricatives voisées et non-voisées on obtient respectivement pour les consonnes /v/, /z/, /ʒ/ des taux de 82,2%, 91,2%, 95,5%. La même observation peut être faite concernant les occlusives voisées qui sont également confondues avec leurs équivalents en non-voisés. De la même manière, en combinant les résultats obtenus pour les occlusives voisées et non-voisées, on obtient pour les consonnes /b/, /d/, /g/, des taux respectifs de 84,4%, 86,7%, 100%. On retrouve également la confusion entre les deux consonnes nasales /ɲ/ et /n/ comme en voix parlée ou criée (1 confusion concerne le mot /ɲ u/ et 9 le mot /ɲ i/). La confusion entre le /k/ et le /t/ persiste dans des proportions légèrement plus grandes mais reste toutefois acceptable. Le reste des confusions que nous ne mentionnons pas reste anecdotique.

Nous disposons à présent des références concernant les voix parlées, chuchotées et criées naturelles permettant de faire la comparaison avec nos voix transformées.

### Intelligibilité des consonnes en voix transformée criée

[f]	82.2	6.7		2.2	2.2													4.4	2.2
[s]	95.6	2.2	2.2																
[ʃ]		97.8																	2.2
[v]			82.2							4.4	2.2	6.7	4.4						
[z]	4.4		4.4	77.8						4.4	4.4	4.4							
[ʒ]			8.9	64.4					6.7	11.1	6.7								2.2
[p]					95.6														4.4
[t]						95.6													4.4
[k]						2.2	8.9	86.7		2.2									
[b]									86.7		6.7		2.2						4.4
[d]									15.6	62.2	2.2	4.4	13.3						2.2
[g]									2.2	2.2	88.9		4.4						2.2
[m]									6.7		80.0	13.3							
[n]											4.4	91.1	4.4						
[ɲ]											2.2	42.2	55.6						
[ʀ]									2.2		6.7	11.1		80.0					
[ʁ]										2.2					93.3				2.2
	[f]	[s]	[ʃ]	[v]	[z]	[ʒ]	[p]	[t]	[k]	[b]	[d]	[g]	[m]	[n]	[ɲ]	[l]	[ʀ]	[h]	[]

Figure 11-7 : Matrice de confusion des consonnes pour la voix transformée criée. Les nombres correspondent aux pourcentages de réponse.

La Figure 11-7 représente la matrice de confusion pour les transformations des voix modales vers des voix criées. On constate, par cette matrice, que les confusions pour les voix transformées criées sont plus nombreuses que pour les voix criées naturelles. Néanmoins, dans l'ensemble, on constate une bonne intelligibilité des transformations (82%). On constate comme précédemment la confusion entre les deux consonnes nasales /ɲ/ et /n/. Toutefois, dans le cas des voix transformées cette confusion concerne 9 fois le mot /ɲa/ et 10 fois le mot /ɲi/ mais aucunement le mot /ɲu/. La plus grosse difficulté de compréhension se situe cependant sur les consonnes fricatives voisées, occlusives voisées ainsi que sur les nasales. Outre la consonne /ʒ/, régulièrement confondue avec la consonne /v/, et la consonne /d/ confondue avec la consonne /b/, la majorité des confusions se déplacent vers les nasales ou les liquides. Un dernier point concerne les consonnes nasales elles-mêmes. Celles-ci sont mal interprétées mais les réponses des sujets restent dans la même classe de consonne (i.e. les nasales). Enfin la liquide /l/ tend à être comprise en consonnes nasales et notamment confondue avec la nasale /n/.

Nous n'avons pas d'explication satisfaisante concernant ces confusions et notamment le fait que des occlusives voisées soit confondues avec des nasales. Bien que ces deux types de consonne soient régulièrement confondus, on n'observe pas dans les matrices de confusion des voix naturelles criées un tel phénomène. Nous estimons que ces confusions sont essentiellement dues à des problèmes liés à

la transformation elle-même ; soit concernant des erreurs de marquage de pitch, soit des difficultés de compréhension liées à la modification de l'intensité relative entre consonnes et voyelles que nous effectuons. Toutefois, malgré ces confusions, on note une bonne intelligibilité des transformations pour les voix criées.

**Intelligibilité des consonnes en voix transformée chuchotée**

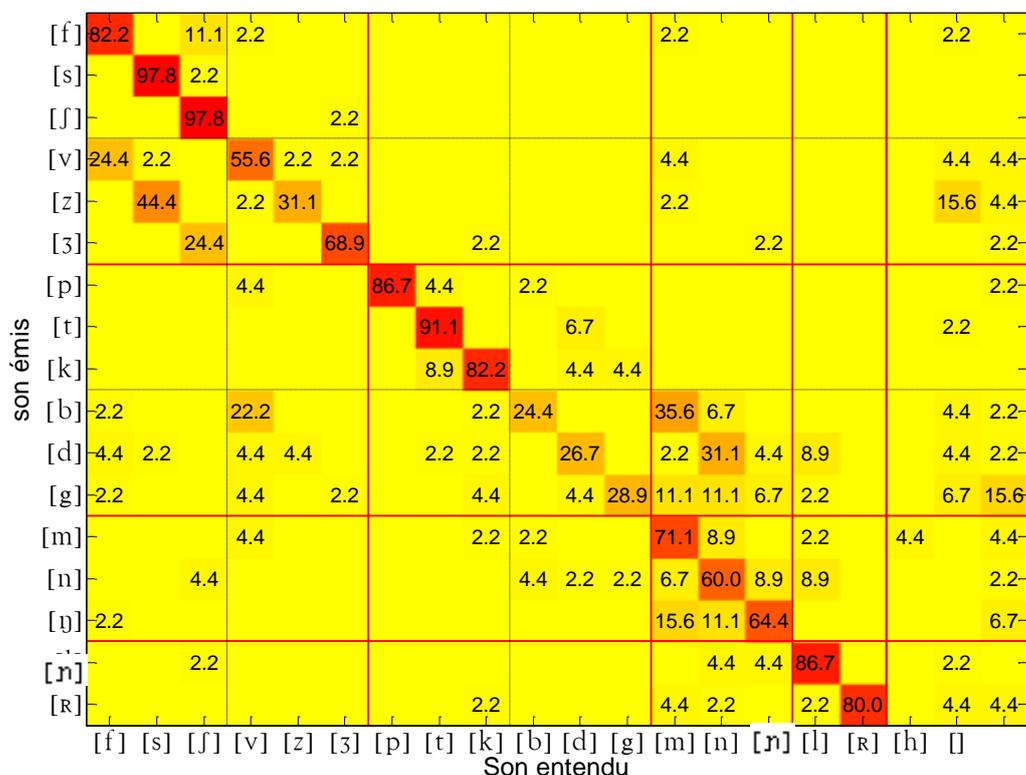


Figure 11-8: Matrice de confusion des consonnes pour la voix transformée chuchotée. Les nombres correspondent aux pourcentages de réponse.

Regardons enfin la matrice de confusion liée à la transformation en voix chuchotée. (cf. Figure 11-8). De la même manière que pour les voix chuchotées naturelles, on observe pour les fricatives non-voisées une confusion fréquente avec leurs équivalents en non-voisés. En combinant les résultats des fricatives voisées et non-voisées on obtient pour les consonnes /v/, /z/, /ʒ/ des taux respectifs de 80%, 75,5%, 93,3%. Notons également des difficultés sur le /z/ qui est régulièrement identifié comme étant plusieurs consonnes. (i.e. les sujets n'ont pas répondu en utilisant une seule consonne phonétique et une voyelle phonétique mais deux consonnes phonétiques (au moins) et une voyelle phonétique).

En revanche on constate une nette dégradation en ce qui concerne les consonnes occlusives voisées. En effet, d'après la matrice de confusion concernant les voix chuchotées naturelles (cf. Figure 11-6), les consonnes occlusives voisées sont confondues avec leurs équivalents en non-voisés. Cette tendance n'est pas du tout observée ici. En effet, ces dernières sont confondues avec l'ensemble des autres familles de manière anarchique. Ceci montre bien une forte perte d'intelligibilité pour les consonnes

occlusives voisées. Toutefois, on remarque que les consonnes occlusives voisées tendent à être perçues comme des nasales. Nous ne voyons cependant pas d'explication fiable pour ce type de confusion. Du fait de la bonne compression des occlusives non-voisées en voix parlée et chuchotée naturelle nous pensons que ce phénomène est lié aux règles de transformation. Nous apportons toutefois ci-dessous un élément de réflexion à ce sujet.

Sur la Figure 11-9 on peut voir un exemple de confusion couramment rencontrée entre /b u/ et /m u/. La colonne de gauche représente le mot /b u/ en voix parlée (a), chuchotée transformée (b) et chuchotée naturelle (c). On remarque alors que la vraie voix chuchotée pour le mot /b u/ présente une forte explosion en début de mot qui ne se retrouve pas dans la transformée chuchotée. Ceci est dû au fait que l'explosion est absente dans les voix parlées naturelles. Ainsi, le /b u/ transformée chuchotée se rapproche du /m u/. En effet, la différence entre le /b/ et le /m/ correspond à une addition de nasalité ainsi qu'une explosion. Sans cette explosion initiale il est probable que le phonème perçu soit différent. Dans le cas des consonnes non voisées ces explosions sont toutefois préservées étant donné qu'aucune modification n'est faite sur les parties non voisées du signal.

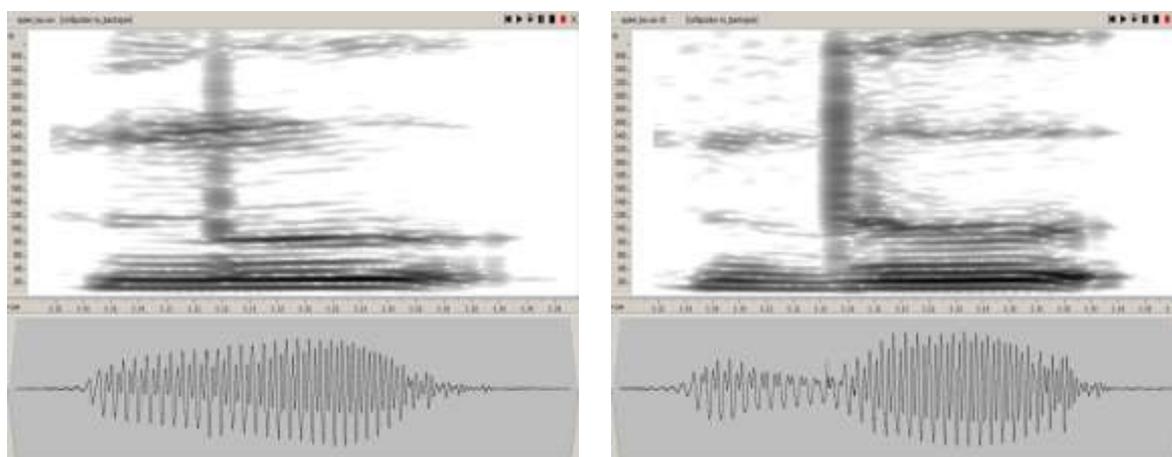
Certaines difficultés semblent également avoir été rencontrées concernant les consonnes nasales. La confusion entre les deux consonnes nasales /ɲ/ et /n/, n'est plus aussi marquée que précédemment et représente 5 confusions sur le mot /ɲ i/, mais la consonne /ɲ/ semble également être confondue avec la consonne /m/. (6 confusions pour le mot /ɲ i/ et une pour le mot /ɲ u/)

En conclusion, hormis pour les occlusives voisées, la matrice de confusion est relativement proche de celle obtenue pour les voix chuchotées naturelles.

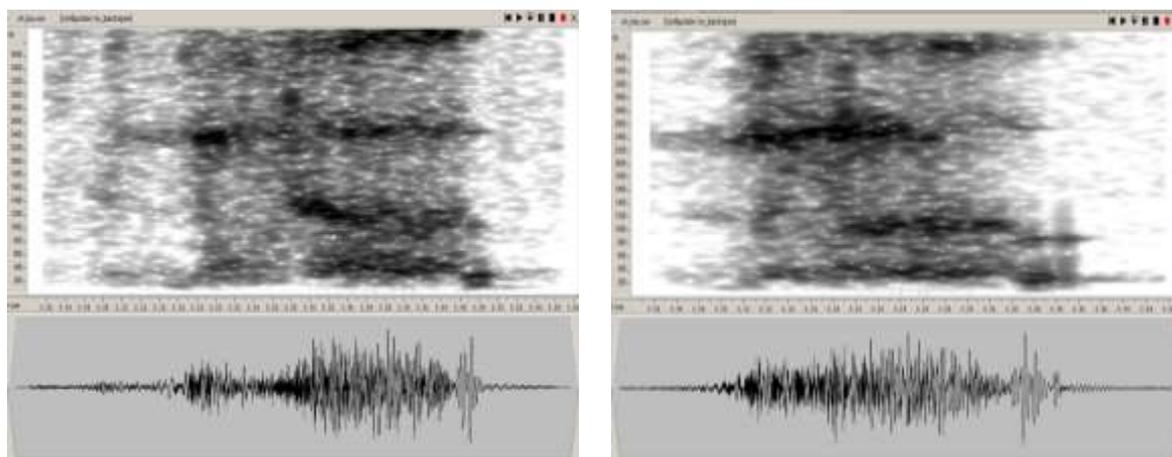
/bu/

/mu/

a- Voix parlée



b- Voix transformées chuchotées



c- Voix chuchotées naturelles

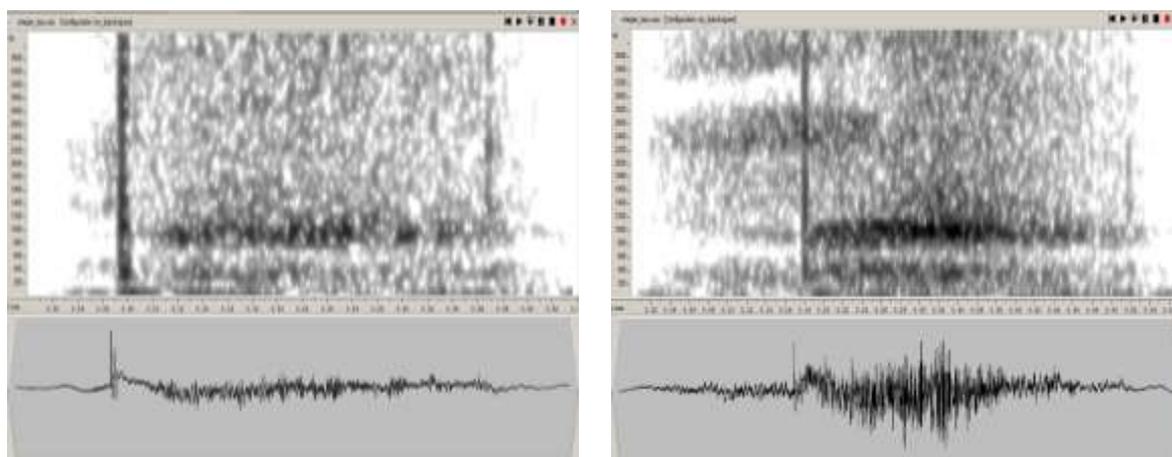


Figure 11-9: Représentation des mots /bu/ (à gauche) et /mu/ (à droite) pour les modes parlées (en haut) les transformée en voix chuchotée (au milieu) ainsi que les voix chuchotées naturelles (en bas). Spectrogramme représenté entre 0 et 4 kHz.

#### 11.1.4 Conclusion des tests d'intelligibilité

On constate donc que les voyelles sont pour les 5 modes étudiés parfaitement reconnues et que l'intelligibilité des mots dans leur globalité dépend essentiellement de la compréhension des consonnes. Les voix parlées naturelles et les voix criées naturelles ne semblent pas poser de problème en termes d'intelligibilité. Les voix chuchotées possèdent toutefois un taux de compréhension bien plus faible que pour celle des voix parlées mais qui dépend fortement de la confusion faite entre les consonnes fricatives voisées et non voisées ainsi que de la confusion entre les occlusives voisées et non voisées. Ces confusions sont prévisibles étant donné que la différenciation entre consonnes voisées et non voisées se fait essentiellement sur la présence ou non d'une F0. Ainsi, en voix chuchotée, la différenciation devient plus difficile à effectuer. Néanmoins, en considérant les confusions entre ces consonnes en voix chuchotée comme correctes, le taux d'intelligibilité correspondrait à 90%.

Les voix criées transformées, quant à elles, possèdent une intelligibilité de 82 % ce qui correspond à un bon taux de compréhension. Ainsi, notre algorithme de transformation de voix modale en voix créée ne semble pas altérer la compréhension des phonèmes, et ce, malgré les modifications drastiques qui sont appliquées. Les voix chuchotées transformées semblent toutefois être plus problématiques, même si le taux de compréhension est identique à celui des voix chuchotées naturelles. Nous avons pu constater que notre algorithme dégrade sensiblement l'intelligibilité des occlusives voisées. Des difficultés sont également rencontrées concernant les consonnes nasales mais dans des proportions moins excessives. Ainsi, même en ne tenant pas compte des confusions entre les occlusives voisées et non-voisées, et celles entre les fricatives voisées et non voisées, le taux d'intelligibilité est de 72%. Nous penchons pour une explication visant à mettre en cause notre algorithme mais également à des phénomènes de reconstruction d'explosion (pour les occlusives) qui ne peuvent être reproduits facilement. Concernant les consonnes nasales nous émettons l'hypothèse que la mauvaise compréhension est liée aux outils que nous utilisons dans l'analyse-synthèse et en particulier par l'utilisation d'un prédicteur linéaire (filtre AR). L'utilisation d'un filtre AR, se base sur l'hypothèse que la fonction de transfert du conduit vocal peut être modélisée par un filtre tout-pôle. Cette supposition est acceptable pour la plupart des sons de la parole voisée, mais n'est pas appropriée pour les nasales et les fricatives. Lors de la production de ce type de son, la fonction de transfert arbore des zéros. L'utilisation de filtres comprenant des zéros est en générale préconisée bien que si l'ordre du filtre AR est suffisant, la plupart des sons de la voix peuvent être modélisés par cette méthode (Makhoul et Wolf, 1972). Néanmoins cette restriction de la modélisation par filtre AR peut engendrer certains problèmes lors de la transformation. De plus la transformation en voix chuchotée, contrairement à la transformation en voix créée (ou les consonnes nasales sont bien reconnues) ne considère plus du tout le résidu de l'analyse, qui par définition contient les éventuelles erreurs

d'approximation. Ainsi toute erreur d'analyse est définitivement perdue et une mauvaise estimation du filtre du conduit vocal engendre des conséquences catastrophiques au niveau de l'intelligibilité.

## 11.2 Perception de distance

Après avoir étudié l'intelligibilité de nos transformations, nous nous penchons à présent sur le cœur de notre étude à savoir la perception de distance à partir de nos transformations. Nous ne considérons dans cette section que les mots transformés en voix criée et pas les transformations en voix chuchotée qui ne présentent pas un grand intérêt dans ce test. Nous expliquons dans un premier temps, les choix faits afin de sélectionner certains des mots de la base de données. Nous continuons ensuite, comme pour la section précédente, par donner les différents réglages faits et les règles choisies pour la transformation des voix parlées. Enfin nous décrivons la procédure choisie pour mener à bien ces tests avant de nous pencher sur les résultats.

### 11.2.1 Choix des signaux tests

Afin de réaliser le test concernant l'estimation de distance via nos transformations nous avons choisi de sélectionner uniquement certains mots. En effet, des tests sur l'ensemble du corpus DB4 seraient bien trop longs et seraient donc difficilement acceptables par les sujets. Ainsi, nous ne considérons que les mots CVC et CVCV. Ce choix est fait afin de valider les transformations de mots monosyllabiques et de mots bi-syllabiques. Étant donné la bonne intelligibilité des 3 voyelles constatée au cours des tests précédents, nous nous limitons à une seule voyelle. Nous avons choisi d'utiliser les mots contenant la voyelle /a/. Nous retenons cette voyelle car c'est elle qui subit le plus de modifications dans notre algorithme et qui de plus, est la voyelle la plus utilisée dans la langue française (Combescure, 1981). En effet, le premier formant de la voyelle /a/ est déplacé sur la troisième harmonique de la F0 (i.e. 3\*F0) se qui correspond à un premier formant placé 990 Hz au niveau du maximum de la F0. D'autre part, nous avons choisi de n'utiliser qu'un seul mot par type de consonne. Ainsi nous retenons les mots dont le taux d'intelligibilité des consonnes est le meilleur. Nous retenons alors, pour chaque structure 7 mots qui sont listés dans le Tableau 11-2. Nous utilisons ces mots prononcés par les 3 locuteurs pour réaliser ce test.

Tableau 11-2: Mots choisis parmi notre corpus pour mener les tests de perception de distance

<b>CVC</b>	/p a p/	/g a g/	/ʃ a ʃ/	/v a v/	/n a n/	/l a l/	/r a r/
<b>CVCV</b>	/p a p a/	/g a g a/	/ʃ a ʃ a/	/v a v a/	/n a n a/	/l a l a/	/r a r a/

### 11.2.2 Règles utilisées pour générer les signaux tests

Nous avons pour ce test créé deux types de transformations. L'unique différence entre ces deux transformations se situe au niveau de la fréquence fondamentale. En effet, la première transformation correspond à la transformation en voix criée la plus poussée. Les contours de F0 utilisés sont tracés sur la Figure 11-10 pour les mots CVC et sur la Figure 11-11 pour les mots CVCV. Les valeurs choisies pour les points caractéristiques correspondent à une approximation issue des analyses réalisées sur le corpus DB4. La deuxième transformation ne subit pas de modification du contour de F0. Nous avons choisi de réaliser une transformation qui modifie la F0 en appliquant simplement un facteur de multiplication (translation du contour de la F0 de la voix parlée vers les hautes fréquences), tout en conservant les autres règles de transformation inchangées. Ainsi cette transformation dispose d'une F0 moyenne égale à la F0 moyenne de la transformation en voix criée (première transformation). Nous avons décidé d'inclure ce type de transformation, afin de comparer l'influence de la modélisation de la dynamique de F0, que nous proposons, aux méthodes observées dans la littérature (à savoir une simple multiplication du contour de F0) (Cheyne et al., 2009). Il s'agit là de la transformation de référence. Les autres paramètres sont communs aux deux transformations. A savoir que l'intensité des voyelles est multipliée par 2, les voyelles sont allongées de 40% et l'intensité de la deuxième voyelle des mots CVCV est 40% plus forte que la première.

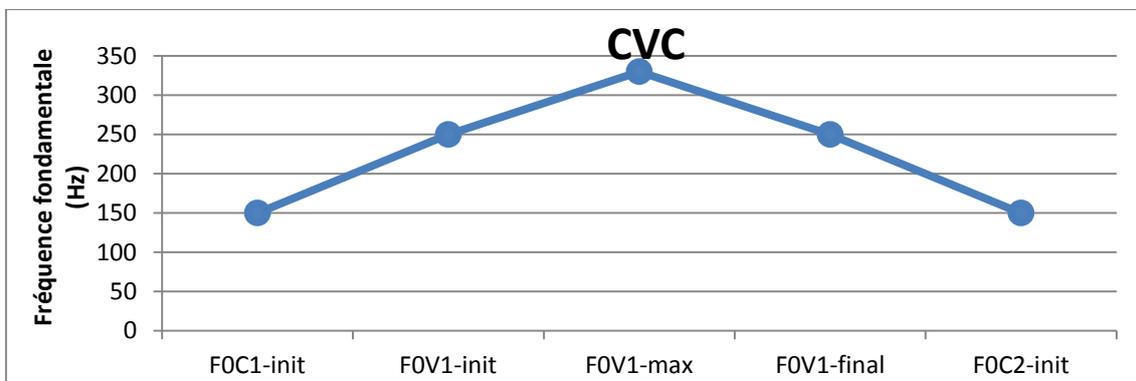


Figure 11-10: Représentation du contour de F0 pour les mots CVC utilisé pour la transformation en voix criée

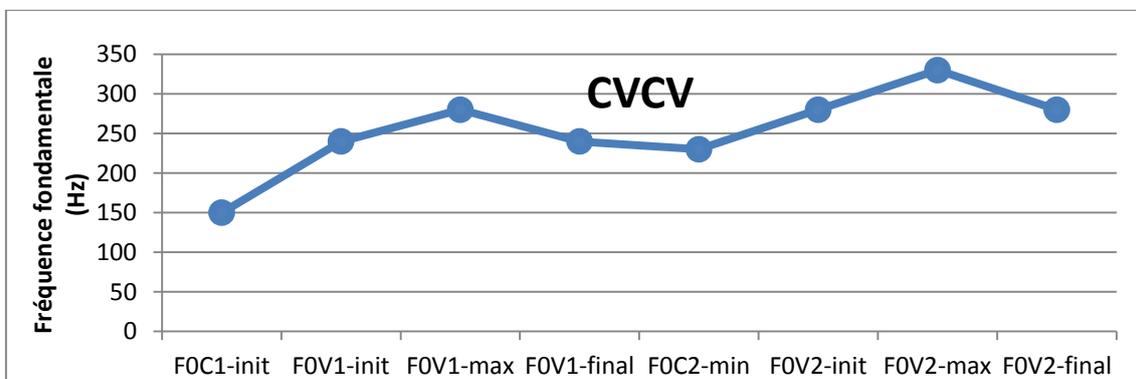


Figure 11-11: Représentation du contour de F0 pour les mots CVCV utilisés pour la transformation en voix criée

### 11.2.3 Description du test de perception de la distance

Nous utilisons pour ce test 4 types de voix : des voix parlées, des voix criées, des transformations en voix criée (appelé ici DELTA\_F0) ainsi que des transformations par multiplication du contour de F0 (appelée ici xF0). Ce test se base sur 2 listes de mots (cf. Tableau 11-2). Une série de mots CVC et une série de mots CVCV et ce pour chacun des 3 locuteurs de la base de données DB4. Soit un total de 6 séries.

Ce test est construit suivant le principe de l'immersion. Ceci signifie que le sujet doit s'imaginer étant placé à distance d'un locuteur qui lui parle (cf. Figure 11-12). Les sujets, en fonction de la voix du locuteur doivent alors répondre à la question : « *A quelle distance se trouve le locuteur ?* ». Pour indiquer la distance perçue, il n'y a que trois choix possibles : *proche, moyen, loin*. Ces trois niveaux sont représentés dans le programme de test par des photos faites d'un sujet placé à des distances différentes et correspondant aux plages qualitatives proposées dans notre cahier des charges. Les photos ont été prises lors des enregistrements de DB1 (cf. Figure 11-13). Le choix des sujets se fait alors en cliquant sur l'image correspondante.

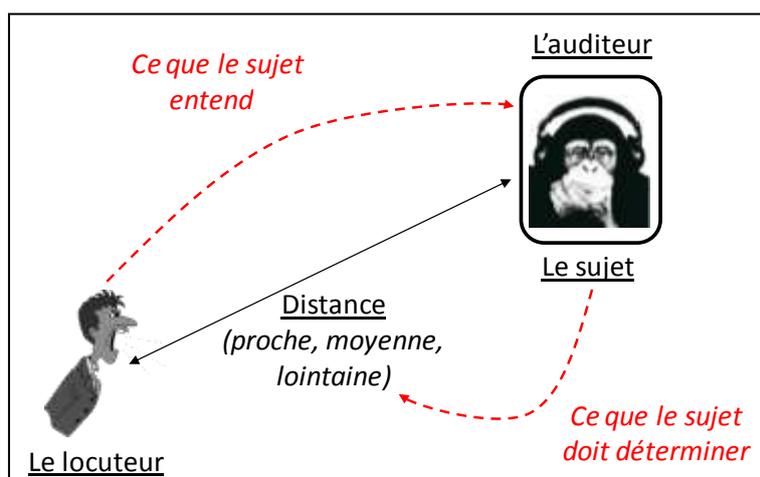


Figure 11-12: Principe du test à la troisième personne utilisé pour le test de perception de distance

Le test s'effectue en 2 étapes. La première étape est une étape d'entraînement. Pour chacune des séries, l'entraînement consiste à entendre l'ensemble des voix parlées et des voix criées naturelles qui servent de référence durant le test (14 voix de référence : 7 parlées et 7 criées). A chaque écoute des ces mots, le programme indiquant aux sujets quelle était la bonne réponse. Ainsi, pour les voix parlées naturelles, la réponse est *proche*, et pour les voix criées naturelles la réponse est *loin*. Seuls les extrêmes du test sont présentés dans l'étape d'entraînement afin de donner aux sujets les limites hautes et basses de l'échelle de distance. Par la suite, le test en lui-même débute. Dans ce test on retrouve dans un ordre aléatoire les 14 voix naturelles de référence, ainsi que les 14 transformations réalisées (7 transformations DELTA-F0 et 7 transformations xF0) soit 28 mots (pour un locuteur donné). A l'écoute de chaque mot, les sujets doivent indiquer la distance perçue du locuteur en cliquant sur

l'image qui convient à leurs choix. Au préalable, les sujets ont été informés qu'il s'agissait de voix de synthèse ayant pour but de refléter une distance de communication. Ceci afin qu'ils ne soient pas perturbés par l'écoute de voix qui peuvent paraître peu naturelles. D'autre part, pour chaque locuteur nous avons pris soin de leur préciser qu'une voix criée est une voix qui est plus aigüe mais qu'une voix aigüe ne signifie pas qu'il s'agit d'une voix criée. Nous précisons alors aux sujets que la différenciation de ces deux types de voix s'effectue par la perception de l'effort vocal. Il est alors demandé aux sujets de se concentrer sur ce point afin de réaliser le test de perception de la distance. Cette précaution est indispensable car sans cette consigne, les sujets basent leurs réponses sur la hauteur de F0 uniquement et non pas sur la sensation d'effort perçue. Ainsi le test se base principalement sur *la perception de l'effort vocal* ressentie par les sujets. L'intensité de l'ensemble des échantillons sonores a été égalisée pour ce test.

Les réponses ont été collectées suivant le principe de la note d'opinion moyenne (en anglais *mean opinion score*). Ainsi à chaque réponse correspond un score entre 1 et 3. La réponse proche correspond à 1, moyen correspond à 2, et loin correspondent à 3. Par la suite pour chaque type de voix, la moyenne de toutes les réponses est calculée.



Figure 11-13: Apparence du programme de test pour la perception auditive de la distance des voix criées transformées

#### 11.2.4 Résultats du test de perception de distance

En tout, **24 sujets** ont participé à ce test. Chaque sujet ayant effectué l'estimation de la distance sur 4 séries : une CVC et une CVCV pour deux locuteurs différents. L'ordre dans lequel les séries ont été entendues a été fait aléatoirement afin que chacune des séries soit entendues un nombre de fois équivalent. De plus l'ordre de passage des séries a également été fait de manière à ce que chacune d'elles soit entendue un même nombre de fois pour une position donnée lors de la présentation des 4

séries. Ainsi chaque série a été entendue 16 fois. Ce qui représente un total de 2688 mots sur lesquelles la distance a été estimée. La Figure 11-14 montre les résultats pour l'ensemble de ces mots. On peut formuler une première constatation concernant les vraies voix parlées et criées. Si les voix parlées sont bien identifiées comme proches, les voix criées ne sont étonnamment plus perçues aussi loin que ce qu'il a été indiqué durant la série d'entraînement. Deux explications peuvent être envisagées. D'une part, la définition des zones ne convenait pas aux sujets. Le terme moyen qui n'est pas représenté durant l'entraînement pouvant éventuellement être déroutant pour l'estimation. D'autre part, il se peut que les voix transformées DELTA\_F0 soient effectivement perçues plus loin que les voix criées naturelles. C'est d'ailleurs ce que l'on observe sur les résultats. Ce résultat est très inattendu. En effet, on s'attendait à ce que nos transformations ne soient pas perçues plus loin que les voix d'origine. Or on observe le contraire dans la majorité des cas.

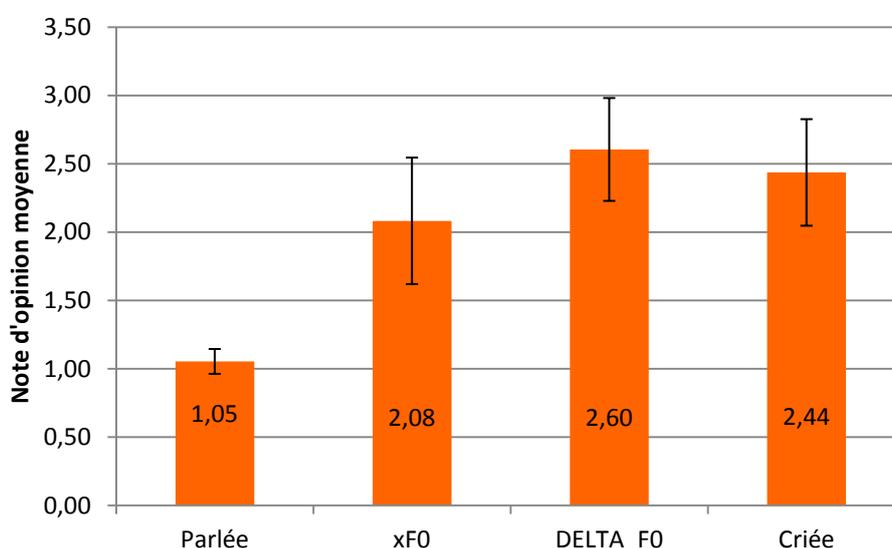


Figure 11-14 : Résultats d'opinion moyenne des sujets sur l'ensemble du test (3 : loin, 2 : moyen, 1 : proche)

Une deuxième constatation, la transformation xF0 (notre « référence ») est perçue moyennement proche du locuteur alors que les mêmes modifications que celles appliquées pour la transformation DELTA\_F0 (la nôtre) sont apportées sur l'intensité, les formants et la durée. Rappelons que les transformations xF0 ne disposent pas de la modélisation du contour de F0 (contrairement aux transformations DELTA\_F0) mais seulement d'une simple multiplication des contours de F0 de la voix parlée, dans le but d'obtenir une valeur moyenne identique à celle de la transformation DELTA\_F0 (pour un mot donné). Plusieurs sujets nous ont avoué que les logatomes CVCV transformées par simple décalage du contour de F0, étaient plus difficilement associables à une distance de communication. En effet, certains locuteurs tels que le locuteur 1 et 2 ont effectué des montées de F0 en fin d'énoncé (cf. CHAPITRE 9). De ce fait, la différence de F0 entre la première syllabe et la deuxième était alors ressemblante aux transformations que nous appliquons par nos règles

(i.e. sur les transformations DELTA\_F0). Cependant, notons que le naturel des transformations xF0 est moindre que pour les transformations DELTA\_F0. Cette différence peut alors être la cause de la gêne mentionnée par les sujets car la prosodie globale était similaire entre les 2 transformations mais sans le naturel de la voix.

Pour mieux comprendre ces résultats, penchons-nous sur les résultats obtenus pour chaque série. La Figure 11-15 montre ces résultats pour les logatomes CVC et la Figure 11-16 pour les logatomes CVCV. On constate que les voix transformées DELTA\_F0 sont perçues plus loin que les vraies voix criées pour le locuteur 1 et 2 mais pas pour le locuteur 3. Nous expliquons ceci par le fait que le contour de la F0 choisi, est plus intense que celui observé en voix créée pour les locuteurs 1 et 2, et est moins intense pour le locuteur 3. Ainsi, par rapport aux voix criées naturelles, les transformations paraissent moins intenses pour le locuteur 3 que pour les locuteurs 1 et 2, car elles ont une F0 moyenne plus faible que celles des transformations DELTA\_F0 contrairement aux voix criées naturelles des locuteurs 1 et 2.

Examinons également les résultats sous une autre forme. Sur la Figure 11-17 sont représentées les répartitions des réponses des sujets en fonction du type de stimuli (CVC et CVCV). A partir de ces figures, on remarque que la voix parlée CVC est relativement bien identifiée comme une voix proche. Concernant la voix créée pour les locuteurs 1 et 2 on constate que le choix est reparti aussi bien sur des voix « loin » que « moyennement loin ». Pour le locuteur 3 en revanche les voix criées sont clairement identifiées comme lointaines. On retrouve ici la problématique de la valeur moyenne de F0 qui est plus élevée pour les transformations que pour les voix criées chez les locuteurs 1 et 2. Ceci se traduit également par des voix transformées DELTA\_F0 (qui est la nôtre) perçues majoritairement comme lointaines. En revanche les transformations xF0 (la « référence ») ne présentent aucune tendance particulière et les résultats se répartissent sur l'ensemble des 3 choix possibles (proche, moyen, loin). Ce qui traduit bien la difficulté d'estimer une distance à partir de ces stimuli. Chez le locuteur 3 les transformations DELTA\_F0 sont réparties sur les distances moyennes et proches tandis que les xF0 sont repartis sur proche et moyen. Ainsi, les transformations xF0 sont clairement perçues moins loin que les voix criées naturelles et que les transformations DELTA\_F0.

Pour les logatomes CVCV on constate le même type de répartition. On constate notamment que les résultats des transformations xF0 sont plus confus pour les transformations DELTA\_F0 que pour les voix naturelles. De la même manière les transformations xF0 sont majoritairement perçues plus proches que les DELTA\_F0.

Soulignons que les résultats diffus sur les transformations xF0 reflètent sans doute un problème lié au naturel de la voix. En effet, à l'écoute la plupart des sujets nous ont confié que ces transformations étaient désagréables et bizarres à l'écoute. Ce sont des chimères ! Ceci va encore une fois dans le sens

de notre hypothèse à savoir que l'évolution de la F0 est indispensable pour la perception de l'effort vocal et donc de la distance mais également du naturel de la voix.

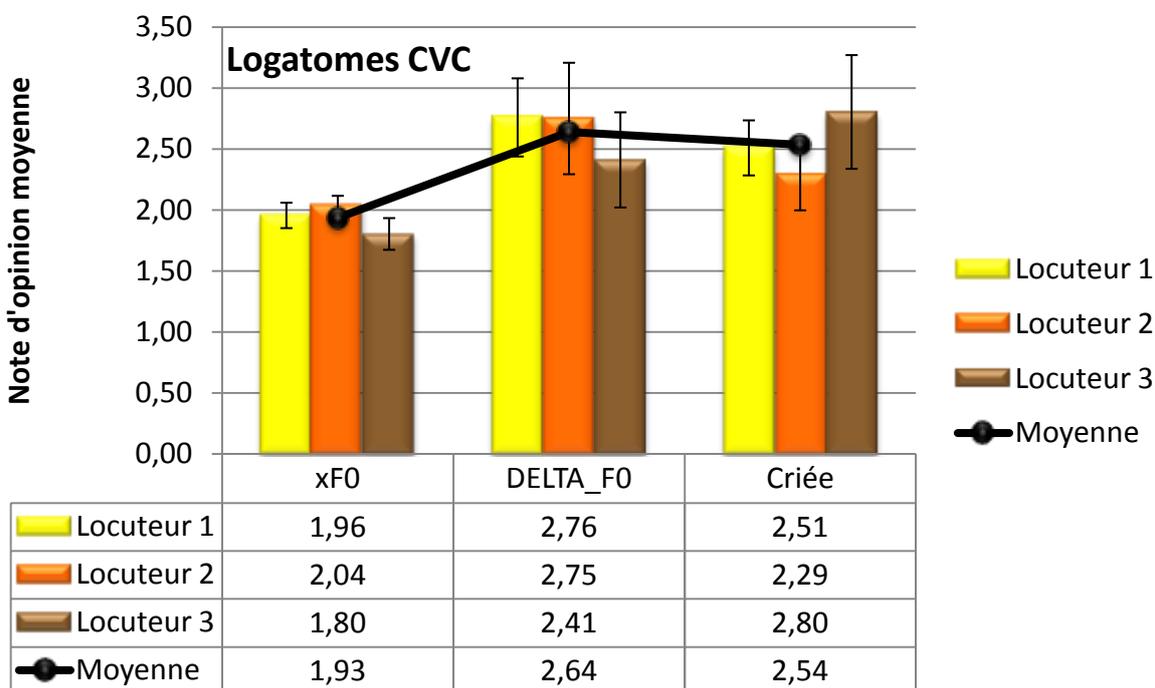


Figure 11-15: Résultats d'opinion moyenne des logatomes CVC des sujets en fonction du locuteur. (3 : loin, 2 : moyen, 1 : proche)

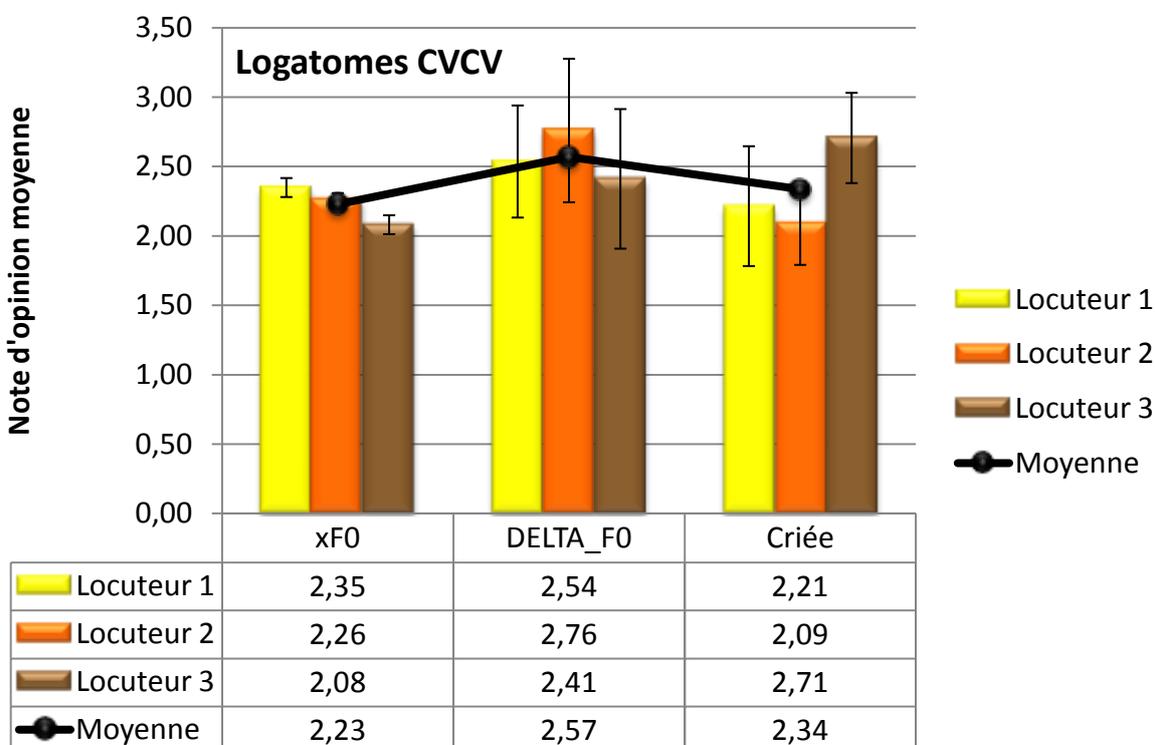


Figure 11-16 : Résultats d'opinion moyenne des logatomes CVCV des sujets en fonction du locuteur (3 : loin, 2 : moyen, 1 : proche)

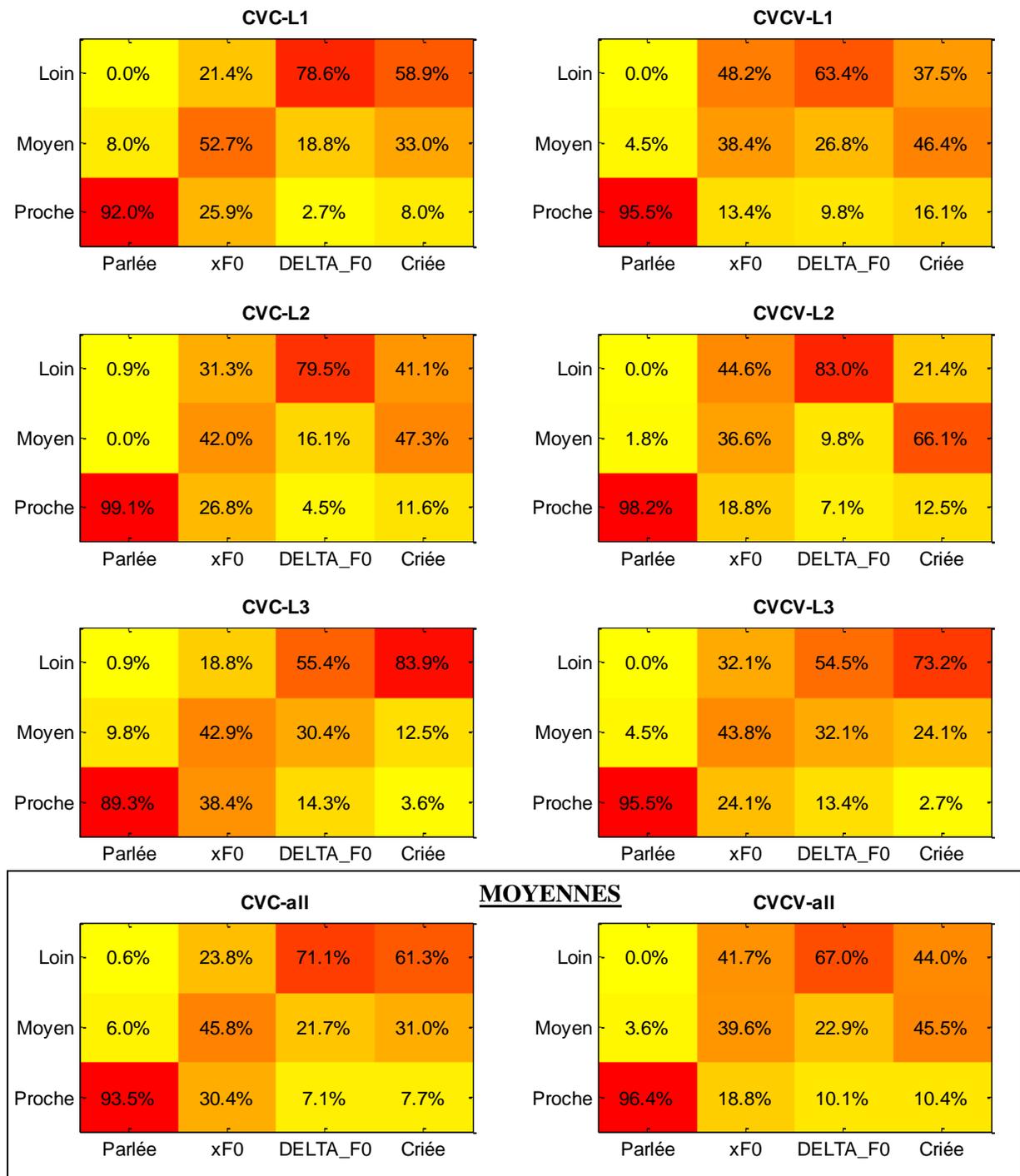


Figure 11-17 : Répartition des réponses en fonction du type de stimuli ( CVC, à gauche et CVCV, à droite) en pourcentage des résultats obtenus concernant la perception de la distance. Pour chaque type de voix on représente la répartition en pourcentage des résultats obtenus concernant la perception de la distance.

**Remarque :**

Dans l'ensemble, les tendances pour chaque type sont respectées et notamment le fait que nos transformations de voix modale en voix criée (DELTA\_F0) sont perçues comme venant de plus loin que les transformations utilisant une simple multiplication du contour de F0 (xF0).

Rappelons que nous n'utilisons que des mots dont la voyelle est un /a/. Cette voyelle subit le plus de modification sur le premier formant en voix criée (Traunmüller et Eriksson, 2000) et ce déplacement est reproduit lors de nos transformations. Ainsi, c'est cette voyelle qui est la plus déformée et qui risque d'être la moins naturelle, ce qui peut alors perturber la perception de la distance. C'est pourquoi nous n'avons pas inclus des mots contenant les voyelles /i/ ou /u/ dans les tests de perception de la distance. Notons toutefois que les transformations des mots contenant les voyelles /i/ et /u/ sont de qualité équivalente. A priori, nous estimons que la perception de la distance à partir de ces transformations aurait été similaire. Cependant il est possible que ces mots ne soient pas perçus de la même manière que ceux que nous avons utilisés. En effet, étant donné que les voyelles /i/ et /u/ subissent moins de modification que la voyelle /a/, la différence entre le mode parlé et le mode transformé est moins flagrante (au niveau phonétique). Ainsi, la faible différence entre la place du premier formant pourrait amener à des résultats différents. Malheureusement, nous n'avons pas pu tester cette hypothèse par manque de temps.

### 11.3 Conclusions du chapitre 11

---

Dans ce chapitre nous avons évalué selon deux critères différents les transformations réalisées à partir des règles que nous proposons.

Le premier test est destiné à évaluer l'intelligibilité des mots CV transformés par nos algorithmes. Nous comparons ainsi l'intelligibilité des transformations des voix parlées en voix chuchotée à l'intelligibilité des voix chuchotées naturelles, et les transformations des voix parlées en voix criée à l'intelligibilité des voix criées naturelles. Nous avons également pris soin de mesurer l'intelligibilité des voix parlées naturelles qui ont été utilisées dans les transformations. Les résultats nous montrent que les différentes étapes proposées pour réaliser ces transformations de voix n'affectent que peu l'intelligibilité globale pour les transformations en voix chuchotée (chuchotée naturelle : 68,7%, chuchotée transformée : 65,1%) mais dégradent légèrement celle des transformations en voix criée (criée naturelle : 95,6%, criée transformée : 82,3 %). Néanmoins, il s'avère que malgré le taux d'intelligibilité acceptable, les transformations en voix chuchotée affectent sensiblement l'intelligibilité des consonnes occlusives. En effet, pour les voix chuchotées naturelles l'intelligibilité est majoritairement dégradée du fait des confusions entre consonnes voisées et non voisées. Or dans

nos transformations, même si la confusion est encore observée pour les fricatives, ce n'est plus le cas pour les occlusives ce qui traduit une perte d'intelligibilité sur ces dernières. Les transformations en voix criée dégradent quant à elles légèrement l'intelligibilité des consonnes voisées mais le taux de compréhension reste très correct. Dans l'ensemble, pour les deux transformations, et ce malgré des modifications drastiques de la position du premier formant, l'intelligibilité des voyelles (/a/, /i/ et /u/) est forte (96,5%). Ainsi le taux de compréhension globale est directement lié à la compréhension des consonnes.

Dans un second temps, nous avons évalué la perception de distance à partir des mots transformés en voix criée. Nous ne considérons pas dans ce test les transformations en voix chuchotée qui, *a priori*, ne devraient pas poser de problème. Pour réaliser ce test, seul 7 mots CVC et 7 mots CVCV pour chaque locuteur ont été utilisés. Ce test consistait alors à juger de la distance du locuteur en écoutant des mots parlées et criées naturelles ainsi que les mots transformés par nos soins en voix criée (DELTA\_F0). Nous avons ajouté à ce test, une transformation de référence permettant de juger de l'apport de notre modélisation du contour de F0, pour la perception de la distance. En effet, une deuxième transformation de voix (xF0) a été réalisée qui n'est différente de la notre (DETLA\_F0) que par son contour de F0. L'ensemble des variations et des règles sont conservées mais le contour de F0 de ces dernières a simplement été décalé vers le haut. Ce décalage de F0 est réalisé afin d'obtenir la même valeur moyenne de F0 que celle des mots qui ont subi une transformation tenant compte du modèle de contour de F0. Cette transformation de référence correspond à la modification de F0 généralement observée dans la littérature qui consiste simplement à décaler le contour de F0. Les transformations DELTA\_F0 sont quant à elles perçues légèrement plus loin ou légèrement plus près que les voix criées naturelles (en fonction du locuteur). En revanche les transformations xF0 ne sont perçues que moyennement loin. On observe cependant une tendance à percevoir les mots CVCV transformés xF0 plus loin que les mots CVC transformés xF0.

Dans l'ensemble ce test montre alors que nos transformations de voix permettent de percevoir une distance d'un locuteur de la même manière qu'une voix criée naturelle. De plus l'utilisation des transformations xF0 nous permet de confirmer l'importance d'une modélisation du contour de F0 pour les transformations en voix criée. Notons que les transformations DELTA\_F0, d'après les sujets, paraissent également plus naturelles que les transformations xF0. Ces dernières sont en effet, à l'écoute, peu probables. Nous n'avons malheureusement pas pu effectuer des tests de naturel de la voix par manque de temps. Ceux-ci auraient toutefois pu nous apporter des informations supplémentaires quant à la perception de la distance en apportant la dimension de naturel. Toutefois ce point constitue l'une de des perspectives de ce travail.

# Conclusions générales et perspective

---

« *Les mots ne sont que les cailloux repères, ils ne sont pas le chemin.* »  
— Jacques Salomé

Notre étude s'est consacrée à la transformation de la voix pour indiquer la distance d'un locuteur. Ceci est possible car la perception de la distance d'un locuteur se fait essentiellement par le biais de l'effort vocal fourni par celui-ci. En effet, pour de courtes distances de communication le locuteur utilisera une voix chuchotée. Pour une distance moyenne (i.e. distance conversationnelle) il utilisera une voix parlée alors qu'il criera pour une grande distance de communication. Cette modification de l'effort vocal du locuteur ainsi que la modification de la nature de la voix engendrent de nombreuses variations sur les paramètres de la voix qui, une fois perçues par son interlocuteur, lui permettent d'estimer la distance du locuteur. Ainsi, nous cherchons à transformer une voix modale en une voix chuchotée ou criée permettant d'apporter la notion de distance d'un locuteur. Ce type d'application peut notamment être utilisé dans les systèmes de radiocommunication en 3D pour apporter la notion de distance qui est manquante dans ces systèmes.

Si les mécanismes de production de la voix chuchotée ainsi que les variations pertinentes sont relativement bien compris, ce n'est pas le cas pour la voix criée. Il nous manque dans la littérature les descriptions précises concernant la voix criée, en particulier concernant la modélisation des mécanismes de production et de perception. Si la voix chuchotée est principalement décrite par l'absence de vibrations des cordes vocales, les paramètres pertinents pour l'identification de voix criée sont nettement moins clairs. Ainsi, la problématique centrale de cette étude est la suivante.

*Quelles sont les variations des caractéristiques de la voix qui permettent de différencier une voix modale d'une voix chuchotée ou d'une voix criée ?*

Ce travail s'est décomposé en 3 grandes parties. La première étant consacrée à l'étude de l'état de l'art des connaissances actuelles sur la voix criée et chuchotée ainsi que sur les méthodes de transformation de la voix parlée. La deuxième partie constitue l'une de nos contributions principales ; à savoir l'enregistrement de plusieurs corpus pour différentes distances de communication et une suite

d'analyses approfondies destinées à extraire les informations les plus pertinentes d'une voix créée. Sur la base de ces deux premières parties, la troisième partie se consacre à l'élaboration de règles de transformation de voix parlée en voix chuchotée ou créée, et à la mise en œuvre des algorithmes de transformation ; ce qui constitue un autre apport majeur de notre étude.

Nous savons que l'estimation de la distance du locuteur est essentiellement basée sur la perception de l'effort vocal. Si ce dernier pouvait être quantifié et mesuré aisément (on pourrait imaginer un « effort-mètre », fonctionnant par le biais de multiples électrodes, qui mesure des paramètres physiologiques liés à l'effort du locuteur), notre tâche serait grandement simplifiée ! Il suffirait alors d'étudier la relation « distance – effort vocal » d'une part et d'autre part « effort vocal – modification des paramètres de la voix ». Mais, dans le cadre de cette thèse, nous ne disposons malheureusement pas d'un tel « effort-mètre ». C'est pourquoi nous sommes contraints à étudier directement la relation « distance – modifications des paramètres de la voix ». Sachant que l'interprétation de cette relation passe obligatoirement par la notion de « l'effort vocal », son absence (d'un point de vue « mesure ») ne facilite pas notre travail. C'est l'un des obstacles majeurs de cette étude.

Nous avons alors cherché à travers la littérature les éléments les plus pertinents quant à la perception de l'effort vocal ou de la distance de communication. Nous complétons également cet état de l'art par une étude permettant de faire notre propre opinion sur la pertinence des paramètres à modifier. Il s'agit de comparer la part d'information fourni par les paramètres prosodiques ou par les paramètres spectraux (formants, pente spectrale, ...) dans la perception de l'effort vocal d'un locuteur. Un algorithme de *matching* permettant de « greffer » sur une voix parlée, soit les paramètres prosodiques soit les paramètres spectraux d'une voix créée a été développé. Cette manipulation nous a permis de juger séparément de l'apport de la prosodie ou des paramètres spectraux dans la perception de l'effort vocal. Ces deux points nous permettent d'affirmer que la prosodie, par opposition aux modifications spectrales, constitue l'élément le plus pertinent pour la perception de l'effort vocal. Notre étude s'est alors concentrée sur la modélisation de la prosodie de la voix créée dans une communication à distance.

Ainsi, pour mieux caractériser les effets de l'effort vocal dans la voix, nous avons opté pour une démarche visant à réaliser une analyse systématique de corpus de voix en communication à distance. L'élaboration de ces corpus a fait l'objet d'une attention particulière afin qu'à travers la distance de communication l'effort vocal puisse être étudié. A l'aide de ces différents corpus, dont certains ont été enregistrés en utilisant un protocole de simulation de communication à distance réalisable en laboratoire dans des conditions idéales, nous avons alors effectué une analyse précise de la prosodie et notamment la prosodie de logatomes. L'étude prosodique des phrases créées nous est rapidement apparue comme étant trop ambitieuse dans notre étude. Des réorganisations prosodiques complexes ont été identifiées en voix créée qui ne permettent pas de mettre directement en place des règles de

transformation, et qui de plus, ne permettent pas d'interpréter facilement les variations observées. C'est pourquoi, nous nous sommes principalement basés sur l'analyse de logatomes CV, CVC, VCV et CVCV.

Grâce à la littérature et surtout grâce à nos propres analyses, nous avons élaboré un jeu de règles permettant de transformer une voix parlée en une voix criée reflétant l'effort vocal (donc la distance). Pour ce faire, nous modélisons dans un premier temps, l'élément le plus pertinent : le contour de F0. Ensuite des règles sur la modification de l'intensité, de la durée, de la pente spectrale mais également de la position du premier formant ont été proposées. Pour le moment, lors de ces transformations nous ne tenons pas compte des modifications de qualité de voix (i.e. des paramètres de l'onde de débit glottique). Nous avons fait le choix, pour une première approche, de ne considérer que les éléments les plus pertinents et notamment le contour de F0. Nous n'ignorons cependant pas la contribution globale des paramètres spectraux. En effet, leur apport n'est pas à négliger du fait que l'effort vocal ne se reflète qu'à travers un tout. D'autre part ces modifications permettraient d'augmenter encore la qualité des transformations, de rendre plus naturel ces transformations mais permettraient également d'accroître la sensation d'effort vocal.

Pour les voix chuchotées, la littérature offre un large aperçu de différentes caractéristiques de celle-ci. On retrouve également dans la littérature des méthodes de transformation de voix parlée en voix chuchotée. La plupart de ces études n'effectuent pourtant qu'un dévoisement du signal de la voix parlée. Cependant, cette unique modification n'apporte pas la qualité que nous espérons. C'est pourquoi nous proposons également un jeu de règles de modification de la voix parlée pour obtenir une voix chuchotée. Nous complétons notamment la technique du dévoisement, par l'utilisation de règles de modification des formants, de la pente spectrale, ainsi que de la dynamique de l'intensité. D'autre part, plutôt que d'utiliser, comme la majorité des études le préconise, un bruit blanc afin de reconstruire le signal de la parole chuchotée, nous avons choisi d'utiliser un bruit caractéristique du bruit de constriction. Nous estimons que l'ensemble de ces modifications, améliorent considérablement la qualité, le naturel et l'intelligibilité des voix chuchotées transformées.

Les différents tests perceptifs réalisés au cours de cette thèse nous ont permis de valider les règles de transformation que nous proposons. De plus, la modélisation du contour de F0 proposée semble être l'élément le plus important dans cet objectif. En effet, un simple décalage du contour de F0, pour la transformation de voix parlée en voix criée, n'est pas suffisant. De plus les précautions que nous avons prises pour réaliser les algorithmes de transformation permettent d'obtenir des taux d'intelligibilité très raisonnables.

Dans l'ensemble, à travers ce travail nous avons voulu mieux décrire les voix chuchotées et les voix criées dans l'objectif de réaliser des transformations de la nature de la voix. Nous avons, pour la transformation en voix chuchotée, proposé des améliorations par rapport aux méthodes de

transformation observées dans la littérature. Pour les transformations de voix parlée en voix criée, face à la difficulté de la tâche, nous avons mis en évidence les paramètres les plus importants pour la perception de l'effort vocal. Nous avons privilégié la modification de la prosodie dans cette tâche.

Rappelons enfin que, bien que la modification de voix parlée en voix chuchotée puisse être appliquée sur n'importe quel type d'énoncé, et peut également être applicable en temps réel, les règles de transformation proposées pour les voix criées ne sont applicables que sur des mots courts et notamment sur les structures que nous avons considérées : CV, CVC, VCV et CVCV.

## 12.1 Perspectives de l'étude

---

L'un des objectifs de la thèse est d'apporter des éléments nouveaux permettant d'orienter les recherches futures sur la voix criée et notamment pour la transformation de voix modale en voix criée. C'est pourquoi les perspectives sont nombreuses. Étant donné que nous ne pouvons pas réaliser dans une seule thèse toutes les analyses nécessaires à la transformation de voix parlée en voix criée tenant compte de tous les phénomènes mentionnés dans la littérature, les perspectives complémentaires sont également envisagées.

Rappelons que pour le cas des transformations en voix criée, nous n'avons pas fait le choix d'appliquer l'ensemble des éléments mentionnés dans la littérature. Ainsi, il serait judicieux, maintenant que nos algorithmes et règles de transformation ont été validés, de modifier l'algorithme proposé en incluant des modifications liées à la qualité de voix. Pour ce faire, l'utilisation de méthodes d'analyse-synthèse plus évoluées est nécessaire. On pourrait alors utiliser des méthodes telles que HNM, LF-ARX ou encore STRAIGHT. Ceci permettrait alors d'apporter les modifications sur les paramètres glottiques que nous ne prenons pas en compte dans cette étude (rapport entre la partie périodique et stochastique de la source, paramètres de la source glottique via le modèle LF, ...).

Les modifications prosodiques appliquées ici, et notamment la modélisation du contour de F0 ne sont pas effectuées à partir des modèles prosodiques connus dans la littérature (type Fujisaki (Fujisaki et Sudo, 1971) ou TILT (Taylor, 2000) ). Pour une première approche, nous considérons le choix d'un modèle de prosodie trop complexe. Nous avons ainsi fait le choix de décrire les contours de F0 en utilisant plusieurs points caractéristiques. Toutefois, il serait intéressant d'utiliser ce type de modèle. En effet, ces derniers offrent généralement l'avantage de générer des contours prosodiques qui paraissent très naturels.

Remarquons également que la modification de la position des formants que nous employons, et notamment dans le cas de la transformation en voix criée, est discutable. En effet, les mécanismes de déplacement des formants en voix criée sont mal connus et plusieurs théories tentent d'expliquer ce

phénomène. Cependant à ce jour, aucune de ces théories ne fait l'unanimité des chercheurs. Cette difficulté quant à la modélisation du déplacement des formants, peut être largement attribuée à la difficulté de mesure des formants pour des voix criées. Le déplacement des formants pour la transformation de voix criée semble pourtant être un élément important permettant d'aboutir à une voix claire et d'améliorer l'intelligibilité. Ainsi une modélisation précise de ces phénomènes semble être indispensable. La problématique de la mesure des formants en voix criée, ou plus généralement en voix avec une F0 élevée, devraient être plus étudiée afin de permettre de quantifier avec précision le déplacement de ces derniers.

Notre travail actuel ne concerne que des mots courts de type CV, CVC, VCV et CVCV. Par ces structures nous avons identifié plusieurs faits marquants et partagés par tous les locuteurs. Néanmoins, nous n'avons pas pu réaliser d'étude précise sur les phrases. Ainsi, de futurs travaux seraient les bienvenues afin de confirmer nos différentes hypothèses et analyses que nous présentons ici. Ceci permettrait également d'ajuster nos règles pour la transformation de phrases.

Nous avons mis en évidence des phénomènes pour les voix criées sur des logatomes enregistrés par lecture de listes. Il serait intéressant de confirmer ces résultats sur des vrais mots dans une situation de communication spontanée. D'autre part, il serait également intéressant de réaliser ce type d'analyse sur un nombre plus élevée de locuteurs permettant de renforcer les hypothèses avancées dans ce travail. De plus, nous n'avons étudié que des voix masculines, mais qu'en est-il de ces modifications prosodiques pour des voix féminines ?

Enfin un seul niveau d'effort a été créé et testé dans cette étude. Nous mentionnons toutefois des pistes quant à l'ajustement du niveau d'effort vocal lors de la transformation. Il serait alors très intéressant de réaliser des tests perceptifs de la distance en utilisant plusieurs niveaux d'effort afin de déterminer si un continuum de la perception de la distance peut être apporté par l'utilisation de nos règles de transformation. Ces tests pourront également être complétés par des tests de qualité de transformation ainsi que de naturel. Toutefois, pour réaliser ce type de tests, des analyses plus précises sur l'évolution des paramètres que nous utilisons dans nos transformation sont nécessaires (dynamique d'intensité, dynamique de F0, déclinaison de l'intensité et de F0, allongement de la durée des voyelles, ...).

## 12.2 D'autres champs d'application possibles

---

En plus de l'application directe que nous destinons à nos recherches, nous pensons que cette étude permet également d'apporter des éléments précieux dans d'autres domaines d'applications.

En premier lieu, une meilleure connaissance des phénomènes liés à l'augmentation de l'effort vocal peut avoir un intérêt non négligeable dans les systèmes de reconnaissance vocale. En effet les

systèmes à commande vocale, de reconnaissance vocale ou de reconnaissance du locuteur ne sont plus aussi efficaces si la voix utilisée est une voix criée ou chuchotée (cf. (Ismail, 2010) pour une discussion sur le sujet). Ainsi notre étude permet d'apporter plus de précisions quant aux modifications observables entre une voix parlée et criée (ou chuchotée) permettant éventuellement de compléter ces systèmes.

D'autres applications peuvent être les jeux vidéo ou le cinéma. De plus en plus de concepteurs de jeux vidéo utilisent la conversion TTS (*Text To Speech*) dans le but d'amoindrir leurs coûts. Cette solution ne demande ni de louer un studio d'enregistrement, ni de payer des acteurs pour enregistrer les voix des différents personnages de leurs jeux. Le premier essai a d'ailleurs été réalisé en 1980 dans le jeu d'arcade « shoot'em up »<sup>1</sup>. Bien que ces techniques soient efficaces, la restitution des émotions ou des différents états de la voix reste encore un problème majeur. Ainsi une meilleure compréhension des voix utilisées pour la communication à distance, et par ce biais l'effort vocal, permettrait d'apporter une pierre à cet édifice.

Un autre intérêt pour les jeux vidéo est la communication orale entre joueur en ligne. Depuis peu, on trouve des logiciels qui permettent de mieux gérer les discussions entre les joueurs. Le projet VIVOX entre autre, permet de représenter en 3 dimensions la voix de son équipier en fonction de la position de celui-ci dans le jeu. Ce projet permet également de modifier sa voix en temps réel. On peut par exemple choisir d'avoir une voix féminine si on est un homme et vice et versa. Bien que cette application inclue une représentation de la distance des joueurs, la gestion de la distance est faite uniquement sur la base du niveau sonore. Plus le joueur s'éloignera, « virtuellement », plus sa voix sera faible jusqu'à n'être plus audible, si le joueur se situe trop loin. Ainsi, une transformation de l'effort vocal permettrait d'apporter une dimension supplémentaire à ce type de discussion en ligne. La distance d'un coéquipier pourrait alors être indiquée par modification de l'effort vocal produit par ce dernier, comme s'il s'agissait d'un monde réel.

Une autre application serait également la reconnaissance de cri d'alarme. En effet, depuis quelques années, des recherches sont menés dans le but d'identifier des voix criées afin de détecter des situations d'urgence (Nanjo et al., 2009). Ces applications permettraient de déclencher automatiquement des systèmes d'alarme et de sécurité. Une meilleure connaissance des voix criées permettrait alors de compléter ces systèmes dans le but de diminuer le nombre de fausses alarmes.

La liste des applications présentées ici n'est pas exhaustive et nous pensons que cette étude peut encore avoir un intérêt dans bien d'autres domaines, même si l'on sait que le chemin pour parvenir à une parfaite compréhension de l'effort vocal, permettant de compléter les applications que nous mentionnons est encore long.

---

<sup>1</sup> Source : [http://en.wikipedia.org/wiki/Speech\\_synthesis](http://en.wikipedia.org/wiki/Speech_synthesis)

# TABLE DES MATIÈRES

Remerciements .....	5
Sommaire .....	7
Abréviations.....	9
Conventions .....	10
<b>CHAPITRE 1 Introduction générale .....</b>	<b>11</b>
1.1 Contexte .....	11
1.2 La problématique .....	15
1.3 Orientation de la thèse et méthodologie.....	16
1.3.1 Enregistrements de corpus.....	17
1.3.2 Analyse du corpus et élaboration de règles de transformation .....	17
1.3.3 Modification de la voix parlée.....	18
1.3.4 Approche retenue .....	19
1.4 Restriction de l'étude .....	20
1.5 Organisation du document.....	21
<b>PARTIE I : FONDAMENTAUX ET ACQUIS.....</b>	<b>23</b>
<b>CHAPITRE 2 Perception auditive de distance (PAD).....</b>	<b>25</b>
2.1 PAD liés par les phénomènes acoustiques .....	26
2.1.1 L'intensité.....	26
2.1.2 Les variations spectrales.....	29
2.1.3 Rapport de l'énergie direct à l'énergie réverbérée.....	34
2.2 Autres phénomènes intervenant dans la PAD .....	35
2.3 PAD par la voix du locuteur .....	36
2.4 Conclusion du chapitre 2 .....	40
<b>CHAPITRE 3 La parole : production, représentation et variabilité.....</b>	<b>43</b>
3.1 Anatomie et physiologie de l'appareil phonatoire .....	43
3.1.1 L'appareil respiratoire .....	45
3.1.2 Le larynx .....	45
3.1.3 Le conduit vocal .....	49
3.2 Théorie acoustique de la production de la parole.....	50
3.3 Modélisation acoustique de la parole .....	54
3.3.1 La théorie source-filtre .....	54
3.3.2 Modélisation acoustique du conduit vocal .....	57
3.3.3 Modélisation de la source périodique : la vibration des cordes vocales .....	57
3.3.4 Modélisation de la source apériodique : turbulence causée par l'air .....	60
3.3.5 Modélisation du rayonnement aux lèvres .....	61
3.4 La parole : information et variabilité.....	63
3.4.1 Théorie de l'information.....	63
3.4.2 Conditions adverses : cas de la communication parlée.....	65
3.4.3 Comportement et adaptation.....	65
3.5 Paramètres paralinguistiques .....	67
3.5.1 La prosodie <sup>1</sup> .....	68
3.5.2 La qualité de voix .....	68
<b>CHAPITRE 4 Effort vocal et effort vocal.....</b>	<b>71</b>
4.1 Causes et circonstances .....	71
4.2 Effort vocal : définition nous concernant.....	73
4.3 La mesure de l'effort vocal.....	76

<b>CHAPITRE 5 Caractéristiques des voix en communication à distance : état de l'art.....</b>	<b>79</b>
5.1 <i>État de l'art des connaissances actuelles sur la voix chuchotée</i> .....	80
5.1.1 Le niveau glottique .....	81
5.1.2 Le niveau acoustique .....	81
5.1.3 Le niveau prosodique .....	84
5.2 <i>État de l'art des connaissances actuelles sur la voix criée</i> .....	85
5.2.1 Niveau acoustique .....	85
5.2.2 Niveau glottique .....	86
5.2.3 Niveau articulatoire .....	89
5.2.4 Niveau spectral .....	89
5.2.5 Prosodie .....	94
5.2.6 Qualité de voix .....	96
5.3 <i>Identification du locuteur et intelligibilité</i> .....	96
5.3.1 Intelligibilité de la voix criée.....	96
5.3.2 Intelligibilité de la voix chuchotée .....	97
5.3.3 Reconnaissance du locuteur.....	98
5.4 <i>Conclusion du chapitre 5</i> .....	98
<b>PARTIE II : CARACTÉRISATION DES VOIX CRIÉES .....</b>	<b>101</b>
<b>CHAPITRE 6 Élaboration des bases de données.....</b>	<b>103</b>
6.1 <i>Protocole d'enregistrement en champ libre : DB1</i> .....	104
6.1.1 Protocole d'enregistrement.....	104
6.1.2 Remarques.....	107
6.2 <i>Protocole d'enregistrement en champ libre version améliorée : DB2</i> .....	107
6.2.1 Protocole d'enregistrement.....	108
6.2.2 Remarques.....	110
6.3 <i>Protocole de simulation de communication orale à distance : DB3</i> .....	112
6.3.1 Protocole d'enregistrement.....	112
6.3.2 Remarques et observations .....	114
6.4 <i>Corpus dédié à l'analyse micro-prosodique : DB4</i> .....	115
6.4.1 Protocole d'enregistrement.....	115
6.4.2 Remarques et observations .....	116
6.5 <i>Conclusion du chapitre 6</i> .....	117
<b>CHAPITRE 7 Perception de l'effort vocal : Prosodie ou paramètres spectraux ? .....</b>	<b>119</b>
7.1 <i>État de l'art sur les paramètres pertinents de l'effort vocal</i> .....	120
7.1.1 Éléments pertinents pour la perception des voix chuchotées.....	121
7.1.2 Éléments pertinents pour la perception des voix criées .....	122
7.2 <i>Le rôle de la prosodie dans la perception auditive de l'effort vocal</i> .....	127
7.2.1 La base de voix utilisée .....	128
7.2.2 Méthodologie.....	129
7.2.3 Dynamic Time Warping (DTW) .....	129
7.2.4 Génération des signaux tests.....	131
7.2.5 Test perceptif.....	133
7.2.6 Résultats .....	134
7.2.7 Discussion .....	135
7.3 <i>Conclusions du chapitre 7</i> .....	135
<b>CHAPITRE 8 Dynamique des paramètres et effort vocal .....</b>	<b>137</b>
8.1 <i>Analyses de l'intensité et de la F0 globales</i> .....	138
8.1.1 Intensité globale .....	138
8.1.2 Fréquence fondamentale.....	140
8.2 <i>Dynamique des paramètres vs effort vocal</i> .....	141
8.2.1 Méthode proposée pour la mesure de la dynamique.....	142
8.2.2 Analyses prosodique de DB2 .....	143
8.2.3 Évaluation du protocole de simulation (DB3) par comparaisons avec le corpus enregistré en champ libre de jour (DB1) .....	148
8.3 <i>Dynamique des formants</i> .....	152
8.4 <i>Conclusion du chapitre 8</i> .....	156

<b>CHAPITRE 9 La prosodie et la micro- prosodie de l'effort vocal.....</b>	<b>159</b>
9.1 <i>Analyses de la durée</i> .....	160
9.1.1 Analyse de la variation de durée totale entre les structures étudiées .....	161
9.1.2 Analyse de la variation de durée des consonnes et des voyelles .....	163
9.1.3 Rapport entre la durée des voyelles pour les logatomes VCV et CVCV .....	166
9.1.4 Résumé .....	168
9.2 <i>Analyses de l'intensité des logatomes</i> .....	169
9.2.1 Variations relatives de l'intensité des consonnes et des voyelles pour les logatomes CV....	169
9.2.2 Variations relatives de l'intensité des consonnes et des voyelles pour les logatomes CVC.	172
9.2.3 Variations relatives de l'intensité des consonnes et des voyelles pour les logatomes VCV.	173
9.2.4 Variations relatives de l'intensité des consonnes et des voyelles pour logatomes CVCV ...	177
9.2.5 Résumé .....	179
9.3 <i>Analyses de la fréquence fondamentale des logatomes</i> .....	179
9.3.1 Les variations relatives de F0 des consonnes et des voyelles pour les logatomes CV .....	179
9.3.2 Les variations relatives de F0 des consonnes et des voyelles pour les logatomes CVC.....	183
9.3.3 Les variations relatives de F0 des consonnes et des voyelles pour les logatomes VCV.....	184
9.3.4 Les variations relatives de F0 des consonnes et des voyelles pour les logatomes CVCV ...	188
9.3.5 Résumé .....	189
9.4 <i>Micro-mélorie des énoncés CV, CVC, VCV et CVCV</i> .....	190
9.4.1 Les contours mélodiques des logatomes CV .....	190
9.4.2 Les contours mélodiques des logatomes CVC .....	195
9.4.3 Les contours mélodiques des logatomes VCV .....	199
9.4.4 Les contours mélodiques des logatomes CVCV.....	204
9.4.5 Résumé .....	207
9.5 <i>Interprétations hypothétiques</i> .....	210
9.5.1 Hypothèse 1 .....	210
9.5.2 Hypothèse 2 .....	211
9.5.3 Hypothèse 3 .....	211
9.6 <i>Conclusions du chapitre 9</i> .....	213
<b>PARTIE III TRANSFORMATION .....</b>	<b>215</b>
<b>CHAPITRE 10 Transformation de la voix.....</b>	<b>217</b>
10.1 <i>Choix de la méthode d'analyse-synthèse</i> .....	218
10.1.1 Modifications dans le domaine temporel : TD-PSOLA .....	219
10.1.2 Modifications dans le domaine temporel complétées d'une représentation source filtre : LP-PSOLA et RELP-PSOLA .....	220
10.1.3 Modifications dans le domaine temporel complétées par d'une représentation source filtre et avec un ré-échantillonnage du résidu .....	221
10.2 <i>Description des procédures d'analyse</i> .....	222
10.2.1 Marquage du pitch.....	223
10.2.2 Mesure et modification des formants .....	225
10.3 <i>Transformation de voix modale vers une voix chuchotée</i> .....	227
10.3.1 Élaboration des règles de transformation .....	227
10.3.2 Réalisation de l'algorithme de transformation de voix modale en voix chuchotée .....	233
10.3.3 Remarques.....	236
10.4 <i>Transformation de voix modale en voix criée</i> .....	236
10.4.1 Élaboration des règles de transformation .....	237
10.4.2 Validité des règles de transformation proposées .....	244
10.4.3 Réalisation des transformations de voix parlée en voix criée .....	248
10.4.4 Réflexions quant à la transformation de plusieurs niveaux d'effort vocal : .....	250
10.5 <i>Conclusion du chapitre 10</i> .....	251
<b>CHAPITRE 11 Évaluation des transformations.....</b>	<b>253</b>
11.1 <i>Tests d'intelligibilité</i> .....	254
11.1.1 Caractéristiques des signaux tests .....	254
11.1.2 Description du protocole de test d'intelligibilité .....	255
11.1.3 Résultats des tests d'intelligibilité.....	256
11.1.4 Conclusion des tests d'intelligibilité .....	266
11.2 <i>Perception de distance</i> .....	267

11.2.1	Choix des signaux tests .....	267
11.2.2	Règles utilisées pour générer les signaux tests.....	268
11.2.3	Description du test de perception de la distance .....	269
11.2.4	Résultats du test de perception de distance .....	270
11.3	<i>Conclusions du chapitre 11</i> .....	275
<b>CHAPITRE 12 Conclusions générales et perspective .....</b>		<b>277</b>
12.1	<i>Perspectives de l'étude</i> .....	280
12.2	<i>D'autres champs d'application possibles</i> .....	281
<b>Table des matières.....</b>		<b>283</b>
<b>Références.....</b>		<b>287</b>
<b>ANNEXES .....</b>		<b>309</b>
<b>Annexe A</b>	<b>Listes des logatomes CV, CVC, VCV ; CVCV .....</b>	<b>311</b>
A.1	<i>Liste des logatomes pour la voyelle [a] :</i> .....	311
A.2	<i>Liste des logatomes pour la voyelle [i] :</i> .....	312
A.3	<i>Liste des logatomes pour la voyelle [u] :</i> .....	312
<b>Annexe B</b>	<b>Mesure du quotient ouvert pour le corpus DB3.....</b>	<b>313</b>
<b>Annexe C</b>	<b>Matrice de confusion des tests d'intelligibilité pour les voyelles .....</b>	<b>315</b>
C.1	<i>Intelligibilité des voyelles en voix parlée</i> .....	315
C.2	<i>Intelligibilité des consonnes en voix chuchotée</i> .....	315
C.3	<i>Intelligibilité des consonnes en voix criée</i> .....	316
C.4	<i>Intelligibilité des consonnes en voix transformée chuchotée</i> .....	317
C.5	<i>Intelligibilité des consonnes en voix transformée criée</i> .....	317
<b>Annexe D</b>	<b>Matrice de confusion des tests d'intelligibilité en fonction des locuteurs .....</b>	<b>319</b>
D.1	<i>Locuteur 1</i> .....	319
D.2	<i>Locuteur 2</i> .....	326
D.3	<i>Locuteur 3</i> .....	333
<b>Annexe E</b>	<b>Valeurs caractéristiques des contours de F0 des logatomes.....</b>	<b>341</b>
E.1	<i>Les logatomes CV</i> .....	342
E.2	<i>Les logatomes CVC</i> .....	345
E.3	<i>Les logatomes VCV</i> .....	348
E.4	<i>Les logatomes CVCV</i> .....	351
<b>Annexe F</b>	<b>Résultats des tests de perception de la distance pour chaque sujet .....</b>	<b>355</b>
F.1	<i>Ordre de passage des séries en fonction des sujets</i> .....	355
F.2	<i>Résultats de chaque locuteur en fonction de la série pour les logatomes CVC</i> .....	356
F.3	<i>Résultats de chaque locuteur en fonction de la série pour les logatomes CVCV</i> .....	357
<b>Résumé.....</b>		<b>358</b>

# RÉFÉRENCES

- ACKER, B. F. (1987). "Vocal tract adjustments for the projected voice," *Journal of Voice* **1**, 77–82.
- AGIOMYRGIANNAKIS, Y., & ROSEC, O. (2009). "ARX-LF-based source-filter methods for voice modification and transformation," *International Conference on Acoustics, Speech and Signal Processing* pp. 3589–3592.
- ALKU, P. (2002). "Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation," *Speech Communication* **38**, 321–334.
- ANDERSSON, A., ERIKSSON, A., & TRAUNMÜLLER, H. (1996). "Cries and whispers: Acoustic effects of variations in vocal effort," *TMH-QPSR* **37**, 127–130.
- ARRABITO, G. R. (2000). "An evaluation of three-dimensional audio displays for use in military environments," *Canadian acoustics* **28**, 5–14.
- ASHMEAD, D. H., DEFORT, L., & ODOM, R. D. (1990). "Perception of the relative distances of nearby sound sources," *Perception & psychophysics* **47**, 326–331.
- ATAL, B. S., HANAVER, S. L., & HANAUER, S. L. (1971). "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *Journal of the Acoustical Society of America* **50**, 637–655.
- ATAL, B. S., & SCHROEDER, M. (1974). "Recent advances in predictive coding – applications to speech synthesis," In Fant (Ed.), *Speech Communication* (Almqvist and Wiksell), pp. 27 – 31.
- AUBERGÉ, V., & RILLIARD, A. (2006). "Le focus prosodique n'est pas que déictique: le modèle VID (Valence-Intensité-Domaine)," *Actes des 26èmes Journées d'étude sur la parole (JEP)* (Dinard, France).
- AUDIBERT, N. (2008). "*Prosodie de la parole expressive : dimensionnalité d'énoncés méthodologiquement contrôlés authentiques et actés*", (Thèse de Doctorat), Université de Grenoble, INP.
- BAILLY, L. (2009). "*Interaction entre cordes vocales et bandes ventriculaires en phonation: exploration in-vivo, modélisation physique, validation in-vitro*", (Thèse de Doctorat), Université du Maine.
- BAILLY, L., & HENRICH, N. (2010). "Contribution des bandes ventriculaires lors d'un effort vocal Impact sur la vibration glottique," *Actes du 10ème Congrès Français d'Acoustique (CFA)* (Lyon, France).

- BAILLY, L., PELORSON, X., HENRICH, N., & RUTY, N. (2008). "Influence of a Constriction in the near Field of the Vocal Folds: Physical Modeling and Experimental Validation," *Journal of the Acoustic Society of America* **124**, 3296–3308.
- BARNEY, A., DE STEFANO, A., & HENRICH, N. (2007). "The Effect of Glottal Opening on the Acoustic Response of the Vocal Tract," *Acta Acustica united with Acustica* **93**, 1046–1056.
- BASS, H. E. (1990). "Atmospheric absorption of sound: Update," *Journal of the Acoustical Society of America* **88**, 2019–2021.
- BEGAULT, D. R. (1991). "Preferred sound intensity increase for sensation of half distance," *Perceptual and motor skills* **72**, 1019–1029.
- BELE, I. V. (2006). "The speaker's formant," *Journal of voice* **20**, 555–578.
- BLATCHFORD, H., & FOULKES, P. (2006). "Identification of voices in shouting," *Journal of speech, language and the law* **16**, 241–254.
- BLAUERT, J. (1997). *Spatial Hearing* (J. Blauert, Ed.) (MIT Press), Vol. 2.
- BOERSMA, P. (1993). "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *IFA Proceedings* **17**, 97–110.
- BOITE, R. (2000). *Traitement de la parole* (Presses polytechniques et universitaires romandes).
- BOND, Z., & MOORE, T. (1990). "A note on loud and lombard speech," *Proc. of the 1st International Conference on Spoken Language Processing (ICSLP)* (Kobe, Japan), pp. 969–972.
- BONNOT, J.-F., & CHEVRIE-MULLER, C. (1991). "Some effects of shouted and whispered conditions on temporal organization," *Journal of Phonetics* **19**, 473–483.
- BOTTE, M. C., CANÉVET, G., DEMANY, L., & SORIN, C. (1990). "Identification de la distance," In *Tech et Doc* (Eds.), *Psychoacoustique et perception auditive*, pp. 100–102.
- BOU-GHAZALE, S. E., & HANSEN, J. H. L. (1995). "A Source Generator Based Modeling Framework for Synthesis of Speech Under Stress," *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Detroit, Michigan, USA), pp. 664–667.
- BOU-GHAZALE, S. E., & HANSEN, J. H. L. (1996a). "Synthesis of stressed speech from isolated neutral speech using HMM-based models," *Proc. of the 4th International Conference on Spoken Language Processing (ICSLP)* **3**, 1860–1863.
- BOU-GHAZALE, S. E., & HANSEN, J. H. L. (1996b). "Generating stressed speech from neutral speech using a modified CELP vocoder," *Speech Communication* **20**, 93–110.

- BOU-GHAZALE, S. E., & HANSEN, J. H. L. (1998). "HMM-based stressed speech modeling with application to improved synthesis and recognition of isolated speech under stress," *IEEE Transactions On Speech And Audio Processing* **6**, 201–2016.
- BOUZID, A., & ELLOUZE, N. (2005). "Glottal Closure Instant Detection from Speech Signal by Multiscale Product," *Proc. of the 17th IMACS world congress (Paris, France)*.
- BRONKHORST, A. W., VELTMAN, J. A., & VAN BREDA, L. (1996). "Application of a three-dimensional auditory display in a flight task," *Human factors* **38**, 23–33.
- BRUNGART, D. S. (1999). "Auditory localization of nearby sources III Stimulus effects," *Journal of the Acoustical Society of America* **106**, 3589–3602.
- BRUNGART, D. S., DURLACH, N. I., & RABINOWITZ, W. M. (1999). "Auditory localization of nearby sources II Localization of a broadband source," *Journal of the Acoustical Society of America* **106**, 1956–1968.
- BRUNGART, D. S., KORDIK, A., DAS, K., & SHAW, A. (2002). "The Effects of F0 Manipulation on the Perceived Distance of Speech," *Proc. of the 7th International Conference on Spoken Language Processing (ICSLP) (Denver, CO, USA)*, pp. 1641–1644.
- BRUNGART, D. S., & RABINOWITZ, W. M. (1999). "Auditory localization of nearby sources Head-related transfer functions," *Journal of the Acoustical Society of America* **106**, 1465–1479.
- BRUNGART, D. S., & SCOTT, K. R. (2000). "Auditory distance perception of speech: the influence of production level," *Nasa Tech briefs*.
- BRUNGART, D. S., & SCOTT, K. R. (2001). "The effects of production and presentation level on the auditory distance perception of speech," *Journal of the Acoustical Society of America* **110**, 425–440.
- BRUNGART, D. S., SCOTT, K. R., & SIMPSON, B. D. (2001). "The influence of vocal effort on human speaker identification," *Proc. of Interspeech (Aalborg, Denmark)*, pp. 747–750.
- BURKHARDT, F. (2009). "Rule-based voice quality variation with formant synthesis," *Proc. of Interspeech (Brighton, UK)*, pp. 2659–2662.
- BURNETT, T. A. (1998). "Voice F0 responses to manipulations in pitch feedback," *Journal of the Acoustical Society of America* **103**, 3153–3161.
- BUTLER, R. A., LEVY, E. T., & NEFF, W. D. (1980). "Apparent distance of sounds recorded in echoic and anechoic chambers," *Journal of Experimental Psychology: Human Perception and Performance* **6**, 745–750.

- BÅVEGÅRD, M., & FANT, G. (1994). "Notes on glottal source interaction ripple," *STL-QPSR* **4**, 63–78.
- VON BÉKÉSY, G. (1938). "Über die Entstehung des Entfernungsempfindung beim Hören, (On the origin of the sensation of distance in hearing)," *Akustische zeitschrift* **3**, 21–31.
- VON BÉKÉSY, G. (1949). "The moon illusion and similar auditory phenomena," *The American journal of psychology* **62**, 540–552.
- CALLIOPE (FIRM) (1989). *La Parole et son traitement automatique* (Paris), Masson.
- CALZADA, À., & SOCORÓ, J. C. (2011). "Vocal effort modification through harmonics plus noise model representation," *Proc. of the 5th International Conference on Advances in Nonlinear Speech Processing (NoLISP)* (Las Palmas de Gran Canaria, Spain), pp. 96–103.
- CAMPBELL, N., & MOKHTARI, P. (2003). "Voice quality: the 4th prosodic dimension," *Proc. of the 5th International Congress of Phonetic Sciences (ICPhS)* (Barcelona, Spain), pp. 2417–2420.
- CHARPENTIER, F., & STELLA, M. (1986). "Diphone Synthesis Using an Overlap-Add Technique for Speech Waveforms Concatenation," *Proc. Int. Conf. on Audio, Speech and Signal Proc.* (Tokyo).
- CHENG, C. I., & WAKEFIELD, G. H. (2001). "Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency, and Space," *Journal of the Audio Engineering Society* **49**, 231–249.
- DE CHEVEIGNÉ, A., & KAWAHARA, H. (2002). "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America* **111**, 1917–1930.
- CHEYNE, H. A., KALGAONKAR, K., CLEMENTS, M. A., & ZUREK, P. (2009). "Talker-to-listener distance effects on speech production and perception," *Journal of the Acoustical Society of America* **126**, 2052–2060.
- CHIEN, C. (1975). "Sound propagation along an impedance plane," *Journal of Sound and Vibration* **43**, 9–20.
- CHILDERS, D. G., & WONG, C. F. (1994). "Measuring and modeling vocal source-tract interaction," *IEEE transactions on bio-medical engineering* **41**, 663–671.
- CHLÁDKOVÁ, K., BOERSMA, P., & PODLIPSKÝ, V. J. (2009). "On-line formant shifting as a function of F<sub>0</sub>," *Proc. of Interspeech* (Brighthon, UK), pp. 464–467.
- COCHRAN, P., THROOP, J., & SIMPSON, W. E. (1968). "Estimation of distance of a source of sound," *The American journal of psychology* **81**, 198–206.
- COLEMAN, P. D. (1962). "Failure to localize the source distance of an unfamiliar sound," *Journal of the Acoustical Society of America* **34**, 345–346.

- COLEMAN, P. D. (1963). "An analysis of cues to auditory depth perception in free space," *Psychological bulletin* **60**, 302–315.
- COLEMAN, P. D. (1968). "Dual role of frequency spectrum in determination of auditory distance," *Journal of the Acoustical Society of America* **44**, 631–632.
- COMBESCORE, P. (1981). "20 listes de dix phrases phonétiquement équilibrées," *Revue d'Acoustique* **54**, 34–38.
- DI CRISTO, A. (1982). *Prolégomènes à l'étude de l'intonation - micromélogie* (CNRS, Paris).
- DI CRISTO, A. (2000). "Interpréter la prosodie," Actes des 23èmes journées d'étude sur la parole (JEP) (Aussois, France), pp. 13–29.
- CUMMINGS, K. E. (1995). "Glottal Models for Digital Speech Processing: A Historical Survey and New Results," *Digital Signal Processing* **5**, 21–42.
- CUMMINGS, K. E., & CLEMENTS, M. A. (1995). "Analysis of the glottal excitation of emotionally styled and stressed speech," *Journal of the Acoustical Society of America* **98**, 88–98.
- DEGOTTEX, G. (2010). "*Glottal source and vocal-tract separation*", (Ph.D. Thesis), Université Pierre et Marie Curie, Paris.
- DELATTRE, P. (1966). "Les 10 intonations de base du français," *French Review* **40**, 1–14.
- VAN DEN BERG, J. W. (1958). "Myoelastic-aerodynamic theory of voice production," *Speech hearing research* **1**, 224–244.
- DIEHL, R. L., & KLUENDER, K. R. (1989). "On the Objects of Speech Perception," *Ecological Psychology* **1**, 121–144.
- DOVAL, B., D'ALESSANDRO, C., & HENRICH, N. (2006). "The spectrum of glottal flow models," *Acta Acustica* **92**, 1026–1046.
- DRUGMAN, T., BOZKURT, B., & DUTOIT, T. (2010). "Analyse et modification de la qualité vocale basée sur l'excitation," Actes des 28èmes Journées d'étude sur la parole (JEP) (Mons, Belgique), pp. 25–28.
- DRUGMAN, T., WILFART, G., & DUTOIT, T. (2009). "A Deterministic plus Stochastic Model of the Residual Signal for Improved Parametric Speech Synthesis," Proc. of the International Conference on Spoken Language Processing (ICSLP) (Brighton, UK), pp. 1779–1782.
- DURBIN, J. (1960). "The Fitting of Time-Series Models," *Review of the International Statistical Institute* **28**, 233–244.

- D'ALESSANDRO, C. (2006). "Voice source parameters and prosodic analysis," In S. Sudhoff et W. de Gruyter (Eds.), *Method in empirical prosody research*, (Walter de Gruyter), pp. 63–87.
- D'ALESSANDRO, C., & DOVAL, B. (1998). "Experiments in voice quality modification of natural speech signals: the spectral approach," *Proc of the 3rd ESCA/COCOSDA international workshop on speech synthesis* (Blue mountains , Australia), pp. 277–282.
- D'ALESSANDRO, N. (2009). "*Realtime and accurate musical control of expression in voice synthesis*", (Ph.D. Thesis), Université de Mons.
- EDGINGTON, M., & LOWRY, A. (1996). "Residual-based speech modification algorithms for text-to-speech synthesis," *Proc. of the 4th International Conference on Spoken Language Processing (ICSLP)* (Philadelphia, USA), pp. 1425–1428.
- EDWARDS, A. S. (1955). "Accuracy of auditory depth perception," *Journal of General Psychology* **52**, 327–329.
- EKLUND, I., & TRAUNMÜLLER, H. (1996). "A comparative study of the male and female whispered and phonated versions of the long vowels of Swedish," *TMH-QPSR* **37**, 131–134.
- ELLIOTT, J. (2000). "Comparing the acoustic properties of normal and shouted speech: a study in forensic phonetics," *Proc. of the International Conference on Speech Science and Technology (SST)* (Canberra, Australia), pp. 154–159.
- EMBLETON, T. F. W. (1996). "Tutorial on sound propagation outdoors," *Journal of the Acoustical Society of America* **100**, 31–48.
- EPPS, J., SMITH, J. R., & WOLFE, J. (1997). "A novel instrument to measure acoustic resonances of the vocal tract during phonation," *Measurement Science and Technology* **8**, 1112–1121.
- ERIKSSON, A., & TRAUNMÜLLER, H. (1999). "Perception Of Vocal Effort And Speaker Distance On The Basis Of Vowel Utterances," *Proc of the International Congress of Phonetic Sciences* pp. 2469–2472.
- ERIKSSON, A., & TRAUNMÜLLER, H. (2002). "Perception of vocal effort and distance from the speaker on the basis of vowel utterances," *Perception psychophysics* **64**, 131–139.
- ERTEN, G. (2005). "Method for manipulating listener perceived distance to voice source in speech signals," *Provisional patent application*, Application Number 60/678,698.
- FAIRBANKS, G. (1957). "Effects of Vocal Effort upon the Consonant-Vowel Ratio within the Syllable," *Journal of the Acoustical Society of America* **29**, 621–626.
- FANT, G. (1960). *Acoustic Theory of Speech Production* (Mouton: The Hague).

- FANT, G. (1982). "Preliminaries to analysis of the human voice source," *STL-QPSR* **4**, 1–27.
- FANT, G., LILJENCANTS, J., & LIN, Q. (1985). "A four-parameter model of glottal flow," *STL-QPSR* **4**, 1–13.
- FARNER, S., ROBEL, A., & RODET, X. (2009). "Natural transformation of type and nature of the voice for extending vocal repertoire in high-fidelity applications," Proc. of the 35th Audio Engineering Society International Conference (AES) (London, UK), pp. 1189–1198.
- FENG, G. (1986). "*Modélisation acoustique et traitement du signal de parole, le cas des voyelles nasales*", (Thèse de Doctorat), INP Grenoble.
- FENG, G. (1996). *Traitement et codage de la parole : entre une discipline et un objet pluridisciplinaire*, (Habilitation à diriger des recherches), INP Grenoble.
- FLANAGAN, J. L. (1968). "Source-system interaction in the vocal tract," *Annals of the New York Academy of Sciences* **155**, 9–17.
- FLANAGAN, J. L. (1972). *Speech Analysis, Synthesis and Perception* (Springer-Verlag, Ed.) (New-York), 2nd editio.
- FLANAGAN, J. L., & LANDGRAF, L. (1968). "Self-oscillating source for vocal-tract synthesizers," *IEEE Transactions on Audio and Electroacoustics* **16**, 57–64.
- FUJISAKI, H., & SUDO, H. (1971). "A model for the generation of fundamental frequency contours of Japanese word accent," *Japan Acoustic Society* **27**, 445–453.
- FUX, T., AUBERGÉ, V., FENG, G., & ZIMPFER, V. (2012a). "Speaker's Prosodic Strategy for a Large Physical Distance Communication Task," International conference of the Gruppo di Studi sulla Comunicazione Parlata (GSCP) (Bel Horizonte, Brazil).
- FUX, T., FENG, G., & ZIMPFER, V. (2010). "Le rôle de la prosodie dans la perception de l'effort vocal," Actes du congrès français d'acoustique (Lyon, France).
- FUX, T., FENG, G., & ZIMPFER, V. (2011a). "Talker-to-listener distance effects on the variations of the intensity and the fundamental frequency of speech," Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (Prague, Czech republic), pp. 4964 – 4967.
- FUX, T., FENG, G., & ZIMPFER, V. (2011b). "Relevant acoustic features of speech signals for natural-to-shouted voice transformation," Proc. of the Forum Acusticum, Acta Acustica united with Acustica, Vol. 97 (S1) (Aalborg, Denmark).

- FUX, T., FENG, G., & ZIMPFER, V. (2012b). "Natural-to-Shouted Voice Transformation for Distance Cues of Monosyllabic Consonant-Vowel-Consonant Words (in press)," *Acta Acustica United With Acustica*.
- FUX, T., & ZIMPFER, V. (2009). *Spatialisation du son: corrélation entre la distance perçue et l'effort vocal* (French-German Research Institute of Saint-Louis, France), ISL-RV 218.
- FÓNAGY, I., & FÓNAGY, J. (1966). "Sound Pressure Level and Duration," *Phonetica* **15**, 14–21.
- GARBER, S. R., & MOLLER, K. T. (1979). "The effects of feedback filtering on nasalization in normal and hypernasal speakers," *Journal of Speech and Hearing Research* **22**, 321–33.
- GARBER, S. R., SIEGEL, G. M., & PICK, H. L. (1981). "Regulation of Vocal Intensity in the Presence of Feedback Filtering and Amplification," *Journal of Speech and Hearing Research* **24**, 104–108.
- GARDNER, M. B. (1969). "Distance Estimation of 0° or apparent 0°-oriented speech signals in anechoic space," *Journal of the Acoustical Society of America* **45**, 47–53.
- GARNIER, M. (2007). "*Communiquer en environnement bruyant: de l'adaptation jusqu'au forçage vocal*", (Thèse de Doctorat), Université Paris 6.
- GARNIER, M. (2008). "May speech modifications in noise contribute to enhance audio-visible cues to segment perception?," *Proc. of the International Conference on Auditory-Visual Speech Processing (AVSP)* (Moreton Island, Australia), pp. 95–100.
- GARNIER, M., HENRICH, N., SMITH, J., & WOLFE, J. (2010). "Vocal tract adjustments in the high soprano range," *Journal of the Acoustical Society of America* **127**, 3771–3780.
- GARNIER, M., WOLFE, J., HENRICH, N., & SMITH, J. (2008). "Interrelationship between vocal effort and vocal tract acoustics: a pilot study," *Proc. of the International Conference on Spoken Language Processing (ICSLP)* (Brisbane, Australia).
- GAUFFIN, J., & SUNDBERG, J. (1989). "Spectral Correlates of Glottal Voice Source Waveform Characteristics," *Journal of Speech and Hearing Research* **32**, 556–565.
- GEUMANN, A. (2001a). "Vocal Intensity: Acoustic and Articulatory Correlates," *Proc. of the 4th International Speech Motor Conference* (Nijmegen, Holland), pp. 70–73.
- GEUMANN, A. (2001b). "*Invariance and variability in articulation and acoustics of natural perturbed speech*", (Ph.D. Thesis), University of Munich.
- GIOVANNI, A., OUAKNINE, M., GARREL, R., AYACHE, S., & ROBERT, D. (2002). "Un modèle non-linéaire de la vibration glottique Implications cliniques potentielles," *Revue de laryngologie, d'otologie et de rhinologie* **123**, 273–277.

- GIOVANNI, A., SACRE, J., & ROBERT, D. (2007). "Forçage vocal," EMC (Elsevier Masson SAS), Oto-rhino-laryngologie **20-720-A-4**, 1–12.
- GRAMBLE, E. A. (1909). "Intensity as a criterion in estimating the distance of sounds," *Psychology Revue* **16**, 416–426.
- HANSEN, J. H. L., & PATIL, S. (2007). "Speech under stress: Analysis, modeling and recognition," *Speaker Classification I* **4343**, 108–137.
- HARDCASTLE, W., & LAVER, J. (1999). *The handbook of phonetic sciences* (W. J. Hardcastle et J. Laver, Eds.) (Blackwell Publishers).
- HARRIS, C. M. (1964). "Effects of Speaking Condition on Pitch," *Journal of the Acoustical Society of America* **36**, 933–936.
- HAUSTEIN, B. G. (1969). "Hypothesen über die einhörige Entfernungswahrnehmung des Menschlichen Gehörs, (Hypotheses about perception of distance in human hearing with one ear)," *Hochfrequenztech. u. Elektroakustik* **78**, 46–57.
- HEEREN, W., & HEUVEN, V. J. V. (2011). "Acoustics of Whispered Boundary Tones: Effects of Vowel Type and Tonal Crowding," *Proc. of the International Congress of Phonetic Sciences (ICPhS)* (Honk-Kong, China), pp. 851–854.
- HENRICH, N. (2001). *Etude de la source glottique en voix parlée et chantée : modélisation et estimation, mesures acoustiques et électroglottographiques, perception*, Université Paris 6.
- HENRICH, N. (2006). "Mirroring the voice from Garcia to the present day: some insights into singing voice registers," *Logopedics, phoniatrics, vocology* **31**, 3–14.
- HENRICH, N., D'ALESSANDRO, C., DOVAL, B., & CASTELLENGO, M. (2004). "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation," *Journal of the Acoustical Society of America* **115**, 1321–1332.
- HILLMAN, R. E., OESTERLE, E., & FETH, L. L. (1983). "Characteristics of the glottal turbulent noise source," *Journal of the Acoustical Society of America* **74**, 691–694.
- HOLMBERG, E. B., HILLMAN, R. E., & PERKELL, J. S. (1988). "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice," *Journal of the Acoustical Society of America* **84**, 511–29.
- HUANG, D. Y., ONG, E. P., RAHARDJA, S., DONG, M., & LI, H. (2009). "Transformation of Vocal Characteristics: A Review of Literature," *Proc. of the World Congress on Science, Engineering, and Technology (WCSET)* (Amsterdam, Netherlands), pp. 271–279.

- HUANG, D. Z., MINIFIE, F. D., KASUYA, H., & LIN, S. X. (1995). "Measures of vocal function during changes in vocal effort level," *Journal of voice official journal of the Voice Foundation* **9**, 429–438.
- HUBER, J. E., STATHOPOULOS, E. T., CURIONE, G. M., ASH, T. A., & JOHNSON, K. (1999). "Formants of children, women, and men: the effects of vocal intensity variation," *Journal of the Acoustical Society of America* **106**, 1532–1542.
- ITU-T G.114 (2003). "One-way transmission time,".
- INGARD, U. (1953). "A review of the influence of meteorological conditions on sound propagation," *Journal of the Acoustical Society of America* **25**, 405–411.
- ISHIZAKA, K., & FLANAGAN, J. L. (1972). "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell Syst. Tech. J.* **51**, 1233 – 1268.
- ISMAIL, S. (2010). "Employing Second-Order Circular Suprasegmental Hidden Markov Models to Enhance Speaker Identification Performance in Shouted Talking Environments," *EURASIP Journal on Audio, Speech, and Music Processing*.
- ISSHIKI, N. (1964). "REGULATORY MECHANISM OF VOICE INTENSITY VARIATION," *Journal of speech and hearing research* **128**, 17–29.
- ISSHIKI, N. (1969). "Remarks on Mechanism for Vocal Intensity Variation," *J Speech Hear Res* **12**, 669–672.
- ITO, T., TAKEDA, K., & ITAKURA, F. (2005). "Analysis and recognition of whispered speech," *Speech Comm* **45**, 139–152.
- JACKENDOFF, R. (2002). *Foundations of Language* Foundations Of Language (Oxford University Press).
- JESSEN, M., KÖSTER, O., & GFROERER, S. (2005). "Influence of vocal effort on average and variability of fundamental frequency," *Speech, Language and the Law*, **12**, 174–213.
- JESTEADT, W., WIER, C. C., & GREEN, D. M. (1977). "Intensity discrimination as a function of frequency and sensation level," *Journal of the Acoustical Society of America* **61**, 169–177.
- JOLIVEAU, E., SMITH, J., & WOLFE, J. (2004a). "Acoustics: tuning of vocal tract resonance by sopranos," *Nature* **427**, 116.
- JOLIVEAU, E., SMITH, J., & WOLFE, J. (2004b). "Vocal tract resonances in singing: the soprano voice," *Journal of the Acoustical Society of America* **116**, 2434–2439.

- JOVIČIĆ, S. T. (1998). "Formant feature differences between whispered and voice sustained vowels," *Acustica Acta Acustica* **84**, 739–743.
- JOVIČIĆ, S. T., & ŠARIĆ, Z. (2008). "Acoustic analysis of consonants in whispered speech," *Journal of voice official journal of the Voice Foundation* **22**, 263–274.
- JUNQUA, J. C. (1993). "The Lombard reflex and its role on human listeners and automatic speech recognizers," *Journal of the Acoustical Society of America* **93**, 510–524.
- JURAFSKY, D., BELL, A., GREGORY, M., & RAYMOND, W. D. (2001). "Probabilistic Relations between Words: Evidence from Reduction in Lexical Production," In J. Bybee et P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (John Benjamins), pp. 229–254.
- KALLAIL, K. J., & EMANUEL, F. W. (1984a). "Formant-frequency differences between isolated whispered and phonated vowel samples produced by adult female subjects," *Journal Of Speech And Hearing Research* **27**, 245–251.
- KALLAIL, K. J., & EMANUEL, F. W. (1984b). "An Acoustic Comparison of Isolated Whispered and Phonated Vowel Samples Produced by Adult Male Subjects," *Journal of Phonetics* **12**, 175–186.
- KALLAIL, K. J., & EMANUEL, F. W. (1985). "The identifiability of isolated whispered and phonated vowel samples," *Journal of phonetics* **13**, 11–17.
- KAMP, Y., & WILLEMS, L. F. (1994). "A Frobenius norm approach to glottal closure detection from the speech signal," *IEEE Transactions on Speech and Audio Processing* **2**, 258–265.
- KAWAHARA, H. (1999). "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication* **27**, 187–207.
- KENNEDY, E. (1998). "Consonant–vowel intensity ratios for maximizing consonant recognition by hearing-impaired listeners," *Journal of the Acoustical Society of America* **103**, 1098–1114.
- KLATT, D. H. (1980). "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustical Society of America* **67**, 971–995.
- KLATT, D. H., & KLATT, L. C. (1990). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America* **87**, 820–857.
- KONDOZ, A. M. (2004). *Digital Speech: Coding for Low Bit Rate Communication Systems* (Wiley), 2nd Ed.

- KOPČO, N., SCHOOLMASTER, M., & SHINN-CUNNINGHAM, B. (2004). "Learning to judge distance of nearby sounds in reverberant and anechoic environments," Joint conference of CFA/DAGA (Strasbourg, France).
- KOUNOUEDES, A., NAYLOR, P. A., & BROOKES, M. (2002). "The DYPSA algorithm for estimation of glottal closure instants in voiced speech," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 349–352.
- LAGIER, A., VAUGOYEAU, M., BOUCHE, C., GHIO, A., LEGOU, T., ASSAIANTE, C., & GIOVANNI, A. (2009). "Etude posturale de sujets normaux en situation expérimentale d'effort vocal," *Revue de laryngologie, d'otologie et de rhinologie* **130**, 11–16.
- LANE, H. (1970). "Regulation of Voice Communication by Sensory Dynamics," *Journal of the Acoustical Society of America* **47**, 618–624.
- LAPRIE, Y., & COLOTTE, V. (1998). "Automatic pitch marking for speech transformations via TD-PSOLA," *Proc. of the European Signal Processing Conference (EUSIPCO)* (Rhodes, Greece), pp. 1133–1136.
- LAVER, J. (1980). *The phonetic description of voice* New York (Cambridge University Press New-York).
- LECUIT, V., & DEMOLIN, D. (1998). "The relationship between intensity and subglottal pressure with controlled pitch," *Proc. of the 5th International Conference on Spoken Language Processing (ICSLP)* (Sydney, Australia), pp. 3083–3086.
- LEE, B. (1950). "Effect of delayed speech feedback," *Journal of the Acoustical Society of America* **22**, 824–826.
- LEHUCHE, F., & ALLALI, A. (2010). *La voix - Tome I : Anatomie et physiologie des organes de la voix et de la parole* (Masson), 4ème édit.
- LEINO, T. (1993). "Long-term Average Spectrum Study on Speaking Voice Quality in Male Actors," *Proc. of the Stockholm Music Acoustic Conference (SMAC)* (Stockholm, Swedish), pp. 206–210.
- LÉON, P., & MARTIN, P. (1969). *Prolégomènes à l'étude des structures intonatives* (M. Didier, Ed.) (Montréal), Suidia pho.
- LIEBERMAN, P., KATZ, W., JONGMAN, A., ZIMMERMAN, R., & MILLER, M. (1985). "Measures of the sentence intonation of read and spontaneous speech in American English," *Journal of the Acoustical Society of America* **77**, 649–57.

- LINDBLOM, B. (1990). "Explaining phonetic variation: a sketch of the H&H theory," In W. Hardcastle et A. Marchal (Eds.), *Speech Production and Speech Modeling* (Kluwer Academic Publishers), pp. 403–439.
- LINDBLOM, B., BROWNLEE, S., DAVIS, B., & MOON, S.-J. (1992). "Speech transforms," *Speech Communication* **11**, 357–368.
- LINDBLOM, B., & SUNDBERG, J. (1971). "Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movement," *Journal of the Acoustical Society of America* **50**, 1166–1179.
- LITTLE, A. D., MERSHON, D. H., & COX, P. H. (1992). "Spectral content as a cue to perceived auditory distance," *Perception* **21**, 405–416.
- LIÉNARD, J.-S. (1977). *Les Processus de la communication parlée - Introduction à l'analyse et la synthèse de la parole* (Masson).
- LIÉNARD, J.-S., & DI BENEDETTO, M. G. (1999). "Effect of vocal effort on spectral properties of vowels," *Journal of the Acoustical Society of America* **106**, 411–422.
- LOBRÉAU, S. (2010). *Etude de l'influence de la propagation du son de la parole dans l'air sur la perception de la distance*, (Thèse de Master), Université de Haute-Alsace, Mulhouse-Colmar.
- LOMBARD, É. (1911). "Le signe de l'élévation de la voix," *Annales des Maladies de L'Oreille et du Larynx* **37**, 101–119.
- LOUNSBURY, B. F., & BUTLER, R. A. (1979). "Estimation of distances of recorded sounds presented through headphones," *Scandinavian Audiology* **8**, 145–149.
- MACCHI, M. J., ALTOM, M. J., KAHN, D., SINGHAL, S., & SPIEGEL, M. F. (1993). "Intelligibility as a function of speech coding method for template-based speech synthesis," *Proc. of the European Conference on Speech Communication and Technology (Eurospeech)* (Berlin, Germany), pp. 893–896.
- MAKHOUL, J. I., & WOLF, J. J. (1972). *Linear prediction and the spectral analysis of speech*, *Advanced Research Projects Agency, BBN Report No. 2304*.
- MARIANI, J. (2002). *Analyse, synthèse et codage de la parole - Traitement automatique du langage parlée I* (Hermès sciences publications).
- MARKEL, J. D., & GRAY, A. H. (1976). *Linear Prediction of Speech* (Springer-Verlag, Ed.) (New-York).

- MATSUDA, M., & KASUYA, H. (1999). "Acoustic Nature of the Whisper," Proc. of the European Conference on Speech Communication and Technology (Eurospeech) (Budapest, Hungary), pp. 137–140.
- MCAULAY, R., & QUATIERI, T. (1986). "Speech analysis/Synthesis based on a sinusoidal representation," *Acoustics, Speech and Signal Processing, IEEE Transactions on* **34**, 744–754.
- MCGREGOR, P., HORN, A. G., & TODD, M. A. (1985). "Are familiar sounds ranged more accurately?," *Perceptual and motor skills* **61**, 1082.
- MERSHON, D. H., BALLENGER, W., LITTLE, A. D., MCMURTRY, P., & BUCHANAN, J. (1989). "Effects of room reflectance and background noise on perceived auditory distance," *Perception* **18**, 403–416.
- MERSHON, D. H., & BOWERS, J. N. (1979). "Absolute and relative cues for the auditory perception of egocentric distance," *Perception* **8**, 311–322.
- MERSHON, D. H., & KING, L. E. (1975). "Intensity and reverberation as factors in auditory-perception of egocentric distance," *Perception And Psychophysics* **18**, 409–415.
- MIDDLEBROOKS, J. C., & GREEN, D. M. (1991). "Sound localization by human listeners," *Annual review of psychology* **42**, 135–159.
- MILLER, G. A. (1947). "Sensitivity to changes in the intensity of white noise and its relation to masking and loudness," *Journal of Acoustic Society of America* **19**, 609–619.
- MIZUNO, H., & ABE, M. (1996). "A formant frequency modification algorithm dealing with the pole interaction," *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)* **79**, 46–55.
- MONSEN, R. B., & ENGBRETSON, A. M. (1983). "The accuracy of formant frequency measurements: a comparison of spectrographic analysis and linear prediction," *Journal of speech and hearing research* **26**, 89–97.
- MOORE, B. C. J. (2004). "The perception of distance," *An introduction to the psychology of hearing* 5th ed., pp. 265–266.
- MOOSHAMMER, C. (2004). "An EGG study on global and local vocal effort changes in German: preliminary results," *Proceedings of the International Conference of Voice Physiology and Biomechanics* pp. 401–404.
- MOOSHAMMER, C. (2006). "Laryngeal correlates of prosodic variation," *Proc. of the 7th International Seminar on Speech Production (ISSP) (Ubatuba, Brazil)*.

- MOOSHAMMER, C. (2010). "Acoustic and laryngographic measures of the laryngeal reflexes of linguistic prominence and vocal effort in German," *Journal of the Acoustical Society of America* **127**, 1047–1058.
- MORRIS, R. W. (2003). "*Enhancement and recognition of whispered speech*", (Ph.D. Thesis), Georgia institute of technology.
- MORRIS, R. W., & CLEMENTS, M. A. (2002). "Modification of formants in the line spectrum domain," *IEEE Signal Processing Letters* **9**, 19–21.
- MOULINES, E., & CHARPENTIER, F. (1990). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication* **9**, 453–467.
- MOULINES, E., & LAROCHE, J. (1995). "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Communication* **16**, 175–205.
- MOZZICONACCI, S. J. L. (1998). "*Speech Variability and Emotion: Production and Perception*", (Ph.D. Thesis), Technical University Eindhoven.
- NAGUIB, M. (2001). "Estimating the distance to a source of sound: mechanisms and adaptations for long-range communication," *Animal Behaviour* **62**, 825–837.
- NANJO, H., MIKAMI, H., KAWANO, K., & NISHIURA, T. (2009). "A fundamental study of shouted speech for acoustic-based security system," *Proc. of Interspeech (Brighthon, UK)*, pp. 1027–1030.
- NAWKA, T., ANDERS, L. C., CEBULLA, M., & ZURAKOWSKI, D. (1997). "The speaker's formant in male voices," *Journal of voice : official journal of the Voice Foundation* **11**, 422–428.
- NETTER, F. (2007). *Atlas d'anatomie du corps humain* (Masson), 4th ed.
- NIETO, O. (2008). "*Voice transformations for extreme vocal effects*", (M.Sc. Thesis), Master's thesis, Pompeu fabra university.
- NORDSTROM, K. I. (2008). "*Transforming high-effort voices into breathy voices using adaptive pre-emphasis linear prediction*", (Ph.D. Thesis), University of Victoria.
- NORDSTROM, K. I., DRIESSEN, P. F., & RUTLEDGE, G. A. (2006). "Influence of the LPC filter upon the perception of breathiness and vocal effort," *Proc. of the 6th IEEE International Symposium on Signal Processing and Information Technology (ISSPIT) (Vancouver, Canada)*, pp. 23–27.
- NORDSTROM, K. I., TZANETAKIS, G., & DRIESSEN, P. F. (2008). "Transforming Perceived Vocal Effort and Breathiness Using Adaptive Pre-Emphasis Linear Prediction," *IEEE transactions on audio, speech, and language processing* **16**, 1087–1096.

- NÌ CHASAIDE, A., & GOBL, C. (1999). "Voice Source Variation," In W. J. Hardcastle et J. Laver (Eds.), *handbook of phonetic science* (Blackwell Oxford, UK), pp. 427–461.
- OHTA, Y., & TAMURA, H. (1999). "Auditory distance perception in real, virtual, and mixed environments," In Y. Ohta et H. Tamura (Eds.), *Mixed reality: merging real and virtual worlds* (Tokyo: Ohmsha), pp. 201–214.
- O'SHAUGHNESSY, D. (1999). *Speech Communications: Human and Machine* (John Wiley & Sons Inc).
- PALIWAL, K. K. (1984). "Performance of the weighted burg methods of ar spectral estimation for pitch-synchronous analysis of voiced speech," *Speech Communication* **3**, 221–231.
- PELEGRÍN-GARCÍA, D., SMITS, B., BRUNSKOG, J., & JEONG, C.-H. (2011). "Vocal effort with changing talker-to-listener distance in different acoustic environments," *Journal of the Acoustical Society of America* **129**, 1981–1990.
- PETERSEN, J. (1990). "Estimation of loudness and apparent distance of pure tones in free-field," *Acta Acustica* **70**, 61–65.
- PETRUSHIN, V. A., TSIRULNIK, L. I., & MAKAROVA, V. (2010). "Whispered speech prosody modeling for TTS synthesis," *Proc. of the 5th International Conference on Speech Prosody* (Chicago, IL, USA), pp. 1151–1154.
- PFITZINGER, H. R. (2006). "Five dimensions of prosody : intensity, intonation, timing, voice quality, and degree of reduction," In H. Hoffmann et R. Mixdorf (Eds.), *Speech Prosody Abstract Book* (Dresden, Germany), TUDpress., pp. 6–9.
- PHILBECK, J. W., & MERSHON, D. H. (2002). "Knowledge about typical source output influences perceived auditory distance (L)," *Journal of the Acoustical Society of America* **111**, 1980–1983.
- PICKETT, J. M. (1956). "Effect of the vocal force on the intelligibility of speech sound," *Journal of the Acoustical Society of America* **28**, 902–905.
- PICKETT, J. M. (1998). "The glottal sound source and the spectra of vowels," *The Acoustics of Speech Communication: Fundamentals, Speech Perception Theory, and Technology* (Allyn & Bacon).
- PIERCE, A. D. (1981). *Acoustics: an introduction to its physical principles and applications* McGraw-Hil.
- PIERCY, J. E. (1977). "Review of noise propagation in the atmosphere," *Journal of the Acoustical Society of America* **61**, 1403–1418.

- PLANT, R. L., & YOUNGER, R. M. (2000). "The interrelationship of subglottic air pressure, fundamental frequency, and vocal intensity during speech," *Journal of voice : official journal of the Voice Foundation* **14**, 170–177.
- RICHARD, G., & D'ALESSANDRO, C. (1996). "Analysis/synthesis and modification of the speech aperiodic component," *Speech Communication* **19**, 221–244.
- RIESZ, R. R. (1933). "The Relationship between Loudness and the Minimum Perceptible Increment of Intensity," *Journal of the Acoustical Society of America* **4**, 211–216.
- ROSENBERG, A. E. (1971). "Effect of glottal pulse shape on the quality of natural vowels," *Journal of the Acoustical Society of America* **49**, 583–590.
- ROSTOLLAND, D. (1982a). "Acoustic features of shouted voice," *Acustica* **50**, 118–125.
- ROSTOLLAND, D. (1982b). "Phonetic structure of shouted voice," *Acustica* **51**, 80–89.
- ROSTOLLAND, D. (1985). "Intelligibility of shouted voice," *Acustica* **57**, 103–121.
- SAKOE, H., & CHIBA, S. (1978). "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing* **26**, 43–49.
- SAWASHIMA, M. (1977). "Fiberoptic observations of the larynx and other speech organs," In M. Sawashima et F. S. Cooper (Eds.), *Dynamic Aspects of Speech Production* (University Tokyo Press), pp. 31–46.
- SCHRÖDER, M., & GRICE, M. (2003). "Expressing vocal effort in concatenative synthesis," *Proc. of the 15th International Conference of Phonetic Sciences (ICPhS) (Barcelona, Spain)*, pp. 2589–2592.
- SCHULMAN, R. (1989). "Articulatory dynamics of loud and normal speech," *Journal of the Acoustical Society of America* **85**, 295–312.
- SCHWARTZ, M. F. (1968). "Identification of Speaker Sex from Isolated, Whispered Vowels," *Journal of the Acoustical Society of America* **44**, 1736–1737.
- SCHWARTZ, M. F. (1970). "Power spectral density measurements of oral and whispered speech," *Journal Of Speech And Hearing Research* **13**, 445–446.
- SCHWARTZ, M. F. (1972). "Bilabial Closure Durations for /p/, /b/, and /m/ in Voiced and Whispered Vowel Environments," *Journal of the Acoustical Society of America* **51**, 2025–2029.
- SESHADRI, G., & YEGNANARAYANA, B. (2009). "Perceived loudness of speech based on the characteristics of glottal excitation source," *Journal of the Acoustical Society of America* **126**, 2061–2071.

- SHARIFZADEH, H. R., MCLOUGHLIN, I. V., & RUSSELL, M. J. (2010). "Toward a comprehensive vowel space for whispered speech," Proc. of the 7th International Symposium on Chinese Spoken Language Processing (ISCSLP) (Nantou, Taiwan), pp. 65–68.
- SHEELINE, C. W. (1983). *"An Investigation of the Effects of Direct and Reverberant Signal Interactions on Auditory Distance Perception"*, (Ph.D. Thesis), Stanford University.
- SHINN-CUNNINGHAM, B. (2000). "Learning Reverberation: Considerations for Spatial Auditory Displays," Proc. of the International Community of Auditory Display (ICAD) (Atlanta, GA), pp. 126–134.
- SHRIBERG, E., LADD, D. R., TERKEN, J., & STOLCKE, A. (1996). "Modeling pitch range variation within and across speakers: predicting fo targets when 'speaking up'," Proc. of the 4th International Conference on Spoken Language Processing (ICSLP) (Philadelphia, PA, USA), pp. 1–4.
- SIGNOL, F. (2009). *"Estimation de fréquences fondamentales multiples en vue de la séparation de signaux de parole mélangés dans un même canal"*, (Thèse de Doctorat), Université Paris-Sud 11.
- SILVERMAN, K., BECKMAN, M., PITRELLI, J., OSTENDORF, M., WIGHTMAN, C., PRICE, P., PIERREHUMBERT, J., ET AL. (1992). "ToBI: A standard for labeling English prosody," Proc. of the 2nd International Conference on Spoken Language Processing (ICSLP) (Alberta, Canada), pp. 867–870.
- SIMPSON, W. E., & STANTON, L. D. (1973). "Head movement does not facilitate perception of the distance of a source of sound," *The American journal of psychology* **86**, 151–159.
- STANTON, B. J., JAMIESON, L. H., & ALLEN, G. D. (1988). "Acoustic-phonetic analysis of loud and Lombard speech in simulated cockpit conditions," ICASSP (New-York, USA), pp. 331–334.
- STEVENS, K. N. (1971). "Airflow and Turbulence Noise for Fricative and Stop Consonants: Static Considerations," *Journal of the Acoustical Society of America* **50**, 1180–1192.
- STEVENS, K. N. (1998). *Acoustic Phonetics* (MIT Press Cambridge, MA).
- STEVENS, S. S., & GUIRAO, M. (1962). "Loudness, reciprocity and partition scales," *Journal of the Acoustical Society of America* **34**, 1466–1471.
- STORY, B. H., & TITZE, I. R. (1995). "Voice simulation with a body-cover model of the vocal folds," *Journal of the Acoustical Society of America* **97**, 1249–1260.
- STRYBEL, T. Z., & PEROTT, D. R. (1984). "Discrimination of the relative distance in the auditory modality: The succes and failure of the loudness discrimination hypothesis," *Journal of the Acoustical Society of America* **76**, 318–334.

- STUART, A., KALINOWSKI, J., RASTATTER, M. P., & LYNCH, K. (2002). "Effect of delayed auditory feedback on normal speakers at two speech rates," *Journal of the Acoustical Society of America* **111**, 2237–2241.
- STURMEL, N. (2011). "*Analyse de la qualité vocale appliquée à la parole expressive*", (Thèse de Doctorat), Université Paris-Sud 11.
- STYLIANOU, Y. (1996). "*Harmonic Plus Noise Models for Speech, Combined with Statistical Methods, for Speech and Speaker Modification*", (Ph.D. Thesis), Ecole Nationale Supérieure des Télécommunications.
- STYLIANOU, Y. (2009). "Voice Transformation: A survey," 2009 IEEE International Conference on Acoustics, Speech and Signal Processing pp. 3585–3588.
- STYLIANOU, Y., LAROCHE, J., & MOULINES, E. (1995). "High-quality speech modification based on a harmonic+ noise model," Proc. of the European Conference on Speech Communication and Technology (Eurospeech) (Madrid, Spain), pp. 550–553.
- SUN, X. (2000). "Voice quality conversion in TD-PSOLA speech synthesis," IEEE International Conference on Acoustics, Speech, and Signal Processing. p. II.953–II.956.
- SUNDBERG, J. (1973). "The source spectrum in professional singing," *Folia phoniatrica* **25**, 71–90.
- SUNDBERG, J. (1977). "The acoustics of the singing voice," *Scientific American* **236**, 82–84, 86, 88–91.
- SUNDBERG, J., SCHERER, R., HESS, M., & MÜLLER, F. (2010). "Whispering--a single-subject study of glottal configuration and aerodynamics," *Journal of voice official journal of the Voice Foundation* **24**, 574–584.
- SUNDBERG, J., TITZE, I. R., & SCHERER, R. (1993). "Phonatory control in male singing: A study of the effects of subglottal pressure, fundamental frequency, and mode of phonation on the voice source," *Journal of Voice* **7**, 15–29.
- SWERDLIN, Y., SMITH, J., & WOLFE, J. (2010). "The effect of whisper and creak vocal mechanisms on vocal tract resonances," *Journal of the Acoustical Society of America* **127**, 2590–2598.
- TALKIN, D. (1995). "A robust algorithm for pitch tracking (RAPT)," In W. B. Kleijn et K. K. Paliwal (Eds.), *Speech coding and synthesis*, (Elsevier), pp. 495–518.
- TARTTER, V. C. (1989). "What's in a whisper?," *Journal of the Acoustical Society of America* **86**, 1678–1683.

- TARTTER, V. C. (1991). "Identifiability of vowels and speakers from whispered syllables," *Perception & psychophysics* **49**, 365–372.
- TASSA, A., & LIÉNARD, J.-S. (2000). "A New Approach to the Evaluation of Vocal Effort by the PSOLA Method," *WEBSLS The European Student Journal of Language and Speech* **1**, 00–01.
- TAYLOR, P. (2000). "Analysis and synthesis of intonation using the Tilt model," *Journal of the Acoustical Society of America* **107**, 1697–1714.
- THOMPSON, S. P. (1882). "On the function of the two ears in the perception of space," *Philosophical Magazine* **13**, 406–416.
- TITZE, I. R. (1994). *Principles of voice production, 2nd printing* (Prentice Hall).
- TITZE, I. R. (2008). "Nonlinear source-filter coupling in phonation: theory," *Journal of the Acoustical Society of America* **123**, 2733–2749.
- TITZE, I. R., & RIEDE, T. (2010). "A cervid vocal fold model suggests greater glottal efficiency in calling at high frequencies," *PLoS computational biology* **6**, .
- TOM, K., TITZE, I. R., HOFFMAN, E. A., & STORY, B. H. (2001). "Three-dimensional vocal tract imaging and formant structure: varying vocal register, pitch, and loudness," *Journal of the Acoustical Society of America* **109**, 742–747.
- TRAUNMÜLLER, H. (1981). "Perceptual dimension of openness in vowels," *Journal of the Acoustical Society of America* **69**, 1465–1475.
- TRAUNMÜLLER, H. (1985). "The role of the fundamental and the higher formants in the perception of speaker size, vocal effort, and vowel openness," *PERILUS* **IV**, 92–102.
- TRAUNMÜLLER, H. (1988). "Paralinguistic variation and invariance in the characteristic frequencies of vowels," *Phonetica* **45**, 1–29.
- TRAUNMÜLLER, H. (1997). "Perception of Speaker Sex, Age, and Vocal Effort," *Phonum* **4**, 183–186.
- TRAUNMÜLLER, H., & ERIKSSON, A. (1997). "A Method of Measuring Formant Frequencies at High Fundamental Frequencies," *Proc. of the 5th European Conference on Speech Communication and Technology (Eurospeech)* (Rhodes, Greece), pp. 477–480.
- TRAUNMÜLLER, H., & ERIKSSON, A. (2000). "Acoustic effects of variation in vocal effort by men, women, and children," *Journal of the Acoustical Society of America* **107**, 3438–3451.
- TSUNODA, K., OHTA, Y., SODA, Y., NIIMI, S., & HIROSE, H. (1997). "Laryngeal adjustment in whispering magnetic resonance imaging study," *The Annals of otology, rhinology, and laryngology* **106**, 41–43.

- TUAN, V. N., & D'ALESSANDRO, C. (1999). "Robust Glottal Closure Detection Using The Wavelet Transform," Proc. of the European Conference on Speech Communication and Technology (Eurospeech) (Budapest, Hungary), pp. 2805–2808.
- TURK, O., SCHRÖDER, M., BOZKURT, B., & ARLSAN, L. M. (2005). "Voice quality interpolation for emotional text-to-speech synthesis," Proc. of Interspeech (Lisbon, Portugal), pp. 797 – 800.
- VALLABHA, G. K., & TULLER, B. (2002). "Systematic errors in the formant analysis of steady-state vowels," *Speech Communication* **38**, 141–160.
- VELDHUIS, R. (1998). "A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation," *Journal of the Acoustical Society of America* **103**, 566–571.
- VERFAILLE, V., & ARFIB, D. (2001). "A-DAFX: Adaptive Digital Audio Effects," Proc. of the International Conference on Digital Audio Effects (DAFX) (Limerick, Ireland), pp. 10–13.
- VERHELST, W., & ROELANDS, M. (1993). "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," Proc. of the IEEE international conference on Acoustics, Speech, and Signal Processing (ICASSP) pp. 554–557.
- VINCENT, D., ROSEC, O., & CHONAVEL, T. (2007). "A New Method for Speech Synthesis and Transformation Based on an ARX-LF Source-Filter Decomposition and HNM Modeling," Proc. of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) (Honolulu, Hawaii, USA), pp. 525–528.
- WARREN, R. (1968). "Vocal compensation for change of distance," Proc. of the 6th International Congress of Acoustics (Tokyo, Japan), pp. A61–A64.
- WARREN, R. (1977). "Subjective loudness and its physical correlate," *Acustica* **37**, 334–346.
- WATSON, P. J., & HUGHES, D. (2006). "The relationship of vocal loudness manipulation to prosodic F0 and durational variables in healthy adults," *Journal of speech language and hearing research JSLHR* **49**, 636–644.
- WATTERSON, T., MCFARLANE, S. C., & DIAMOND, K. L. (1993). "Phoneme Effects on Vocal Effort and Vocal Quality," *Am J Speech Lang Pathol* **2**, 74–78.
- WHALEN, D. H., & LEVITT, A. G. (1995). "The universality of intrinsic F0 of vowels," *Journal of Phonetics* **23**, 349–366.
- WILKENS, H., & BARTEL, H.-H. (1977). "Wiedererkennbarkeit der Originallautstärke eines Sprechers bei elektroakustischer Wiedergabe," *Acustica* **37**, 45–49.

- WOLFE, J., & GARNIER, M. (2009). "Vocal tract resonances in speech, singing and playing musical instruments," Human Frontier Science Program journal **3**, 6–23.
- YI, S., KIM, H. S., & LEE, O. G. (2000). "Glottal parameters contributing to the peception of the loud voices," Proc. of the 6th International Conference on Spoken Language Processing (ICSLP) (Beijing, China), pp. 580–583.
- ZAHORIK, P. (1996). "*Auditory distance perception: A literature review*", (Ph.D. Preliminary Examination), University of West.
- ZAHORIK, P. (2002). "Direct-to-reverberant energy ratio sensitivity," Journal of the Acoustical Society of America **112**, 2110–2117.
- ZAHORIK, P., BRUNGART, D. S., & BRONKHORST, A. W. (2005). "Auditory distance perception in humans: A summary of past and present research," Acta Acustica United With Acustica **91**, 409–420.
- ZAHORIK, P., & KELLY, J. W. (2007). "Accurate vocal compensation for sound intensity loss with increasing distance in natural environments," Journal of the Acoustical Society of America **122**, EL143–EL150.
- ZAHORIK, P., & WIGHTMAN, F. L. (2001). "Loudness constancy with varying sound source distance," Nature Neuroscience **4**, 78–83.
- ZELINKA, P., & SIGMUND, M. (2010). "Automatic vocal effort detection for reliable speech recognition," Proc. of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP) (Kittilä, Finland), pp. 349–354.
- ZEROUAL, C., ESLING, J. H., & CREVIER-BUCHMAN, L. (2005). "Physiological study of whispered speech in Moroccan Arabic," Proc. of Interspeech (Lisbon, Portugal), pp. 1069–1072.
- ZHANG, C., & HANSEN, J. H. L. (2007). "Analysis and Classification of Speech Mode : Whispered through Shouted," Proc. of Interspeech (Antwerp, Belgium), pp. 2289–2292.
- ZHANG, C., & HANSEN, J. H. L. (2011). "Whisper-Island Detection Based on Unsupervised Segmentation With Entropy-Based Speech Feature Processing," IEEE transactions on audio, speech, and language processing **19**, 883–894.
- ZWICKER, E., & FELDTKELLER, R. (1981). *Psychoacoustique : l'oreille, récepteur d'information* (Masson).
- ZÖLZER, U., AMATRIAIN, X., ARFIB, D., BONADA, J., POLI, G. D., DUTILLEUX, P., EVANGELISTA, G., ET AL. (2002). *DAFX: Digital Audio Effects* (U. Zölzer, Ed.) (Wiley), p. 554.

# **ANNEXES**



# Annexe A

## Listes des logatomes CV, CVC, VCV ; CVCV

---

### A.1 Liste des logatomes pour la voyelle [a] :

	CV	CVC	VCV	CVCV
	pa [pa]	pape [pap]	apa [apa]	papa [papa]
	ba [ba]	babe [bab]	aba [aba]	baba [baba]
	ma [ma]	mame [mam]	ama [ama]	mama [mama]
	ta [ta]	tate [tat]	ata [ata]	tata [tata]
	da [da]	dade [dad]	ada [ada]	dada [dada]
	na [na]	nane [nan]	ana [ana]	nana [nana]
	gna [na]	gnagne [nan]	agna [ana]	gnagna [nana]
	ka [ka]	kafe [kak]	aka [aka]	kaka [kaka]
[a]	ga [ga]	gague [gag]	aga [aga]	gaga [gaga]
	fa [fa]	fafe [faf]	afa [afa]	fafa [fafa]
	va [va]	vave [vav]	ava [ava]	vava [vava]
	la [la]	lale [lal]	ala [ala]	lala [lala]
	sa [sa]	sasse [sas]	assa [asa]	sassa [sasa]
	za [za]	zaze [zaz]	aza [aza]	zaza [zaza]
	cha [ʃa]	chache [ʃaʃ]	acha [aʃa]	chacha [ʃaʃa]
	ja [ʒa]	jaje [ʒaʒ]	aja [aʒa]	jaja [ʒaʒa]
	ra [Ra]	rare [RAR]	ara [ARA]	rara [RARA]

### A.2 Liste des logatomes pour la voyelle [i] :

	CV	CVC	VCV	CVCV
	pi [pi]	pipe [pip]	ipi [ipi]	pipi [pipi]
	bi [bi]	bibe [bib]	ibi [ibi]	bibi [bibi]
	mi [mi]	mime [mim]	imi [imi]	mimi [mimi]
	ti [ti]	tite [tit]	iti [iti]	titi [titi]
	di [di]	dide [did]	idi [idi]	didi [didi]
	ni [ni]	nine [nin]	ini [ini]	nini [nini]
	gni [ɲi]	gnigne [ɲin]	igni [ɲni]	gnigni [ɲini]
	ki [ki]	kike [kik]	iki [iki]	kiki [kiki]
[i]	gi [gi]	guigue [gig]	igi [igi]	gigi [gigi]
	fi [fi]	fife [fif]	ifi [ifi]	fifi [fifi]
	vi [vi]	vive [viv]	ivi [ivi]	vivi [vivi]
	li [li]	lile [lil]	ili [ili]	lili [lili]
	si [si]	sisse [sis]	issi [isi]	sissi [sisi]
	zi [zi]	zize [ziz]	izi [izi]	zizi [zizi]
	chi [ʃi]	chiche [ʃiʃ]	ichi [iʃi]	chichi [ʃiʃi]
	ji [ʒi]	jje [ʒiʒ]	iji [iʒi]	jiji [ʒiʒi]
	ri [Ri]	rire [RiR]	iri [iRi]	riiri [RiRi]

### A.3 Liste des logatomes pour la voyelle [u]

	CV	CVC	VCV	CVCV
	pou [pu]	poupe [pup]	oupou [upu]	poupou [pupu]
	bou [bu]	boube [bub]	oubou [ubu]	boubou [bubu]
	mou [mu]	moume [mum]	oumou [umu]	moumou [mumu]
	tou [tu]	toute [tut]	outou [utu]	toutou [tutu]
	dou [du]	doude [dud]	oudou [udu]	doudou [dudu]
	nou [nu]	noune [nun]	ounou [unu]	nounou [nunu]
	gnou [ɲu]	gnougne [ɲuɲ]	ougnou [uɲu]	gnougnou [ɲuɲu]
	kou [ku]	kouke [kuk]	oukou [uku]	koukou [kuku]
[ou]	gou [gu]	guougue [gug]	ougou [ugu]	gougou [gugu]
	fou [fu]	foufe [fuf]	oufou [ufu]	foufou [fufu]
	vou [vu]	vouve [vuv]	ouvou [uvu]	vouvou [vuvu]
	lou [lu]	loule [lu]	oulou [ulu]	loulou [lulu]
	sou [su]	sousse [sus]	oussou [usu]	soussou [susu]
	zou [zu]	zouze [zuz]	ouzou [uzu]	zouzou [zuzu]
	chou [ʃu]	chouche [ʃuʃ]	ouchou [uʃu]	chouchou [ʃuʃu]
	jou [ʒu]	jouje [ʒuʒ]	oujou [uʒu]	joujou [ʒuʒu]
	rou [Ru]	roure [RuR]	ourou [uRu]	rourou [RuRu]

# Annexe B

## Mesure du quotient ouvert pour le corpus DB3

Dans cette annexe nous donnons les valeurs brutes ainsi que les variations relatives du quotient ouvert mesuré sur les voyelles isolée du corpus DB3.

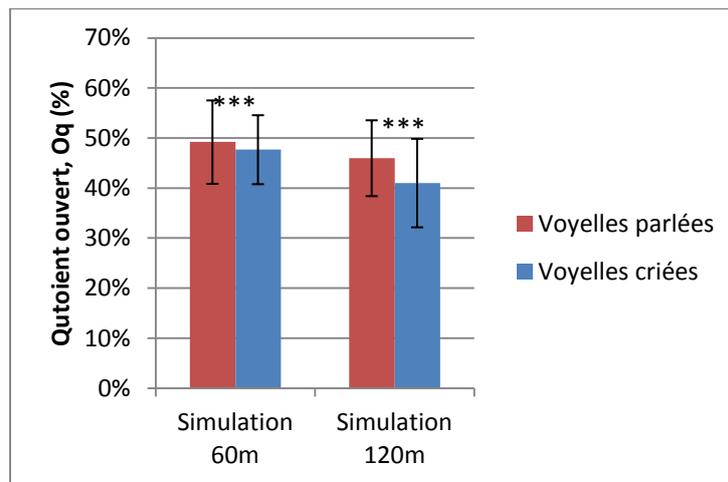


Figure B-1 : Valeurs absolues du quotient ouvert pour les voyelles parlées et créées du corpus DB3 qui contient une simulation de communication à distance de 60 m e de 120 m.

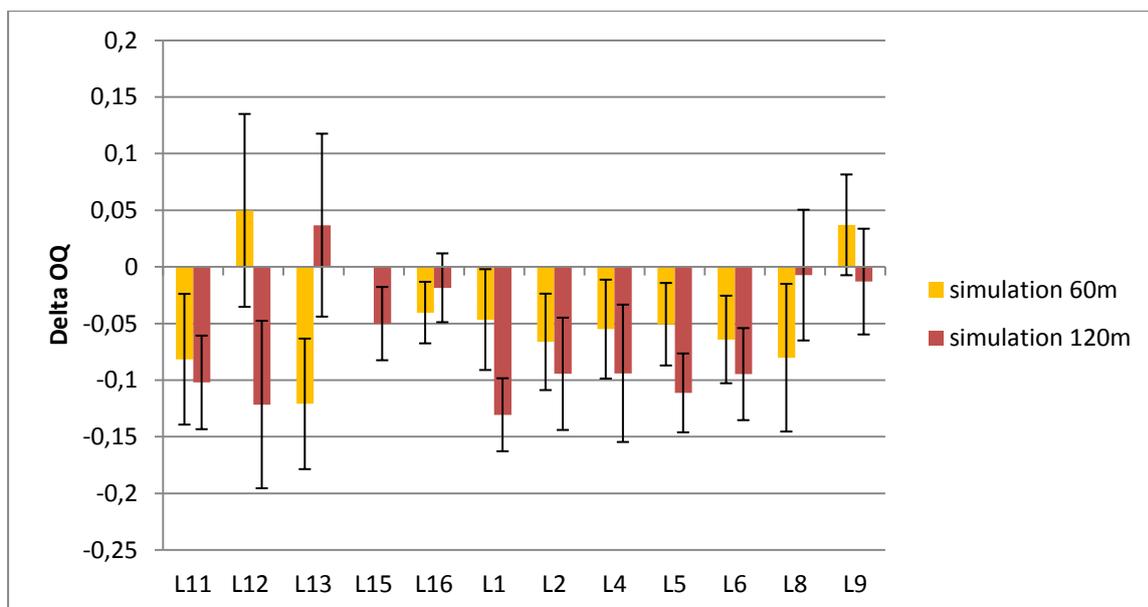


Figure B-2 : Variations du quotient ouvert entre les voyelles parlées et créées du corpus DB3 qui contient une simulation de communication à distance de 60 m e de 120 m.



# Annexe C

## Matrice de confusion des tests d'intelligibilité pour les voyelles

---

Cette annexe donne les matrices de confusion concernant l'intelligibilité des voyelles réalisées au cours du chapitre 11. Pour chaque mode de phonation présent dans le test une matrice de confusion a été tracée. Ces matrices englobent les résultats des 25 sujets pour l'ensemble du test (tous locuteurs confondus)

### C.1 Intelligibilité des voyelles en voix parlée

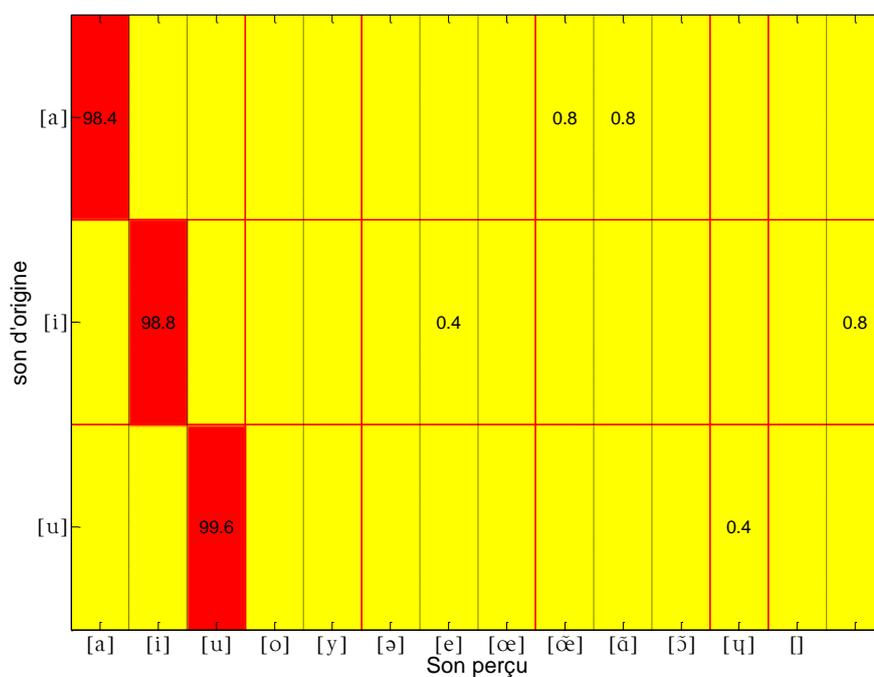


Figure C-1 : Matrice de confusion des voyelles pour la voix parlée, le symbole [] correspond à une intelligibilité nulle

### C.2 Intelligibilité des consonnes en voix chuchotée

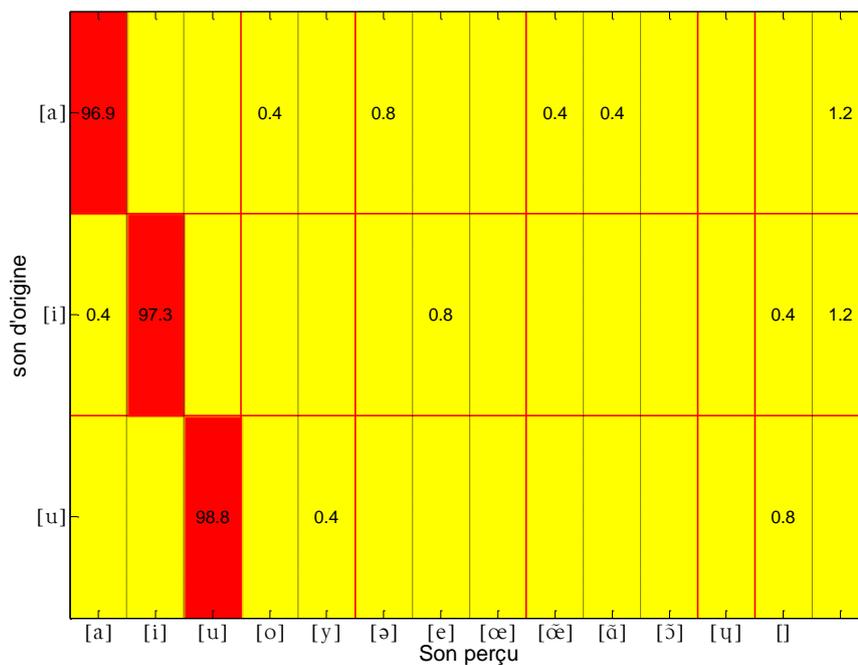


Figure C-2 : Matrice de confusion des voyelles pour la voix chuchotée, le symbole [] correspond à une intelligibilité nulle

### C.3 Intelligibilité des consonnes en voix criée

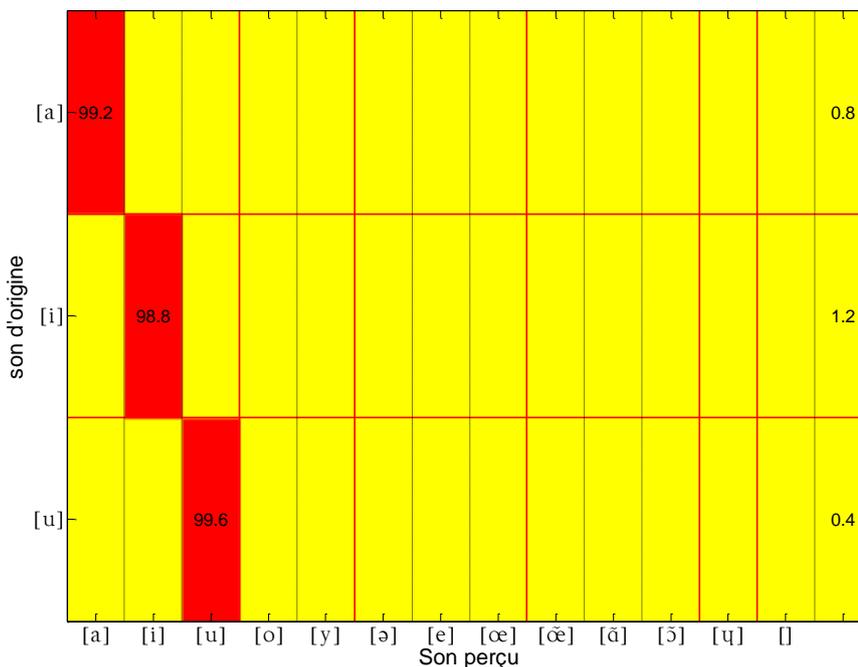


Figure C-3 : Matrice de confusion des voyelles pour la voix criée, le symbole [] correspond à une intelligibilité nulle

### C.4 Intelligibilité des consonnes en voix transformée chuchotée

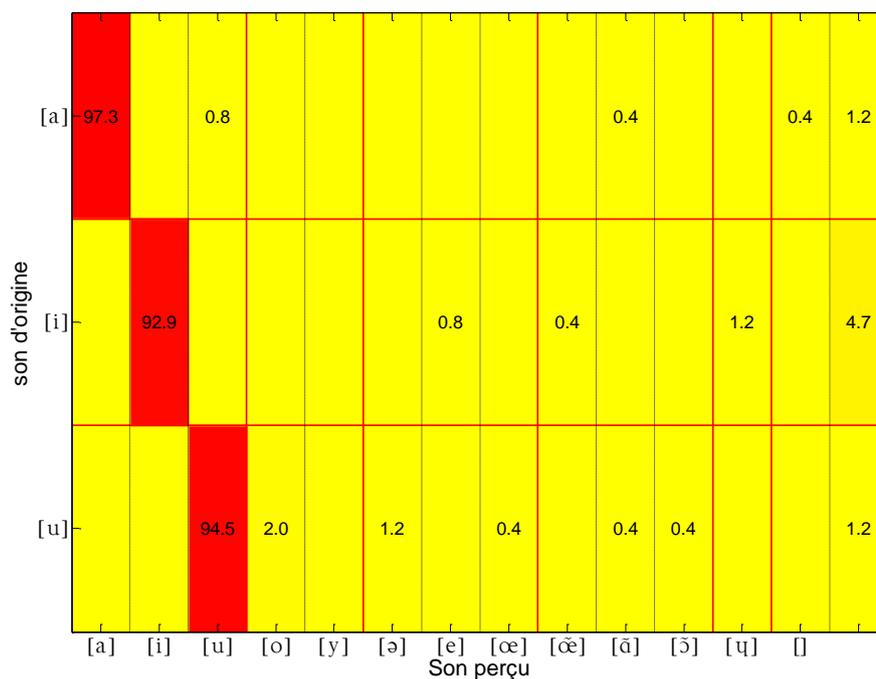


Figure C-4 : Matrice de confusion des voyelles pour la voix transformée chuchotée, le symbole [] correspond à une intelligibilité nulle

### C.5 Intelligibilité des consonnes en voix transformée criée

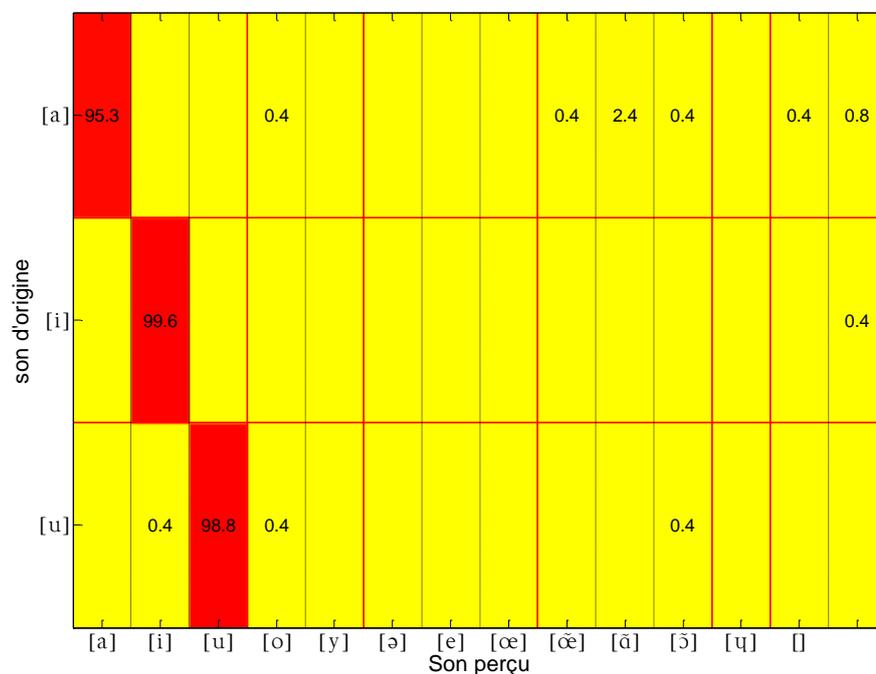


Figure C-5 : Matrice de confusion des voyelles pour la voix transformée criée, le symbole [] correspond à une intelligibilité nulle



# Annexe D

## Matrice de confusion des tests d'intelligibilité en fonction des locuteurs

---

Cette annexe, donne les différents taux d'intelligibilité en fonction des locuteurs. On retrouve ainsi pour chaque locuteur, le taux d'intelligibilité des mots entiers, des consonnes et des voyelles pour les 5 modes de phonations

### D.1 Locuteur 1

#### D.1.1 Intelligibilité globale

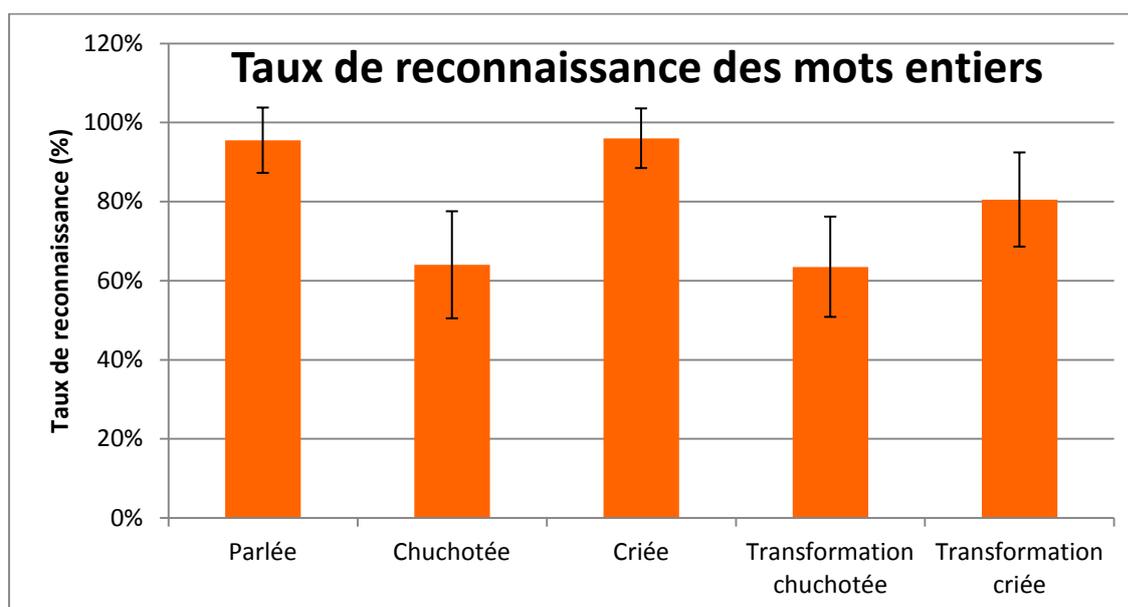


Figure D-1 : Moyennes des taux d'intelligibilité, obtenues par 25 sujets, des mots en fonction des modes de phonation ou des transformations pour le locuteur 1

### D.1.2 Intelligibilité globale des consonnes

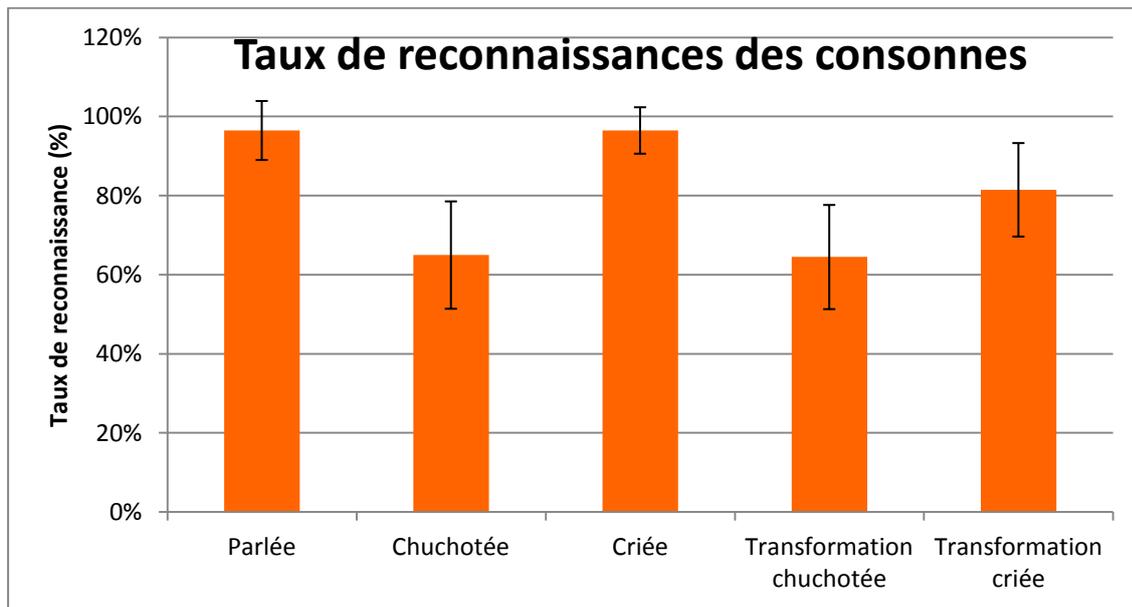


Figure D-2 : Moyennes des taux d'intelligibilité, obtenues par 25 sujets, des consonnes en fonction des modes de phonation ou des transformations pour le locuteur 1

### D.1.3 Intelligibilité globale des voyelles

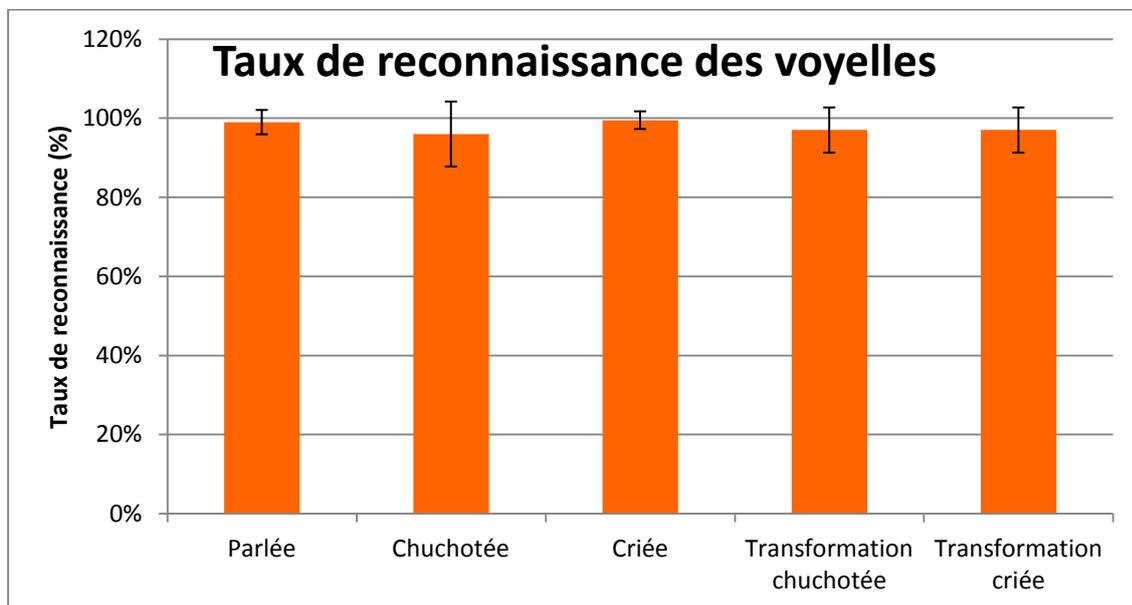


Figure D-3 : Moyennes des taux d'intelligibilité, obtenues par 25 sujets, des voyelles en fonction des modes de phonation ou des transformations pour le locuteur 1

### D.1.4 Intelligibilité de la voix parlée

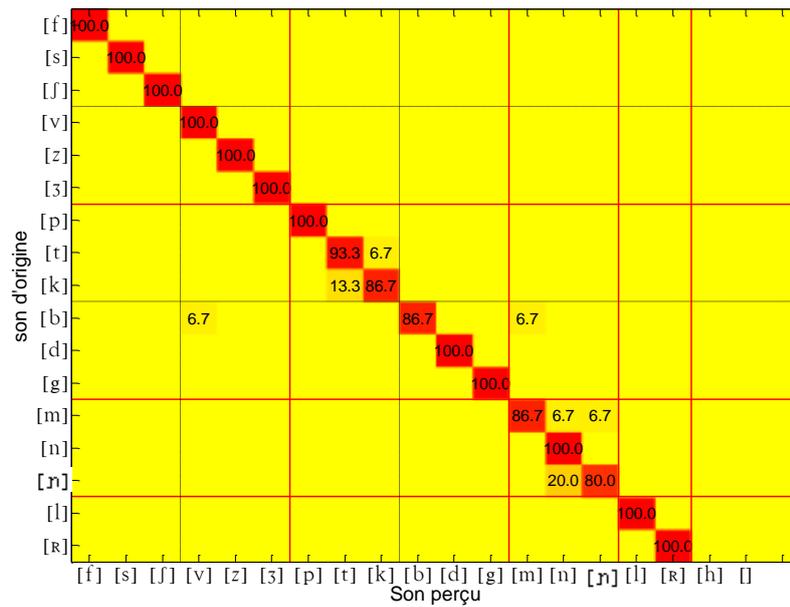


Figure D-4 : Matrice de confusion des consonnes pour la voix parlée du locuteur 1. Les nombres correspondent aux pourcentages de réponse.

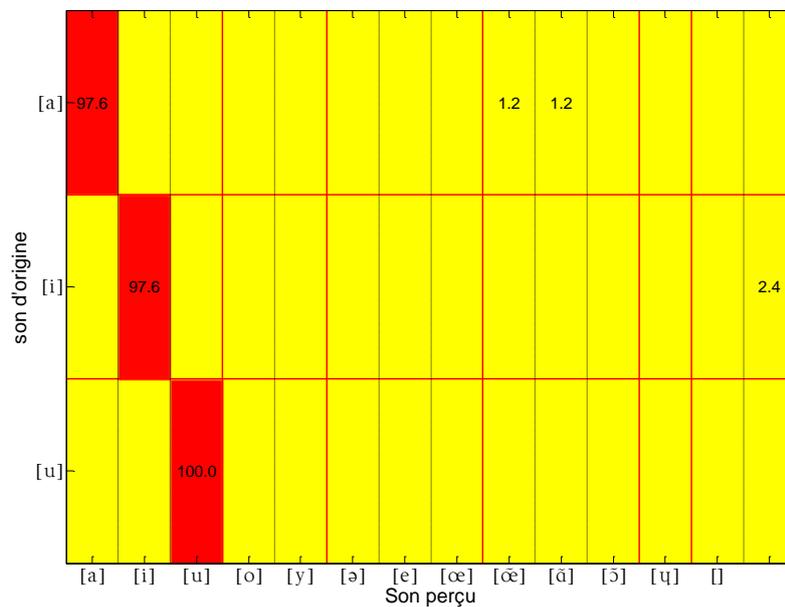


Figure D-5 : Matrice de confusion des voyelles pour la voix parlée du locuteur 1. Les nombres correspondent aux pourcentages de réponse.

### D.1.5 Intelligibilité de la voix chuchotée

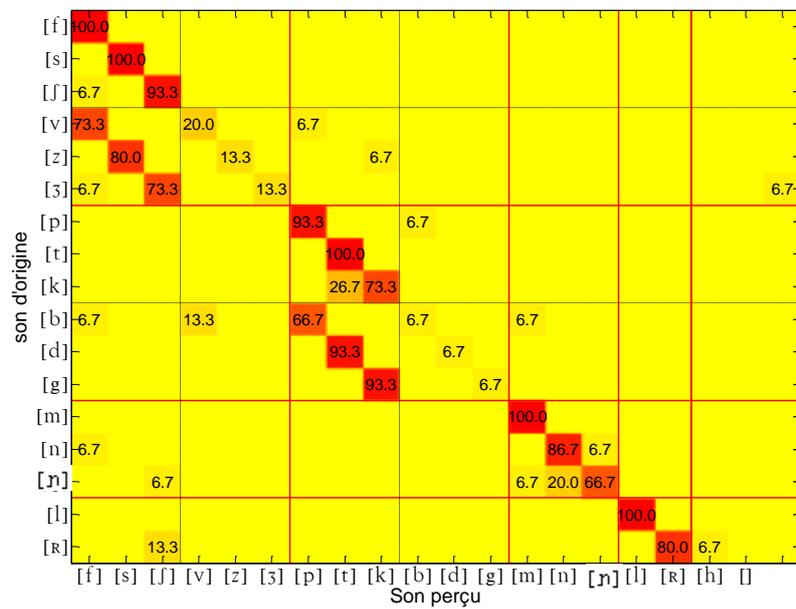


Figure D-6 : Matrice de confusion des consonnes pour la voix chuchotée du locuteur 1. Les nombres correspondent aux pourcentages de réponse.

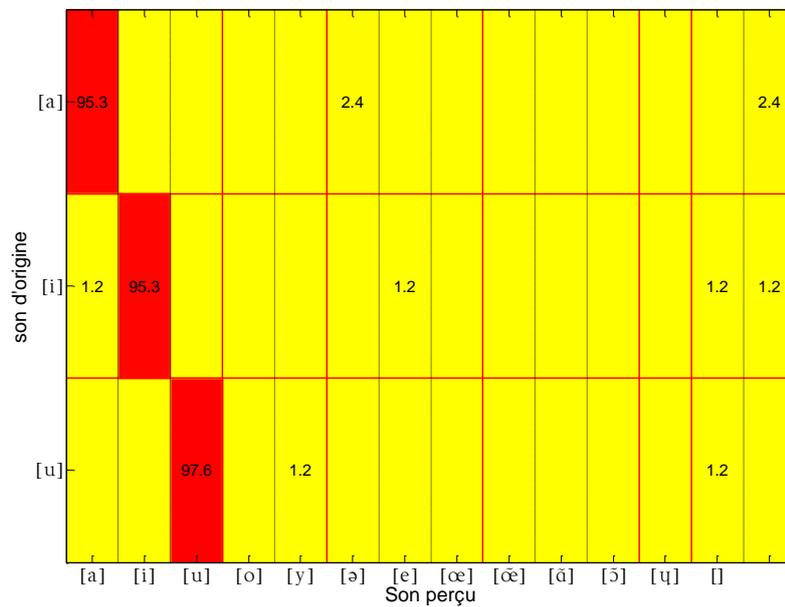


Figure D-7 : Matrice de confusion des voyelles pour la voix chuchotée du locuteur 1. Les nombres correspondent aux pourcentages de réponse.

### D.1.6 Intelligibilité de la voix criée

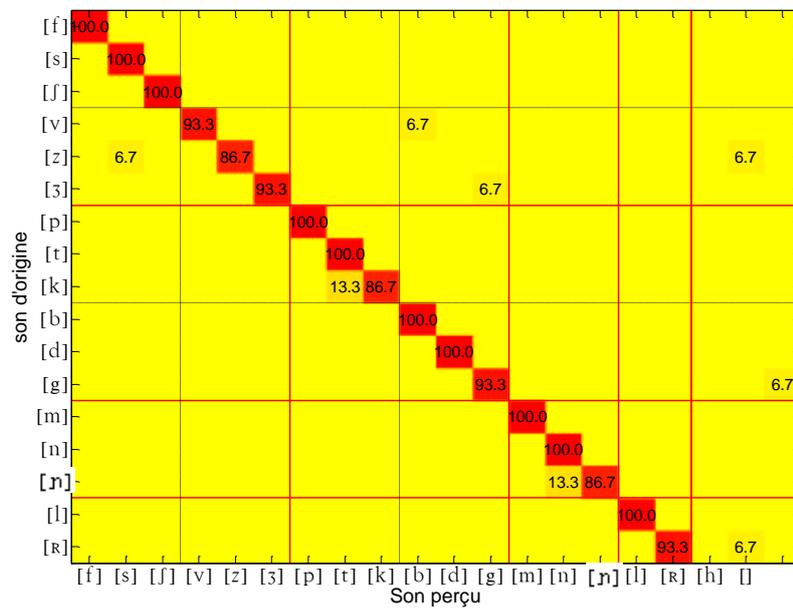


Figure D-8 : Matrice de confusion des consonnes pour la voix criée du locuteur 1. Les nombres correspondent aux pourcentages de réponse.

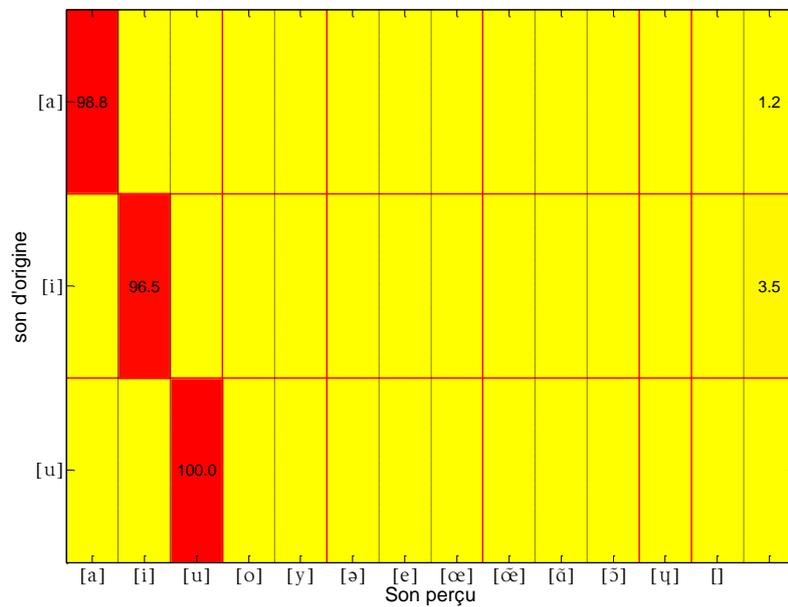


Figure D-9 : Matrice de confusion des voyelles pour la voix criée du locuteur 1. Les nombres correspondent aux pourcentages de réponse.

**D.1.7 Intelligibilité de la voix transformée chuchotée**

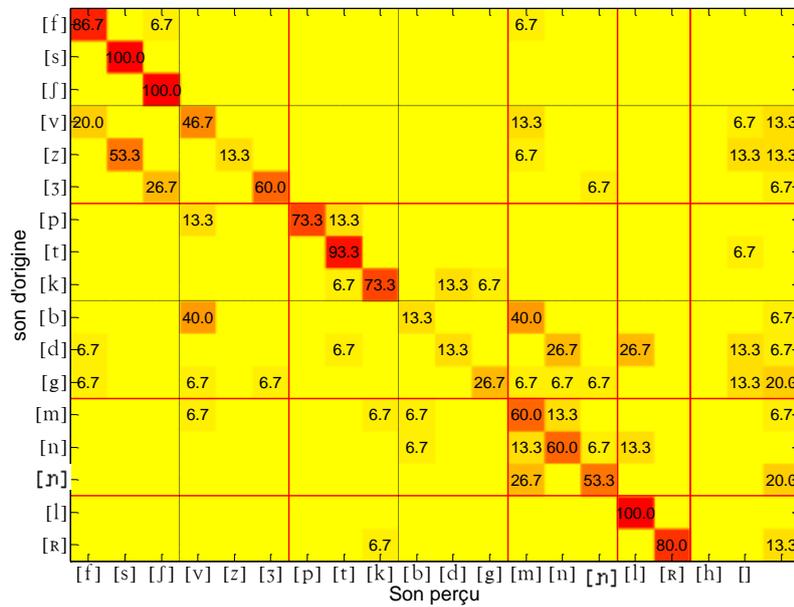


Figure D-10 : Matrice de confusion des consonnes pour la voix transformée chuchotée du locuteur 1. Les nombres correspondent aux pourcentages de réponse.

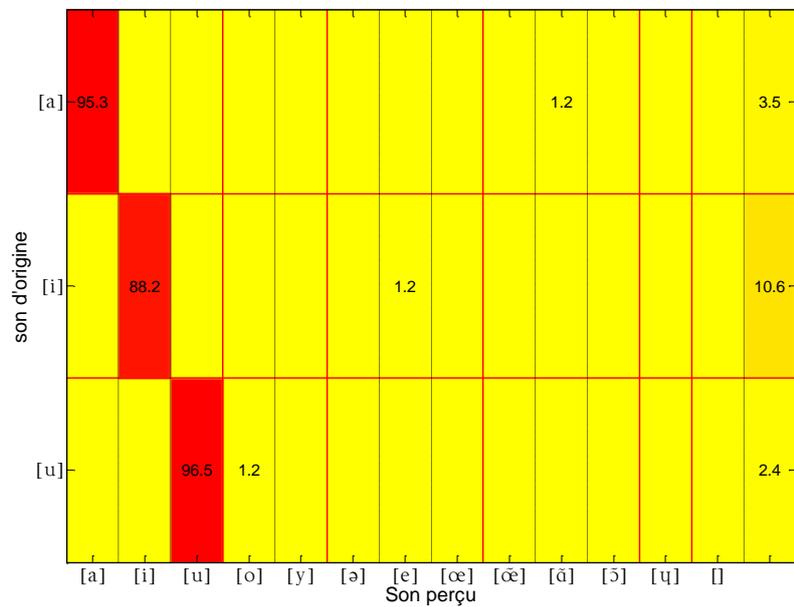


Figure D-11 : Matrice de confusion des voyelles pour la voix transformée chuchotée du locuteur 1. Les nombres correspondent aux pourcentages de réponse.

**D.1.8 Intelligibilité de la voix transformée criée**

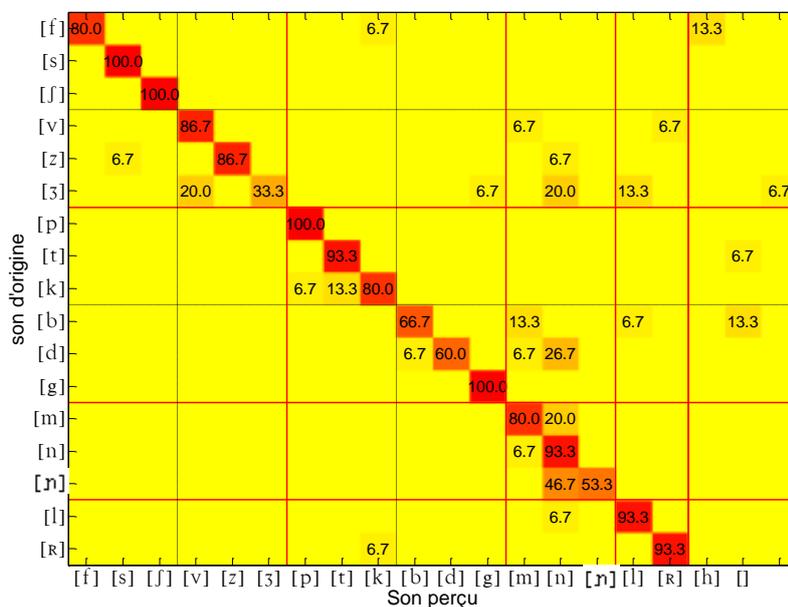


Figure D-12 : Matrice de confusion des consonnes pour la voix transformée criée du locuteur 1. Les nombres correspondent aux pourcentages de réponse.

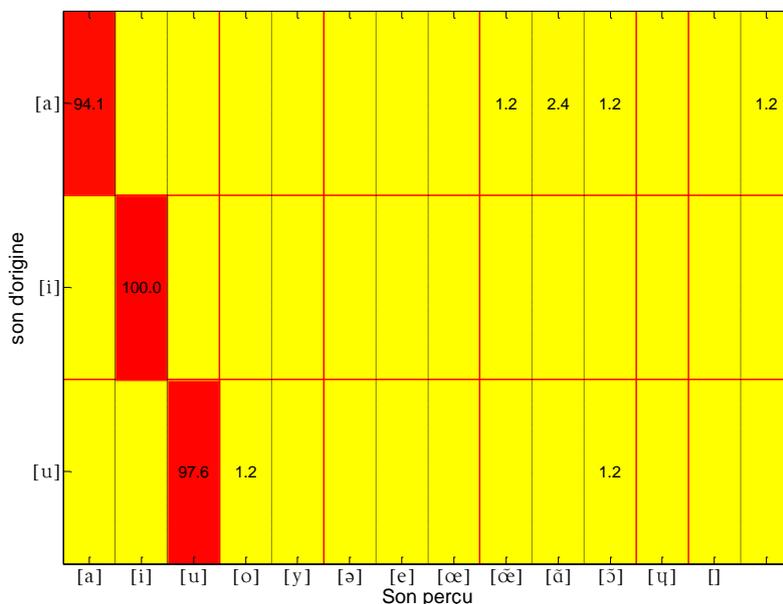


Figure D-13 : Matrice de confusion des voyelles pour la voix transformée criée du locuteur 1. Les nombres correspondent aux pourcentages de réponse.

## D.2 Locuteur 2

### D.2.1 Intelligibilité globale des logatomes

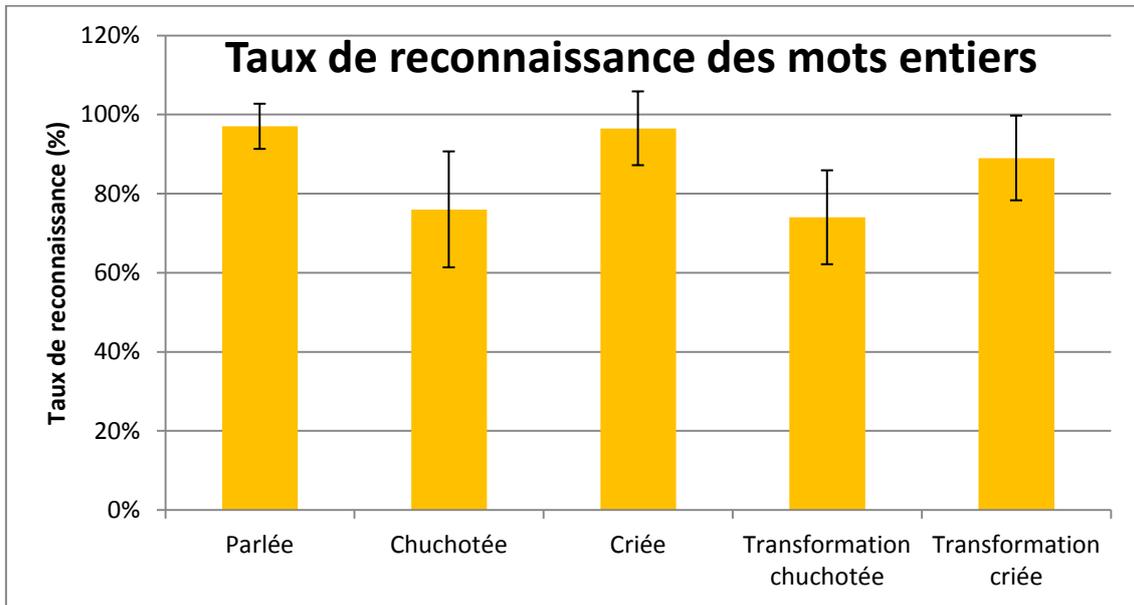


Figure D-14 : Moyennes des taux d'intelligibilité, obtenues par 25 sujets, des mots en fonction des modes de phonation ou des transformations pour le locuteur 2

### D.2.2 Intelligibilité globale des consonnes

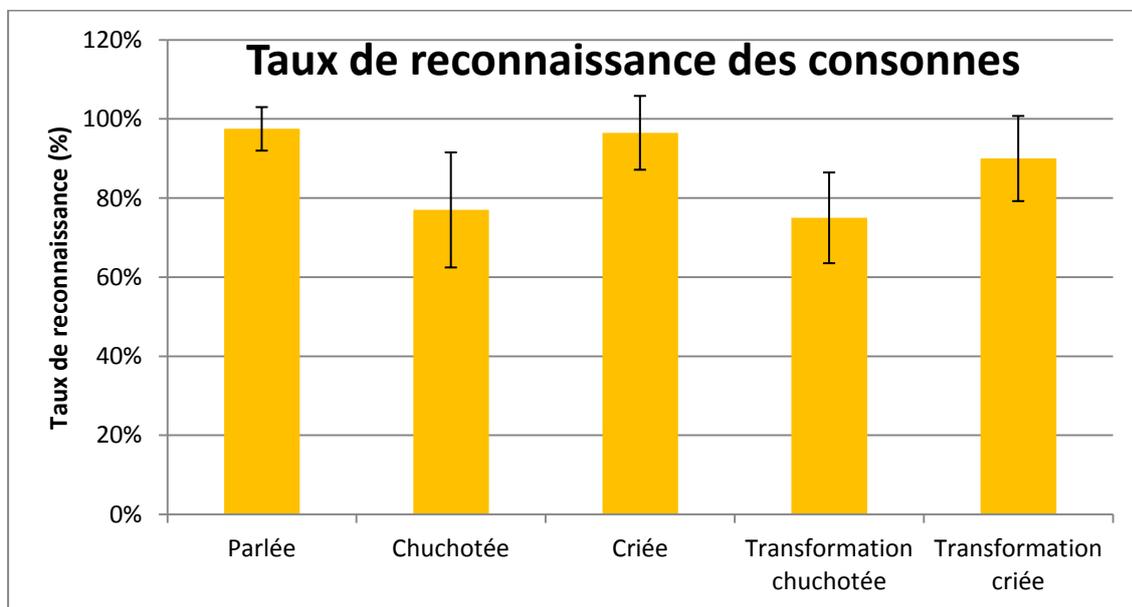


Figure D-15 : Moyennes des taux d'intelligibilité, obtenues par 25 sujets, des consonnes en fonction des modes de phonation ou des transformations pour le locuteur 2

### D.2.3 Intelligibilité globale des voyelles

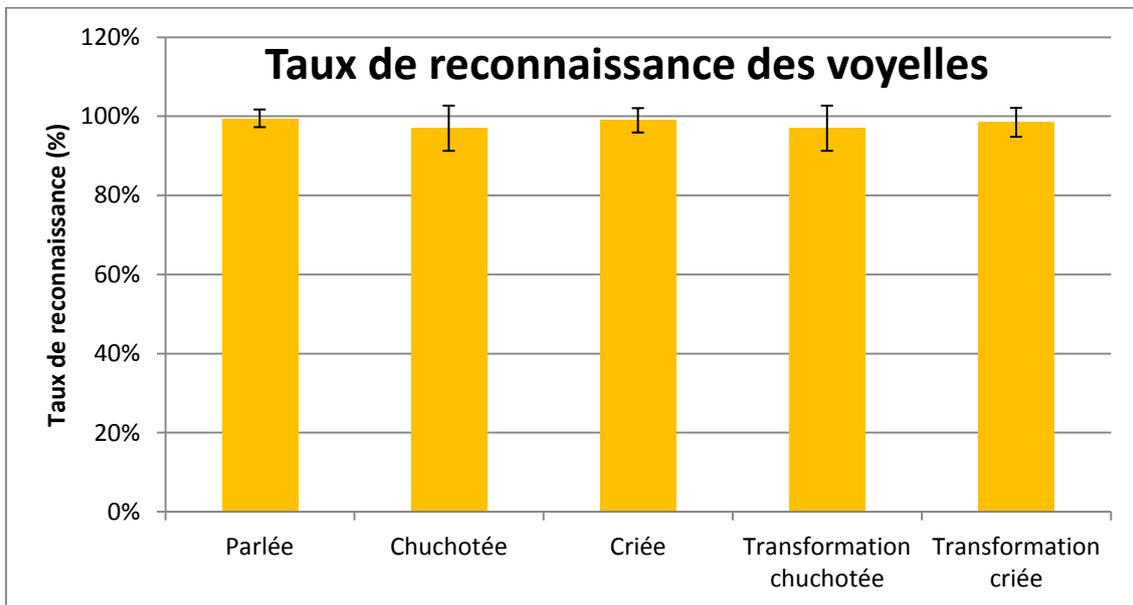


Figure D-16 : Moyennes des taux d'intelligibilité, obtenues par 25 sujets, des voyelles en fonction des modes de phonation ou des transformations pour le locuteur 2

### D.2.4 Intelligibilité de la voix parlée

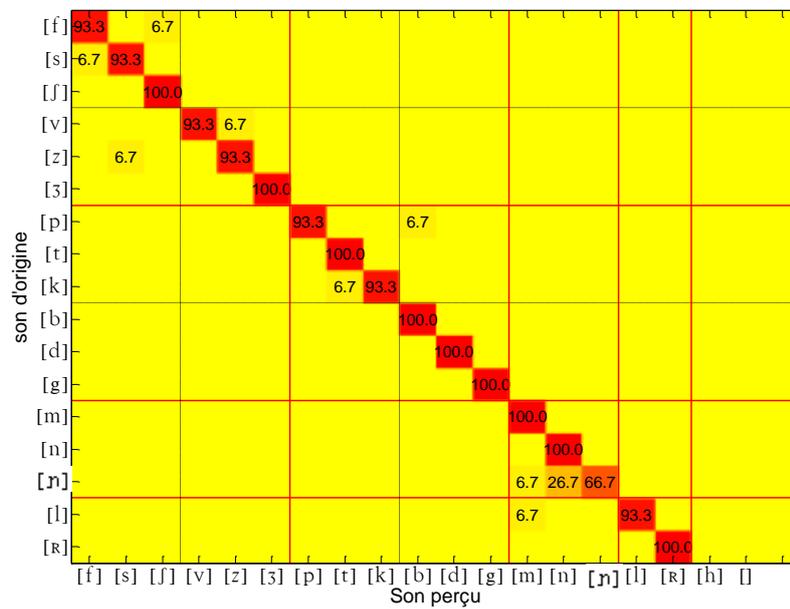


Figure D-17 : Matrice de confusion des consonnes pour la voix parlée du locuteur 2. Les nombres correspondent aux pourcentages de réponse.

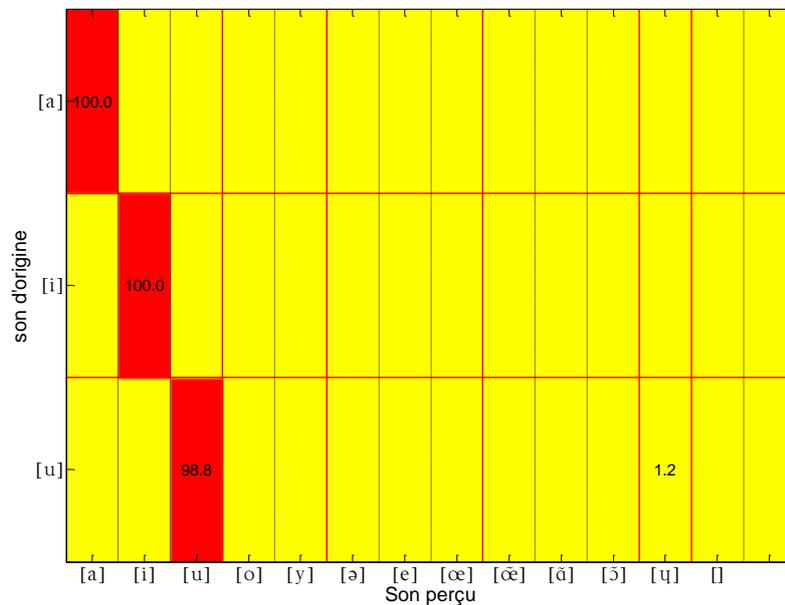


Figure D-18 : Matrice de confusion des voyelles pour la voix parlée du locuteur 2. Les nombres correspondent aux pourcentages de réponse.

**D.2.5 Intelligibilité de la voix chuchotée**

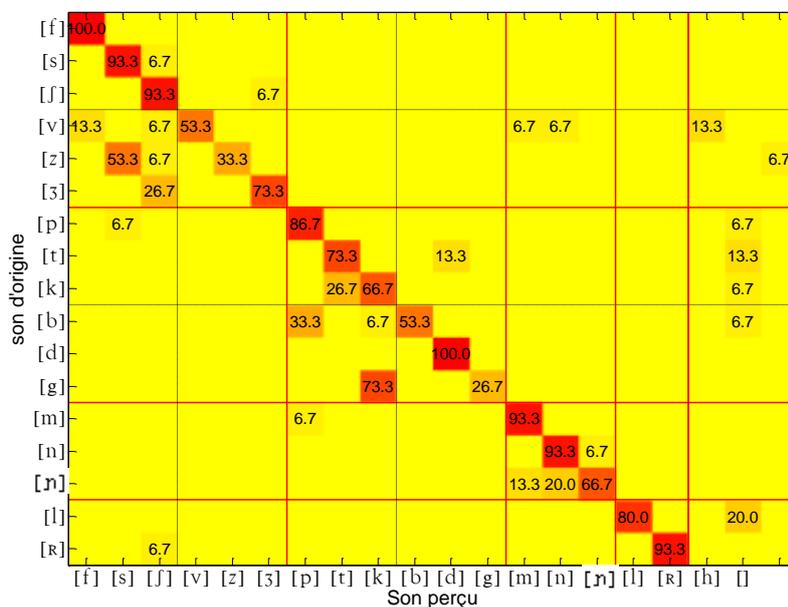


Figure D-19 : Matrice de confusion des consonnes pour la voix chuchotée du locuteur 2. Les nombres correspondent aux pourcentages de réponse.

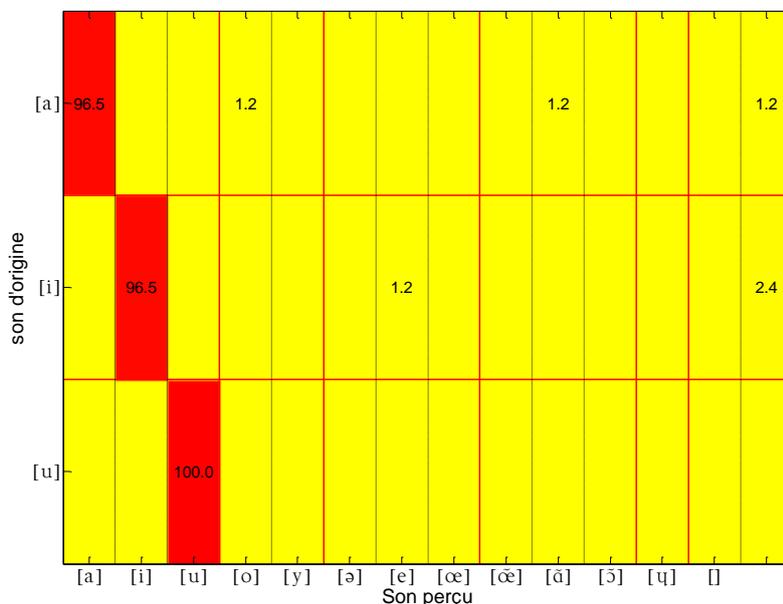


Figure D-20 : Matrice de confusion des voyelles pour la voix chuchotée du locuteur 2. Les nombres correspondent aux pourcentages de réponse.

### D.2.6 Intelligibilité de la voix criée

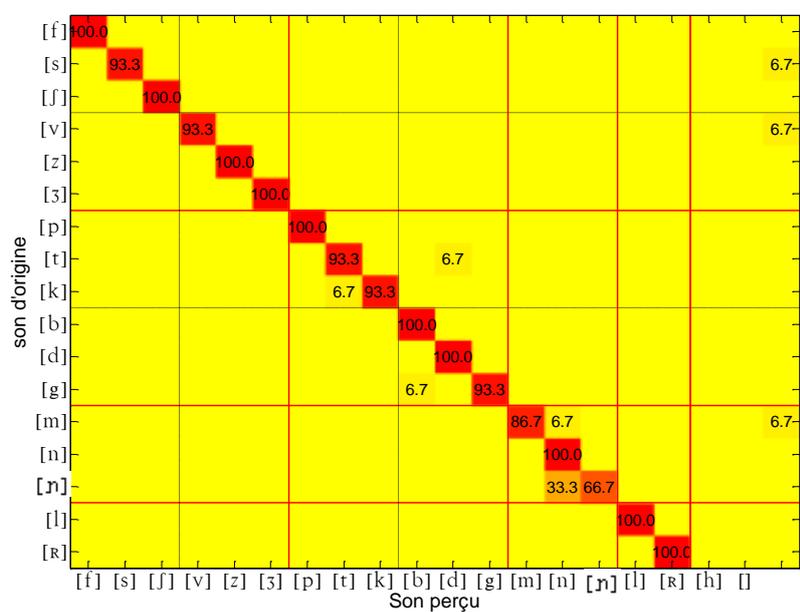


Figure D-21 : Matrice de confusion des consonnes pour la voix criée du locuteur 2. Les nombres correspondent aux pourcentages de réponse.

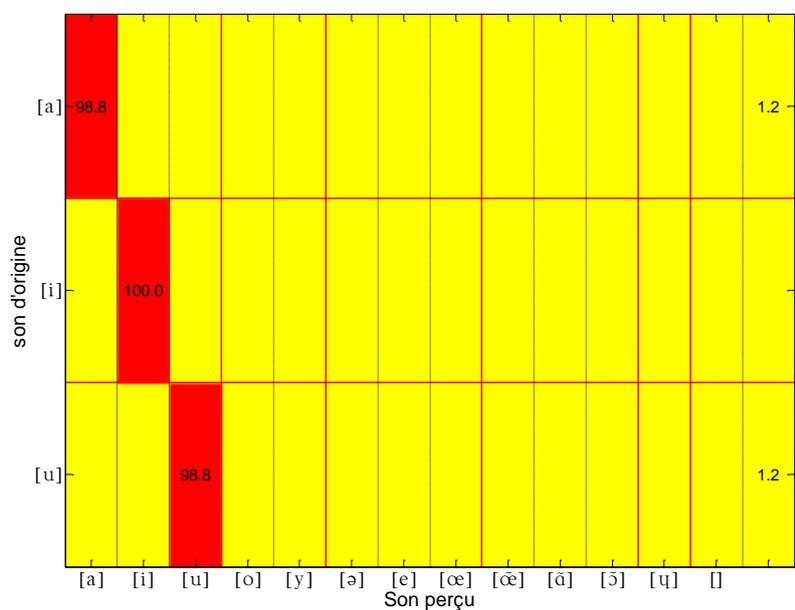


Figure D-22 : Matrice de confusion des voyelles pour la voix criée du locuteur 2. Les nombres correspondent aux pourcentages de réponse.

### D.2.7 Intelligibilité de la voix transformée chuchotée

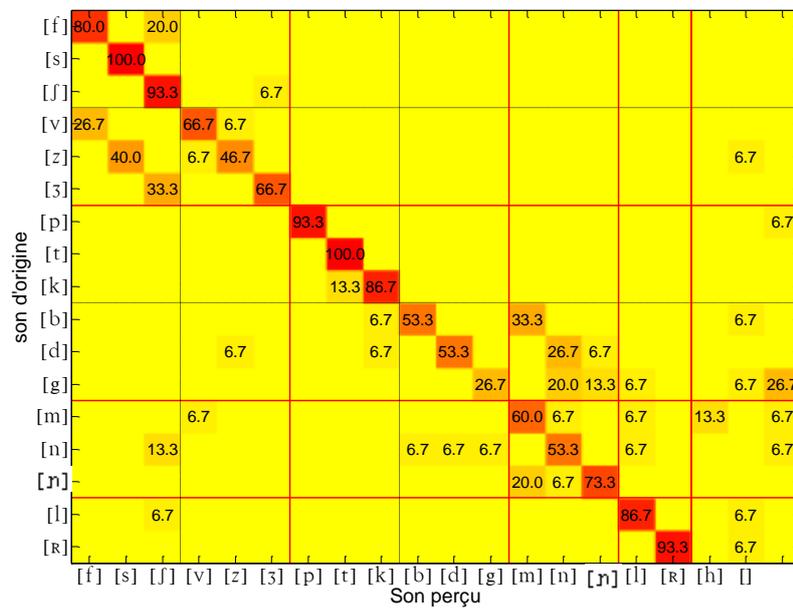


Figure D-23 : Matrice de confusion des consonnes pour la voix transformée chuchotée du locuteur 2. Les nombres correspondent aux pourcentages de réponse.

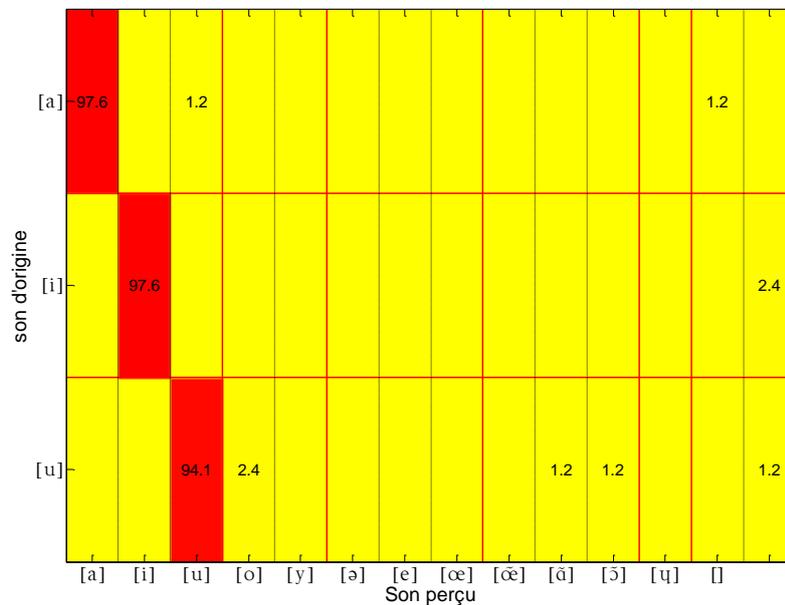


Figure D-24 : Matrice de confusion des voyelles pour la voix transformée chuchotée du locuteur 2. Les nombres correspondent aux pourcentages de réponse.

**D.2.8 Intelligibilité de la voix transformée criée**

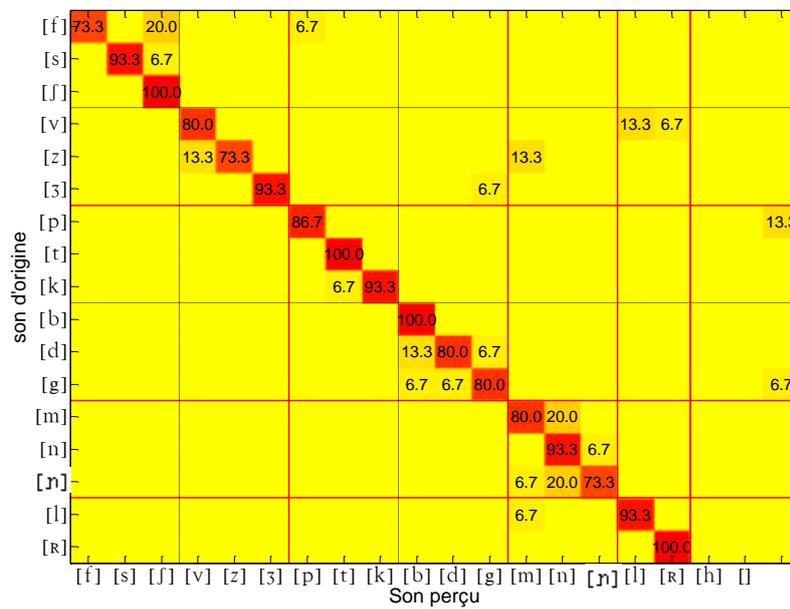


Figure D-25 : Matrice de confusion des consonnes pour la voix transformée criée du locuteur 2. Les nombres correspondent aux pourcentages de réponse.

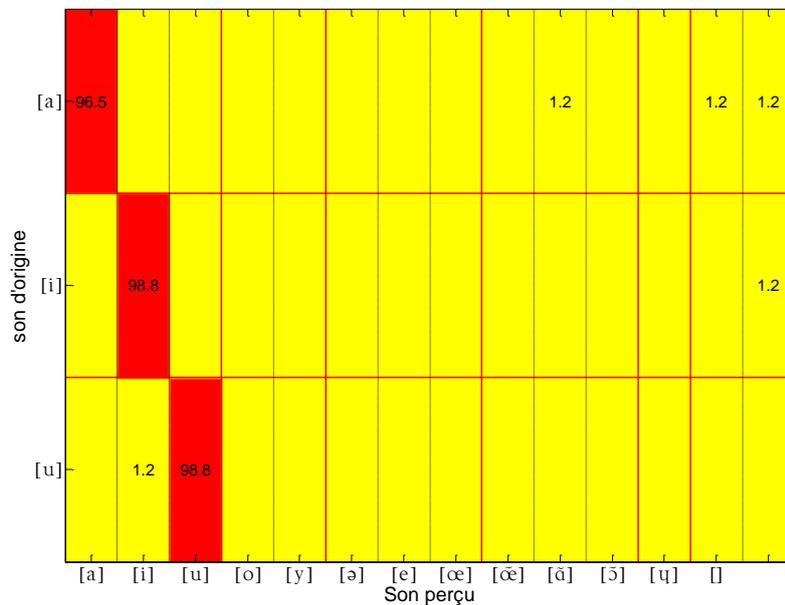


Figure D-26 : Matrice de confusion des voyelles pour la voix transformée criée du locuteur 2. Les nombres correspondent aux pourcentages de réponse.

### D.3 Locuteur 3

#### D.3.1 Intelligibilité globales des logatomes

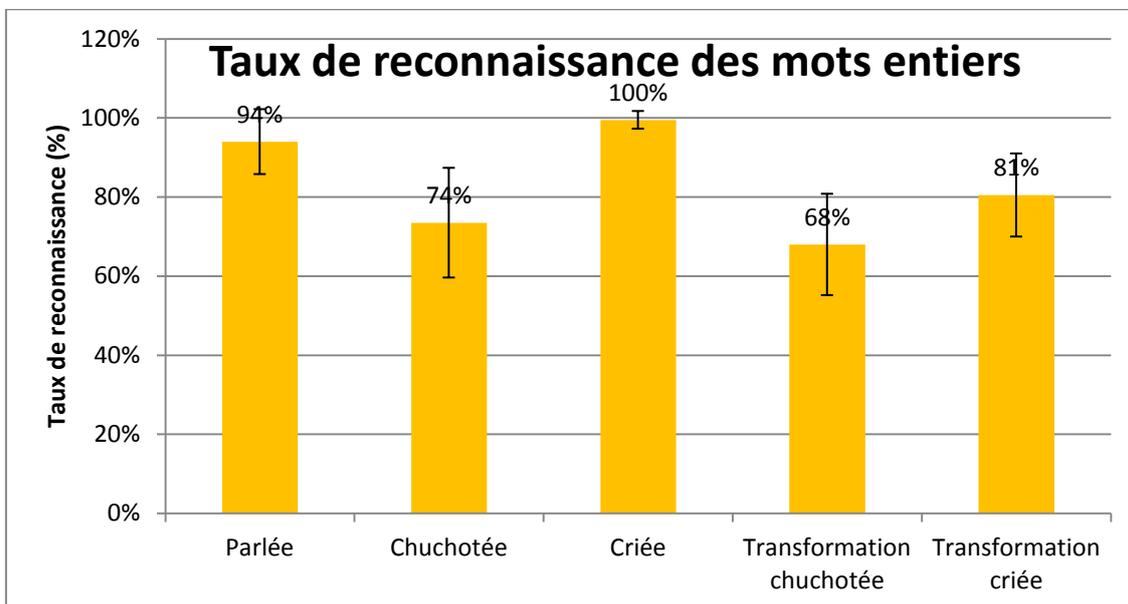


Figure D-27 : Moyennes des taux d'intelligibilité, obtenues par 25 sujets, des mots en fonction des modes de phonation ou des transformations pour le locuteur 3

#### D.3.2 Intelligibilité globales des consonnes

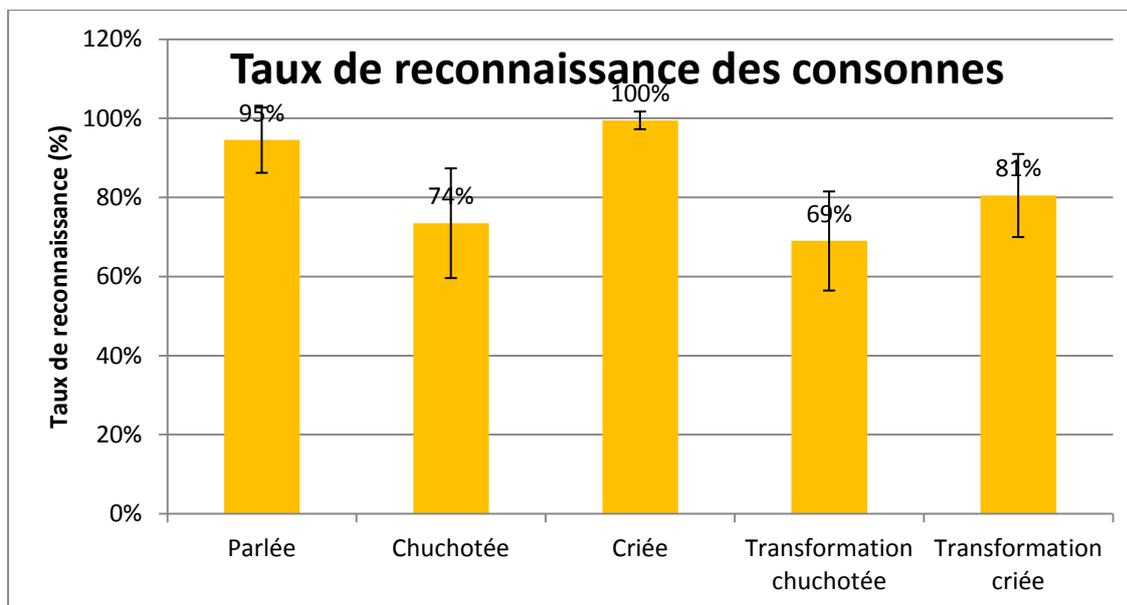


Figure D-28 : Moyennes des taux d'intelligibilité, obtenues par 25 sujets, des consonnes en fonction des modes de phonation ou des transformations pour le locuteur 3

### D.3.3 Intelligibilité globales des voyelles

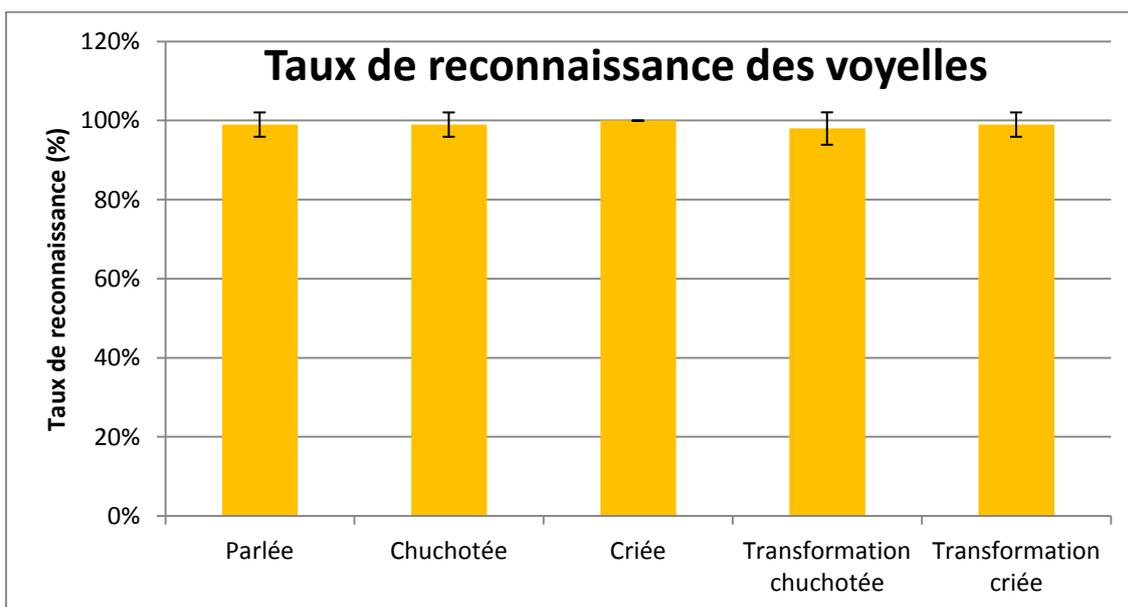


Figure D-29 : Moyennes des taux d'intelligibilité, obtenues par 25 sujets, des voyelles en fonction des modes de phonation ou des transformations pour le locuteur 3

**D.3.4 Intelligibilité de la voix parlée**

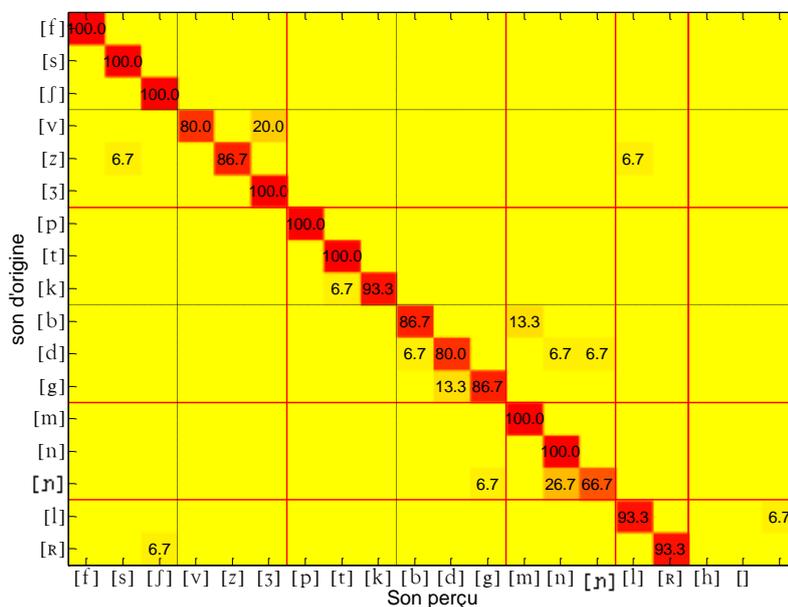


Figure D-30 : Matrice de confusion des consonnes pour la voix parlée du locuteur 3. Les nombres correspondent aux pourcentages de réponse.

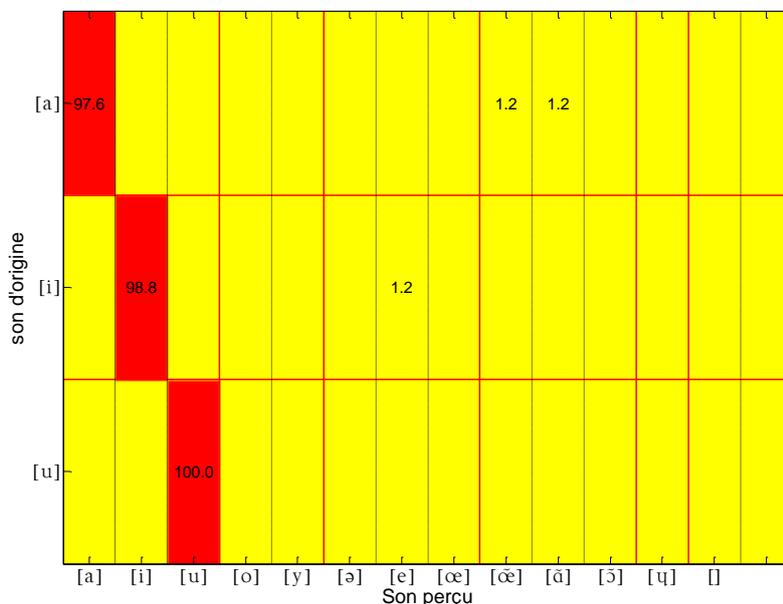


Figure D-31 : Matrice de confusion des voyelles pour la voix parlée du locuteur 3. Les nombres correspondent aux pourcentages de réponse.

### D.3.5 Intelligibilité de la voix chuchotée

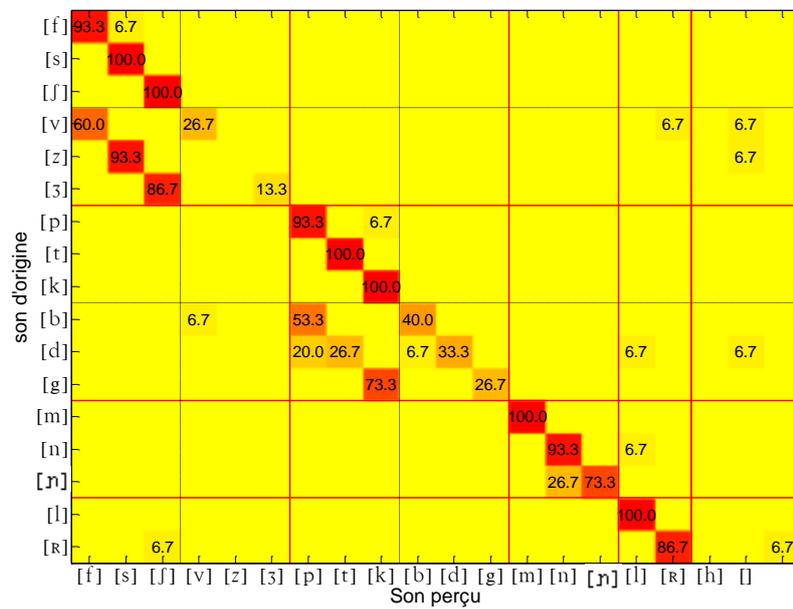


Figure D-32 : Matrice de confusion des consonnes pour la voix chuchotée du locuteur 3. Les nombres correspondent aux pourcentages de réponse.

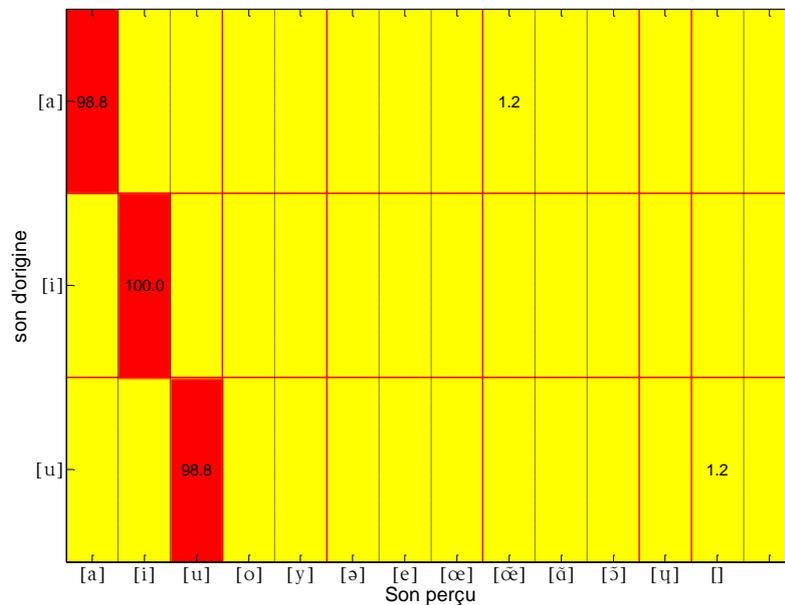


Figure D-33 : Matrice de confusion des voyelles pour la voix chuchotée du locuteur 3. Les nombres correspondent aux pourcentages de réponse.

### D.3.6 Intelligibilité de la voix criée

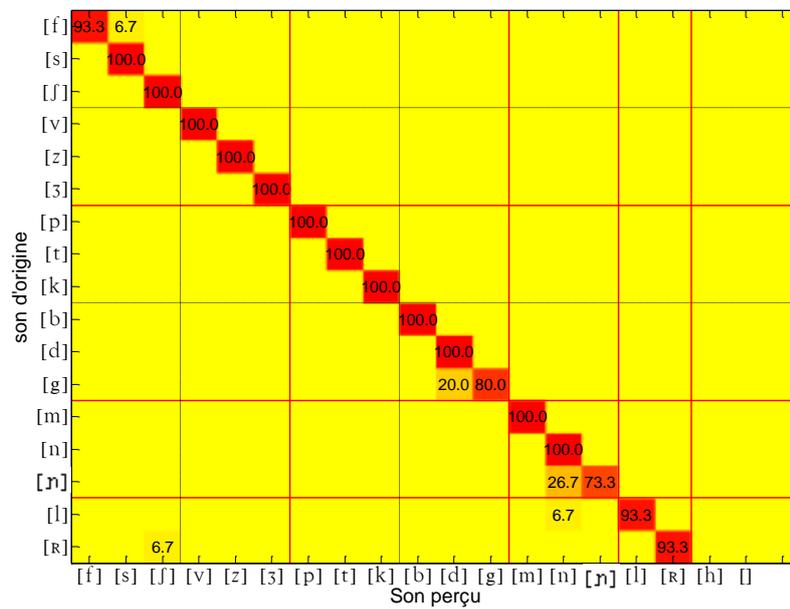


Figure D-34 : Matrice de confusion des consonnes pour la voix criée du locuteur 3. Les nombres correspondent aux pourcentages de réponse.

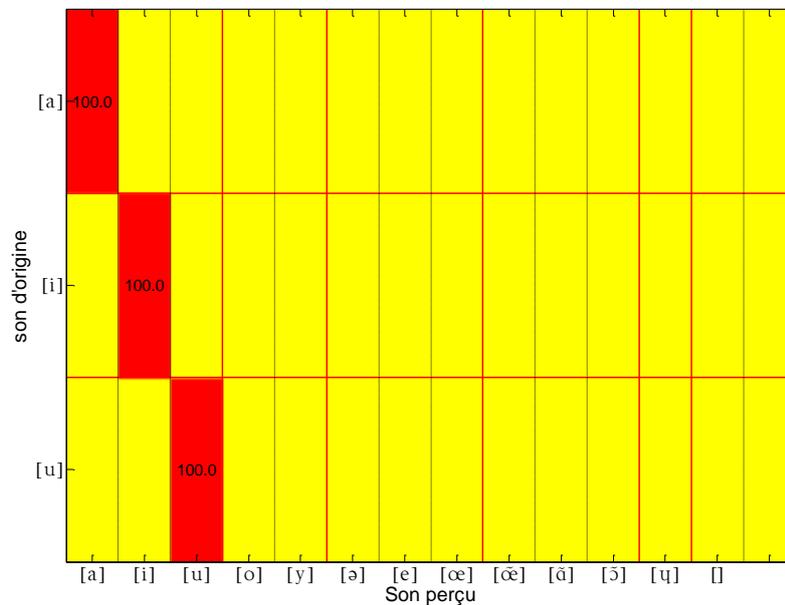


Figure D-35 : Matrice de confusion des voyelles pour la voix criée du locuteur 3. Les nombres correspondent aux pourcentages de réponse.

**D.3.7 Intelligibilité de la voix transformée chuchotée**

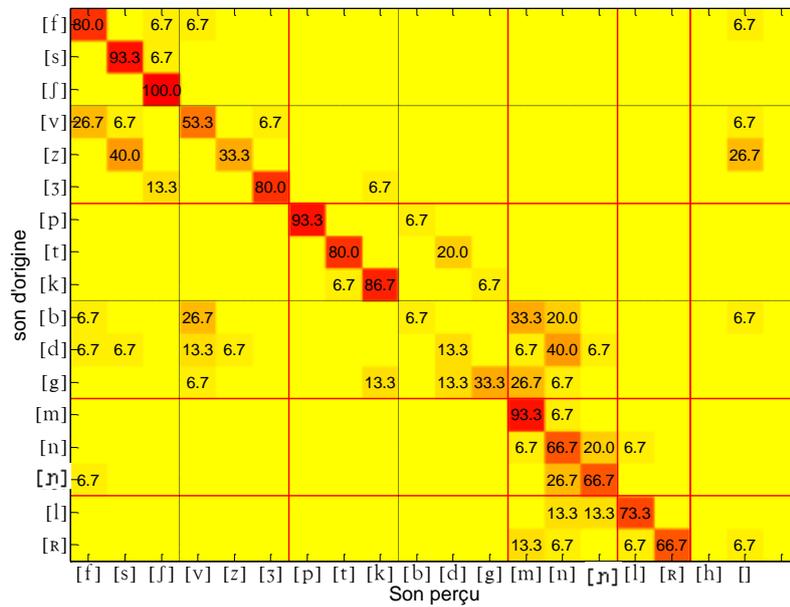


Figure D-36 : Matrice de confusion des consonnes pour la voix transformée chuchotée du locuteur 3. Les nombres correspondent aux pourcentages de réponse.

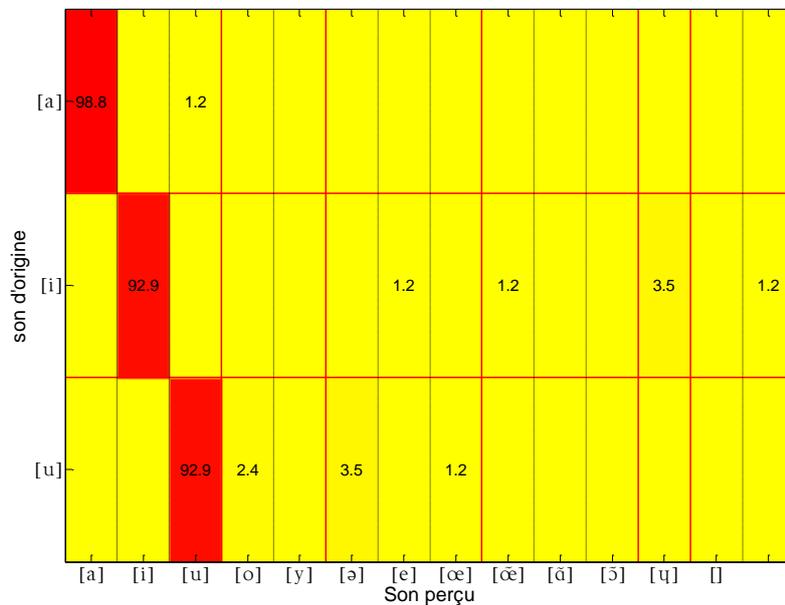


Figure D-37 : Matrice de confusion des voyelles pour la voix transformée chuchotée du locuteur 3. Les nombres correspondent aux pourcentages de réponse.

### D.3.8 Intelligibilité de la voix transformée criée

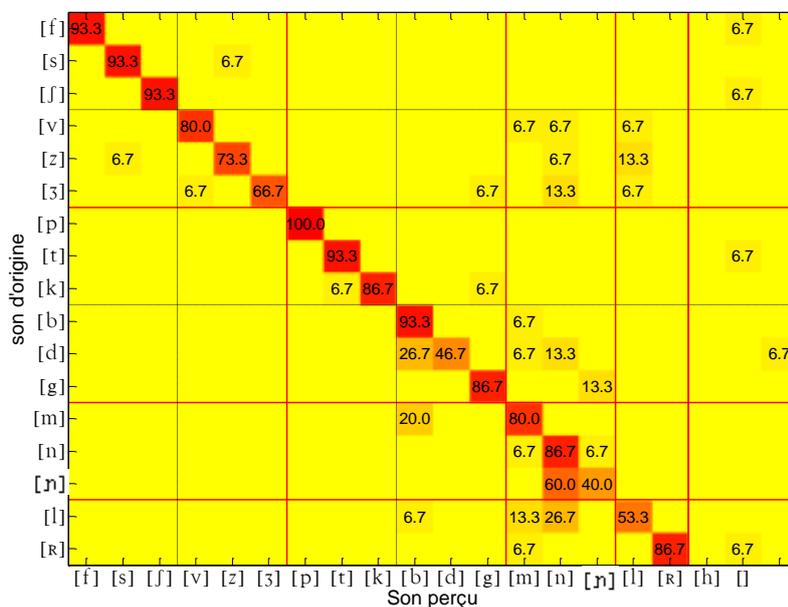


Figure D-38 : Matrice de confusion des consonnes pour la voix transformée criée du locuteur 3. Les nombres correspondent aux pourcentages de réponse.

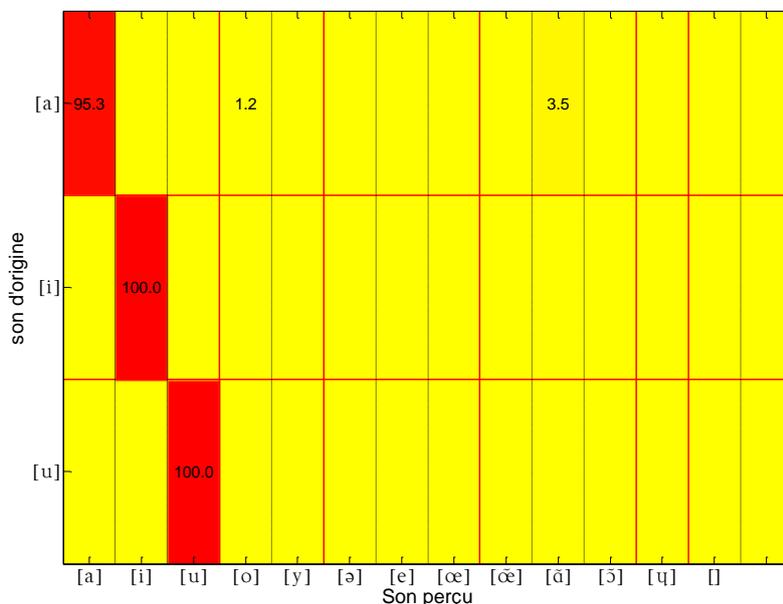


Figure D-39 : Matrice de confusion des voyelles pour la voix transformée criée du locuteur 3. Les nombres correspondent aux pourcentages de réponse.



# Annexe E

## **Valeurs caractéristiques des contours de F0 des logatomes**

---

Cette annexe donne les différents points caractéristiques de F0 présentés dans le chapitre 9 pour l'ensemble des logatomes et l'ensemble des locuteurs. Les valeurs sont exprimées en demi-tons dont la référence est 50 Hz.

## E.1 Les logatomes CV

Locuteur 1	/a/			
	F0 <sub>C1</sub> -init	F0 <sub>V1</sub> -init	F0 <sub>V1</sub> -max	F0 <sub>V1</sub> -finale
Fricatives voisées	11,8 0,8	22,9 0,5	29,7 0,9	29,5 0,6
Fricatives non-voisées	/ /	25,9 0,9	30,1 1,4	29,7 0,7
Occlusives voisées	13,9 2,0	25,4 1,9	30,7 1,0	30,1 0,7
Occlusives non-voisées	/ /	25,3 2,3	29,2 3,2	28,6 2,2
Nasales	16,5 2,1	25,8 0,2	30,9 0,8	30,3 0,4
Liquides	11,9 0,5	18,4 8,5	31,3 0,4	30,3 0,7
Moyenne	13,5	23,9	30,3	29,7
	/i/			
Fricatives voisées	13,5 1,2	25,8 5,2	31,7 2,0	29,0 1,2
Fricatives non-voisées	/ /	27,3 1,4	32,6 0,3	30,3 1,4
Occlusives voisées	15,8 3,3	22,0 3,8	31,0 1,3	29,3 0,7
Occlusives non-voisées	/ /	27,7 2,4	30,7 2,0	29,5 2,0
Nasales	17,5 1,5	29,1 1,2	32,1 1,1	30,3 1,2
Liquides	13,4 2,9	24,4 2,0	31,3 1,8	29,5 0,8
Moyenne	15,0	26,0	31,6	29,7
	/u/			
Fricatives voisées	16,3 2,4	27,2 2,1	31,4 0,3	30,1 0,6
Fricatives non-voisées	/ /	27,7 1,8	31,4 1,4	30,1 0,7
Occlusives voisées	14,5 1,9	21,9 2,3	31,7 0,7	30,2 0,4
Occlusives non-voisées	/ /	29,2 1,1	32,1 2,3	30,0 0,6
Nasales	17,1 1,3	27,2 0,6	32,0 1,2	30,9 0,2
Liquides	11,1 6,1	18,7 12,1	32,5 1,3	30,0 0,7
Moyenne	14,7	25,3	31,9	30,2

Locuteur 2	/a/			
	FO <sub>C1</sub> -init	FO <sub>V1</sub> -init	FO <sub>V1</sub> -max	FO <sub>V1</sub> -finale
Fricatives voisées	14,9 1,6	23,0 1,1	30,2 2,1	24,4 2,5
Fricatives non-voisées	/ /	26,6 1,0	30,0 1,3	24,8 1,7
Occlusives voisées	15,4 0,5	24,4 0,6	30,5 0,5	22,4 1,5
Occlusives non-voisées	/ /	25,9 1,4	31,0 0,7	22,7 2,8
Nasales	16,1 0,6	27,7 2,6	31,3 1,1	22,4 2,9
Liquides	15,6 0,0	25,8 3,2	31,1 0,6	20,8 1,2
Moyenne	15,5	25,6	30,7	22,9
	/i/			
Fricatives voisées	16,6 0,5	27,8 4,2	32,5 1,0	26,0 5,9
Fricatives non-voisées	/ /	27,6 1,5	30,6 0,8	26,9 2,0
Occlusives voisées	15,5 1,2	27,8 3,4	32,1 1,8	28,2 1,4
Occlusives non-voisées	/ /	28,6 2,1	31,6 0,5	25,4 3,0
Nasales	16,3 0,8	31,2 1,5	33,1 0,3	28,3 1,2
Liquides	17,6 0,1	30,4 0,9	32,7 0,2	25,7 2,3
Moyenne	16,5	28,9	32,1	26,7
	/u/			
Fricatives voisées	16,0 0,6	29,9 0,4	32,7 0,6	26,7 4,1
Fricatives non-voisées	/ /	30,3 1,3	32,5 0,3	25,7 2,2
Occlusives voisées	16,3 0,6	29,8 1,5	32,4 1,1	27,1 1,0
Occlusives non-voisées	/ /	29,8 0,4	32,4 0,7	29,4 1,0
Nasales	17,0 0,6	29,9 0,7	32,7 0,5	29,3 2,0
Liquides	16,5 0,2	29,0 1,1	32,3 0,4	27,8 2,2
Moyenne	16,4	29,8	32,5	27,7

<b>Locuteur 3</b>	<b>/a/</b>			
	<b>F0<sub>C1</sub>-init</b>	<b>F0<sub>V1</sub>-init</b>	<b>F0<sub>V1</sub>-max</b>	<b>F0<sub>V1</sub>-finale</b>
<b>Fricatives voisées</b>	<b>19,8</b> 0,6	<b>26,7</b> 1,5	<b>33,6</b> 1,4	<b>31,0</b> 1,1
<b>Fricatives non-voisées</b>	<b>/</b> /	<b>30,9</b> 0,7	<b>34,1</b> 0,1	<b>30,0</b> 0,3
<b>Occlusives voisées</b>	<b>20,0</b> 1,4	<b>25,8</b> 3,3	<b>33,6</b> 1,2	<b>31,2</b> 0,9
<b>Occlusives non-voisées</b>	<b>/</b> /	<b>29,7</b> 0,7	<b>34,1</b> 0,4	<b>31,2</b> 1,7
<b>Nasales</b>	<b>22,8</b> 0,7	<b>29,9</b> 1,4	<b>33,5</b> 0,7	<b>30,8</b> 0,8
<b>Liquides</b>	<b>20,0</b> 0,1	<b>26,9</b> 1,0	<b>33,5</b> 1,2	<b>31,5</b> 1,0
<b>Moyenne</b>	<b>20,6</b>	<b>28,3</b>	<b>33,7</b>	<b>30,9</b>
	<b>/i/</b>			
<b>Fricatives voisées</b>	<b>20,1</b> 1,5	<b>27,8</b> 2,0	<b>34,4</b> 0,6	<b>33,2</b> 1,0
<b>Fricatives non-voisées</b>	<b>/</b> /	<b>31,0</b> 0,5	<b>34,4</b> 0,8	<b>32,4</b> 1,9
<b>Occlusives voisées</b>	<b>21,1</b> 0,5	<b>26,1</b> 2,4	<b>34,6</b> 0,6	<b>32,1</b> 2,4
<b>Occlusives non-voisées</b>	<b>/</b> /	<b>30,8</b> 0,6	<b>34,8</b> 0,1	<b>32,4</b> 2,0
<b>Nasales</b>	<b>22,7</b> 1,1	<b>30,3</b> 0,7	<b>34,7</b> 1,0	<b>31,1</b> 1,7
<b>Liquides</b>	<b>22,9</b> 2,5	<b>31,0</b> 2,0	<b>34,8</b> 0,8	<b>32,0</b> 3,2
<b>Moyenne</b>	<b>21,7</b>	<b>29,5</b>	<b>34,6</b>	<b>32,2</b>
	<b>/u/</b>			
<b>Fricatives voisées</b>	<b>22,5</b> 0,9	<b>29,7</b> 1,6	<b>36,5</b> 0,6	<b>32,3</b> 0,7
<b>Fricatives non-voisées</b>	<b>/</b> /	<b>31,1</b> 0,7	<b>35,3</b> 0,2	<b>31,2</b> 1,7
<b>Occlusives voisées</b>	<b>21,3</b> 0,7	<b>27,9</b> 2,3	<b>36,1</b> 0,4	<b>32,9</b> 0,9
<b>Occlusives non-voisées</b>	<b>/</b> /	<b>31,7</b> 0,8	<b>35,6</b> 0,9	<b>32,2</b> 1,2
<b>Nasales</b>	<b>22,7</b> 0,3	<b>30,3</b> 2,7	<b>35,3</b> 0,6	<b>33,2</b> 0,3
<b>Liquides</b>	<b>24,7</b> 0,7	<b>31,2</b> 0,7	<b>37,5</b> 0,9	<b>32,5</b> 0,6
<b>Moyenne</b>	<b>22,8</b>	<b>30,3</b>	<b>36,0</b>	<b>32,4</b>

## E.2 Les logatomes CVC

Locuteur 1	/a/				
	F0 <sub>C1</sub> -init	F0 <sub>V1</sub> -init	F0 <sub>V1</sub> -max	F0 <sub>V1</sub> -finale	F0 <sub>C2</sub> -finale
Fricatives voisées	13,8 0,6	23,3 2,5	31,3 1,1	24,1 1,4	13,4 2,9
Fricatives non-voisées	/	26,8 2,3	31,4 1,2	24,7 0,4	/
Occlusives voisées	15,1 0,8	22,7 1,9	30,3 0,5	26,2 1,5	16,8 1,9
Occlusives non-voisées	/	27,0 1,4	31,8 1,8	23,0 1,8	/
Nasales	16,4 0,4	27,9 1,3	31,7 1,1	30,2 0,8	16,6 6,2
Liquides	14,6 2,6	27,2 0,8	32,3 0,1	27,8 3,1	17,4 2,1
Moyenne	15,0	25,8	31,5	26,0	16,1
	/i/				
Fricatives voisées	16,6 2,2	26,7 1,3	32,4 1,0	25,6 3,0	18,8 0,7
Fricatives non-voisées	/	30,2 1,2	34,8 0,9	30,0 3,8	/
Occlusives voisées	16,3 2,3	25,4 1,2	33,6 0,9	29,4 1,2	20,9 3,1
Occlusives non-voisées	/	29,1 1,2	33,6 1,8	26,9 2,2	/
Nasales	16,1 3,6	27,6 0,6	33,3 0,1	31,6 0,5	18,8 9,9
Liquides	11,6 7,4	26,2 1,6	33,3 0,3	29,5 2,5	26,0 1,9
Moyenne	15,2	27,5	33,5	28,8	21,1
	/u/				
Fricatives voisées	16,5 1,6	26,1 2,2	32,9 1,8	26,7 1,5	18,2 3,1
Fricatives non-voisées	/	29,9 0,9	34,2 0,8	28,2 0,5	/
Occlusives voisées	16,3 0,5	24,9 3,1	33,7 0,6	29,2 0,4	21,3 1,8
Occlusives non-voisées	/	29,0 0,6	33,1 1,3	26,8 0,5	/
Nasales	17,7 1,3	27,1 2,2	33,1 1,8	31,0 0,2	21,6 3,8
Liquides	14,7 2,4	28,9 0,8	33,9 0,2	29,2 3,5	15,0 9,1
Moyenne	16,3	27,7	33,5	28,5	19,0

<b>Locuteur 2</b>	<b>/a/</b>				
	<b>F0<sub>C1</sub>-init</b>	<b>F0<sub>V1</sub>-init</b>	<b>F0<sub>V1</sub>-max</b>	<b>F0<sub>V1</sub>-finale</b>	<b>F0<sub>C2</sub>-finale</b>
<b>Fricatives voisées</b>	<b>16,0</b> 0,5	<b>20,8</b> 1,3	<b>29,1</b> 1,5	<b>24,9</b> 2,9	<b>15,7</b> 1,0
<b>Fricatives non-voisées</b>	/	<b>25,5</b> 0,4	<b>29,4</b> 0,5	<b>26,3</b> 2,7	/
<b>Occlusives voisées</b>	<b>16,1</b> 0,4	<b>19,7</b> 0,8	<b>27,0</b> 2,1	<b>25,7</b> 1,4	<b>21,4</b> 3,7
<b>Occlusives non-voisées</b>	/	<b>24,4</b> 0,8	<b>28,5</b> 0,4	<b>25,3</b> 0,5	/
<b>Nasales</b>	<b>17,1</b> 0,7	<b>24,0</b> 3,9	<b>29,7</b> 0,4	<b>28,0</b> 1,2	<b>19,9</b> 4,3
<b>Liquides</b>	<b>15,8</b> 0,3	<b>23,1</b> 2,2	<b>29,4</b> 0,0	<b>28,5</b> 1,1	<b>12,7</b> 15,9
<b>Moyenne</b>	<b>16,3</b>	<b>22,9</b>	<b>28,9</b>	<b>26,4</b>	<b>17,4</b>
	<b>/i/</b>				
<b>Fricatives voisées</b>	<b>16,6</b> 0,8	<b>23,9</b> 2,3	<b>29,8</b> 0,9	<b>27,0</b> 2,5	<b>21,0</b> 7,8
<b>Fricatives non-voisées</b>	/	<b>26,8</b> 0,4	<b>30,2</b> 0,5	<b>25,8</b> 0,7	/
<b>Occlusives voisées</b>	<b>16,2</b> 0,8	<b>22,9</b> 0,6	<b>29,7</b> 0,3	<b>26,5</b> 2,6	<b>20,7</b> 3,7
<b>Occlusives non-voisées</b>	/	<b>27,7</b> 1,4	<b>30,3</b> 0,1	<b>26,5</b> 1,6	/
<b>Nasales</b>	<b>18,4</b> 0,6	<b>26,9</b> 1,7	<b>31,4</b> 0,3	<b>30,3</b> 0,7	<b>19,1</b> 5,7
<b>Liquides</b>	<b>16,3</b> 0,0	<b>23,9</b> 0,5	<b>29,6</b> 0,0	<b>27,6</b> 0,6	<b>20,9</b> 8,9
<b>Moyenne</b>	<b>16,9</b>	<b>25,4</b>	<b>30,2</b>	<b>27,3</b>	<b>20,4</b>
	<b>/u/</b>				
<b>Fricatives voisées</b>	<b>15,9</b> 0,4	<b>23,8</b> 1,3	<b>30,7</b> 0,6	<b>26,2</b> 1,9	<b>16,3</b> 1,4
<b>Fricatives non-voisées</b>	/	<b>28,3</b> 0,7	<b>30,8</b> 0,8	<b>27,4</b> 1,1	/
<b>Occlusives voisées</b>	<b>16,5</b> 0,5	<b>24,3</b> 2,7	<b>31,1</b> 0,6	<b>28,1</b> 1,6	<b>17,7</b> 0,9
<b>Occlusives non-voisées</b>	/	<b>27,8</b> 0,5	<b>31,2</b> 0,5	<b>27,2</b> 1,6	/
<b>Nasales</b>	<b>18,4</b> 1,1	<b>25,3</b> 2,0	<b>30,8</b> 1,1	<b>30,1</b> 0,9	<b>26,3</b> 2,8
<b>Liquides</b>	<b>17,2</b> 0,3	<b>26,3</b> 0,4	<b>30,8</b> 0,0	<b>27,2</b> 0,5	<b>21,0</b> 9,4
<b>Moyenne</b>	<b>17,0</b>	<b>26,0</b>	<b>30,9</b>	<b>27,7</b>	<b>20,3</b>

<b>Locuteur 3</b>	<b>/a/</b>				
	<b>F0<sub>C1</sub>-init</b>	<b>F0<sub>V1</sub>-init</b>	<b>F0<sub>V1</sub>-max</b>	<b>F0<sub>V1</sub>-finale</b>	<b>F0<sub>C2</sub>-finale</b>
<b>Fricatives voisées</b>	<b>19,6</b> 0,6	<b>27,6</b> 1,0	<b>34,5</b> 0,3	<b>27,6</b> 2,5	<b>16,7</b> 3,1
<b>Fricatives non-voisées</b>	/	<b>29,6</b> 1,2	<b>33,8</b> 0,9	<b>29,3</b> 2,3	/
<b>Occlusives voisées</b>	<b>19,7</b> 0,3	<b>24,8</b> 2,0	<b>34,2</b> 0,4	<b>29,3</b> 1,9	<b>22,3</b> 3,0
<b>Occlusives non-voisées</b>	/	<b>29,2</b> 2,0	<b>33,7</b> 1,9	<b>29,5</b> 1,8	/
<b>Nasales</b>	<b>21,8</b> 0,9	<b>28,6</b> 2,5	<b>34,7</b> 0,6	<b>32,5</b> 0,6	<b>16,0</b> 1,6
<b>Liquides</b>	<b>20,4</b> 2,2	<b>27,2</b> 3,4	<b>35,1</b> 0,3	<b>30,2</b> 3,0	<b>20,4</b> 4,2
<b>Moyenne</b>	<b>20,4</b>	<b>27,8</b>	<b>34,3</b>	<b>29,7</b>	<b>18,8</b>
	<b>/i/</b>				
<b>Fricatives voisées</b>	<b>21,0</b> 0,3	<b>27,2</b> 1,6	<b>34,9</b> 0,8	<b>30,8</b> 0,5	<b>18,1</b> 5,9
<b>Fricatives non-voisées</b>	/	<b>30,9</b> 0,5	<b>35,1</b> 1,2	<b>29,7</b> 2,9	/
<b>Occlusives voisées</b>	<b>20,8</b> 0,7	<b>28,7</b> 1,4	<b>34,6</b> 0,8	<b>29,8</b> 2,6	<b>20,7</b> 2,8
<b>Occlusives non-voisées</b>	/	<b>30,1</b> 1,1	<b>34,6</b> 2,2	<b>29,9</b> 3,0	/
<b>Nasales</b>	<b>22,4</b> 0,5	<b>30,8</b> 1,4	<b>35,5</b> 0,8	<b>34,2</b> 0,2	<b>15,5</b> 0,6
<b>Liquides</b>	<b>19,7</b> 4,6	<b>29,1</b> 3,4	<b>35,1</b> 2,3	<b>30,5</b> 6,2	<b>14,9</b> 0,5
<b>Moyenne</b>	<b>21,0</b>	<b>29,5</b>	<b>35,0</b>	<b>30,8</b>	<b>17,3</b>
	<b>/u/</b>				
<b>Fricatives voisées</b>	<b>21,4</b> 1,4	<b>29,1</b> 2,8	<b>35,4</b> 1,7	<b>31,2</b> 1,4	<b>14,9</b> 2,7
<b>Fricatives non-voisées</b>	/	<b>30,9</b> 0,3	<b>35,4</b> 1,0	<b>28,8</b> 5,0	/
<b>Occlusives voisées</b>	<b>20,8</b> 0,1	<b>28,2</b> 0,4	<b>35,4</b> 0,5	<b>29,7</b> 2,2	<b>20,8</b> 2,7
<b>Occlusives non-voisées</b>	/	<b>30,2</b> 1,4	<b>34,6</b> 1,7	<b>29,1</b> 1,7	/
<b>Nasales</b>	<b>23,2</b> 0,7	<b>29,9</b> 0,5	<b>35,5</b> 1,5	<b>33,1</b> 0,8	<b>15,7</b> 1,8
<b>Liquides</b>	<b>23,4</b> 1,2	<b>32,0</b> 0,0	<b>36,1</b> 0,8	<b>30,1</b> 1,9	<b>20,1</b> 9,7
<b>Moyenne</b>	<b>22,2</b>	<b>30,0</b>	<b>35,4</b>	<b>30,3</b>	<b>17,9</b>

## E.3 Les logatomes VCV

Locuteur 1	/a/						
	FO <sub>V1</sub> -init	FO <sub>V1</sub> -max	FO <sub>V1</sub> -finale	FO <sub>c1</sub> -min	FO <sub>V2</sub> -init	FO <sub>V2</sub> -max	FO <sub>V2</sub> -finale
Fricatives voisées	16,9	27,7	27,5	24,2	28,3	31,9	30,7
	4,4	0,2	0,0	0,4	0,2	0,4	0,3
Fricatives non-voisées	21,9	26,5	25,8		28,4	30,8	29,5
	1,5	1,1	0,7		0,9	0,6	0,4
Occlusives voisées	19,1	26,8	25,5	23,1	27,4	31,5	30,2
	1,1	0,6	1,1	0,7	1,0	1,5	1,0
Occlusives non-voisées	20,7	26,0	23,5		29,2	31,5	29,9
	2,9	1,8	1,8		1,8	2,1	2,5
Nasales	19,9	28,0	28,0	27,9	29,9	32,9	31,4
	5,2	2,1	2,1	2,2	1,6	0,3	0,9
Liquides	19,6	27,5	27,5	26,1	28,5	32,3	30,7
	0,4	0,1	0,2	2,1	0,9	0,5	0,6
Moyenne	19,7	27,1	26,3	25,3	28,6	31,8	30,4
	/i/						
Fricatives voisées	21,3	29,2	28,0	25,3	29,5	33,2	31,4
	5,0	1,3	2,0	1,9	1,4	1,4	0,9
Fricatives non-voisées	23,9	30,2	27,5		30,6	33,0	30,9
	2,2	0,2	1,2		0,6	0,7	0,6
Occlusives voisées	23,8	30,2	27,6	25,4	29,7	33,2	30,3
	4,0	2,1	0,8	1,6	1,8	1,1	0,8
Occlusives non-voisées	21,1	28,9	23,6		30,9	32,4	30,3
	3,4	2,4	3,8		2,5	2,0	1,9
Nasales	24,5	29,9	29,6	29,4	32,0	33,8	31,0
	2,8	0,5	0,1	0,1	1,4	0,3	0,3
Liquides	23,9	29,0	28,0	27,1	30,7	32,6	30,7
	1,0	0,8	1,4	2,5	0,4	1,3	2,2
Moyenne	23,1	29,6	27,4	26,8	30,6	33,0	30,8
	/u/						
Fricatives voisées	20,3	28,9	27,0	23,8	30,2	33,2	30,5
	2,9	0,8	1,7	1,5	1,7	0,8	0,3
Fricatives non-voisées	19,9	29,5	26,6		30,3	32,7	30,5
	8,0	1,0	0,2		1,0	0,7	0,9
Occlusives voisées	21,2	29,2	27,5	23,9	29,1	32,9	30,8
	2,1	1,8	1,2	1,8	0,7	0,5	0,3
Occlusives non-voisées	20,2	29,3	23,7		29,9	32,4	29,7
	3,4	1,5	1,4		0,8	1,7	0,9
Nasales	20,9	29,3	28,4	28,2	30,7	33,0	30,6
	4,7	2,0	2,1	1,9	0,8	0,9	0,0
Liquides	24,6	29,8	28,6	27,9	31,2	33,4	30,8
	1,4	1,3	0,3	1,3	1,2	1,4	0,4
Moyenne	21,2	29,3	27,0	26,0	30,2	32,9	30,5

<b>Locuteur 2</b>	<b>/a/</b>						
	<b>FO<sub>V1</sub>-init</b>	<b>FO<sub>V1</sub>-max</b>	<b>FO<sub>V1</sub>-finale</b>	<b>FO<sub>c1</sub>-min</b>	<b>FO<sub>V2</sub>-init</b>	<b>FO<sub>V2</sub>-max</b>	<b>FO<sub>V2</sub>-finale</b>
<b>Fricatives voisées</b>	<b>21,8</b> 1,1	<b>27,1</b> 1,0	<b>26,2</b> 0,6	<b>24,4</b> 1,2	<b>26,6</b> 1,1	<b>28,0</b> 1,2	<b>23,5</b> 2,9
<b>Fricatives non-voisées</b>	<b>21,8</b> 1,4	<b>26,3</b> 2,0	<b>25,6</b> 2,1		<b>27,0</b> 0,9	<b>27,4</b> 0,6	<b>25,3</b> 2,1
<b>Occlusives voisées</b>	<b>22,8</b> 0,7	<b>27,6</b> 1,0	<b>25,9</b> 1,4	<b>24,0</b> 1,1	<b>25,3</b> 1,1	<b>27,2</b> 1,9	<b>23,7</b> 3,5
<b>Occlusives non-voisées</b>	<b>22,3</b> 1,7	<b>26,5</b> 1,5	<b>23,9</b> 0,7		<b>26,8</b> 0,9	<b>26,8</b> 0,9	<b>22,7</b> 2,0
<b>Nasales</b>	<b>22,8</b> 2,9	<b>28,0</b> 1,2	<b>27,8</b> 1,0	<b>27,3</b> 1,0	<b>27,9</b> 0,8	<b>28,6</b> 1,3	<b>25,3</b> 6,3
<b>Liquides</b>	<b>22,4</b> 1,9	<b>27,6</b> 3,5	<b>27,2</b> 3,6	<b>26,3</b> 4,5	<b>27,8</b> 2,4	<b>28,2</b> 2,2	<b>26,4</b> 3,6
<b>Moyenne</b>	<b>22,3</b>	<b>27,2</b>	<b>26,1</b>	<b>25,5</b>	<b>26,9</b>	<b>27,7</b>	<b>24,5</b>
	<b>/i/</b>						
<b>Fricatives voisées</b>	<b>25,8</b> 2,2	<b>29,3</b> 1,3	<b>28,4</b> 1,8	<b>27,8</b> 1,9	<b>28,9</b> 1,8	<b>29,8</b> 1,2	<b>27,3</b> 3,9
<b>Fricatives non-voisées</b>	<b>26,5</b> 0,5	<b>29,3</b> 0,8	<b>28,0</b> 0,5		<b>30,2</b> 0,9	<b>30,4</b> 1,2	<b>28,4</b> 0,6
<b>Occlusives voisées</b>	<b>29,6</b> 1,5	<b>30,2</b> 0,9	<b>29,5</b> 0,5	<b>29,2</b> 0,2	<b>29,9</b> 0,5	<b>30,3</b> 0,7	<b>28,8</b> 1,0
<b>Occlusives non-voisées</b>	<b>27,8</b> 1,4	<b>29,4</b> 0,5	<b>28,1</b> 0,7		<b>29,6</b> 1,4	<b>30,1</b> 0,7	<b>28,3</b> 1,0
<b>Nasales</b>	<b>27,4</b> 2,2	<b>30,5</b> 0,7	<b>30,5</b> 0,7	<b>30,5</b> 0,7	<b>31,0</b> 0,5	<b>31,5</b> 0,4	<b>30,8</b> 1,0
<b>Liquides</b>	<b>27,8</b> 3,6	<b>29,8</b> 0,9	<b>29,1</b> 0,4	<b>28,6</b> 1,2	<b>29,1</b> 1,3	<b>29,7</b> 0,4	<b>28,8</b> 0,3
<b>Moyenne</b>	<b>27,5</b>	<b>29,7</b>	<b>28,9</b>	<b>29,0</b>	<b>29,8</b>	<b>30,3</b>	<b>28,7</b>
	<b>/u/</b>						
<b>Fricatives voisées</b>	<b>25,8</b> 1,8	<b>29,7</b> 1,2	<b>27,9</b> 1,5	<b>26,0</b> 2,7	<b>29,0</b> 1,9	<b>30,0</b> 0,9	<b>28,4</b> 0,3
<b>Fricatives non-voisées</b>	<b>25,7</b> 1,1	<b>29,9</b> 0,2	<b>27,7</b> 0,2		<b>30,1</b> 0,2	<b>30,5</b> 0,5	<b>28,1</b> 0,5
<b>Occlusives voisées</b>	<b>26,3</b> 1,7	<b>30,6</b> 0,5	<b>29,9</b> 0,8	<b>29,7</b> 1,1	<b>30,6</b> 0,4	<b>30,9</b> 0,1	<b>28,0</b> 0,7
<b>Occlusives non-voisées</b>	<b>28,1</b> 0,5	<b>30,5</b> 0,6	<b>28,7</b> 0,7		<b>30,7</b> 0,8	<b>30,7</b> 0,8	<b>29,1</b> 0,4
<b>Nasales</b>	<b>26,4</b> 0,3	<b>30,7</b> 0,1	<b>30,4</b> 0,4	<b>30,3</b> 0,5	<b>31,1</b> 1,2	<b>31,4</b> 0,8	<b>29,6</b> 2,0
<b>Liquides</b>	<b>25,6</b> 1,0	<b>30,0</b> 0,6	<b>29,3</b> 1,1	<b>28,8</b> 0,7	<b>30,3</b> 0,9	<b>30,8</b> 0,3	<b>28,5</b> 1,7
<b>Moyenne</b>	<b>26,3</b>	<b>30,2</b>	<b>29,0</b>	<b>28,7</b>	<b>30,3</b>	<b>30,7</b>	<b>28,6</b>

<b>Locuteur 3</b>	<b>/a/</b>						
	<b>FO<sub>V1</sub>-init</b>	<b>FO<sub>V1</sub>-max</b>	<b>FO<sub>V1</sub>-finale</b>	<b>FO<sub>c1</sub>-min</b>	<b>FO<sub>V2</sub>-init</b>	<b>FO<sub>V2</sub>-max</b>	<b>FO<sub>V2</sub>-finale</b>
<b>Fricatives voisées</b>	<b>22,6</b> 0,4	<b>27,1</b> 0,6	<b>25,3</b> 0,4	<b>22,4</b> 0,3	<b>25,6</b> 1,1	<b>31,7</b> 0,4	<b>29,3</b> 0,2
<b>Fricatives non-voisées</b>	<b>20,9</b> 0,9	<b>25,8</b> 0,5	<b>24,2</b> 0,8		<b>28,9</b> 0,2	<b>31,6</b> 0,6	<b>29,5</b> 0,7
<b>Occlusives voisées</b>	<b>21,6</b> 1,6	<b>27,2</b> 0,4	<b>25,7</b> 0,3	<b>23,3</b> 0,5	<b>24,8</b> 2,4	<b>31,9</b> 0,4	<b>29,0</b> 3,1
<b>Occlusives non-voisées</b>	<b>20,9</b> 2,5	<b>25,9</b> 3,0	<b>22,9</b> 2,1		<b>28,6</b> 1,5	<b>32,0</b> 0,8	<b>28,2</b> 2,1
<b>Nasales</b>	<b>21,6</b> 1,0	<b>27,1</b> 0,3	<b>26,9</b> 0,3	<b>26,6</b> 0,4	<b>27,5</b> 1,1	<b>31,4</b> 0,6	<b>28,8</b> 1,2
<b>Liquides</b>	<b>22,3</b> 2,3	<b>28,4</b> 0,7	<b>27,0</b> 1,3	<b>26,6</b> 1,8	<b>28,9</b> 0,3	<b>32,8</b> 0,1	<b>30,8</b> 1,0
<b>Moyenne</b>	<b>21,7</b>	<b>26,9</b>	<b>25,3</b>	<b>24,7</b>	<b>27,4</b>	<b>31,9</b>	<b>29,3</b>
	<b>/i/</b>						
<b>Fricatives voisées</b>	<b>25,3</b> 0,3	<b>29,1</b> 0,4	<b>26,7</b> 0,0	<b>23,1</b> 1,4	<b>27,6</b> 1,0	<b>31,4</b> 0,9	<b>29,7</b> 1,9
<b>Fricatives non-voisées</b>	<b>24,9</b> 0,1	<b>28,5</b> 1,1	<b>25,5</b> 0,6		<b>30,2</b> 0,6	<b>32,2</b> 0,5	<b>30,5</b> 0,3
<b>Occlusives voisées</b>	<b>25,1</b> 1,6	<b>29,8</b> 0,5	<b>27,8</b> 1,0	<b>25,2</b> 2,2	<b>28,8</b> 1,2	<b>31,8</b> 0,7	<b>28,2</b> 1,0
<b>Occlusives non-voisées</b>	<b>25,0</b> 1,4	<b>27,6</b> 0,4	<b>23,6</b> 0,4		<b>29,8</b> 0,8	<b>31,5</b> 0,6	<b>28,9</b> 0,4
<b>Nasales</b>	<b>24,9</b> 0,4	<b>28,4</b> 0,5	<b>27,4</b> 1,2	<b>26,8</b> 0,9	<b>29,5</b> 0,3	<b>32,1</b> 0,8	<b>29,5</b> 1,6
<b>Liquides</b>	<b>25,6</b> 1,0	<b>27,4</b> 1,9	<b>26,7</b> 2,0	<b>24,3</b> 5,1	<b>26,5</b> 2,4	<b>31,6</b> 0,2	<b>28,9</b> 1,2
<b>Moyenne</b>	<b>25,1</b>	<b>28,5</b>	<b>26,3</b>	<b>24,9</b>	<b>28,7</b>	<b>31,8</b>	<b>29,3</b>
	<b>/u/</b>						
<b>Fricatives voisées</b>	<b>24,4</b> 2,0	<b>28,6</b> 1,4	<b>26,9</b> 2,2	<b>25,0</b> 1,7	<b>28,8</b> 0,7	<b>32,5</b> 0,4	<b>29,3</b> 0,8
<b>Fricatives non-voisées</b>	<b>26,4</b> 1,6	<b>30,0</b> 0,3	<b>25,6</b> 1,0		<b>30,5</b> 0,5	<b>32,4</b> 0,0	<b>29,8</b> 0,1
<b>Occlusives voisées</b>	<b>25,0</b> 0,3	<b>30,2</b> 0,3	<b>28,1</b> 0,3	<b>25,9</b> 1,0	<b>28,5</b> 0,5	<b>32,8</b> 0,3	<b>28,3</b> 0,3
<b>Occlusives non-voisées</b>	<b>26,1</b> 1,2	<b>29,8</b> 1,2	<b>24,9</b> 1,7		<b>30,9</b> 1,0	<b>32,6</b> 0,8	<b>30,3</b> 1,2
<b>Nasales</b>	<b>25,1</b> 0,5	<b>29,8</b> 0,9	<b>29,3</b> 1,1	<b>28,8</b> 0,7	<b>29,8</b> 0,7	<b>32,8</b> 0,2	<b>28,2</b> 1,9
<b>Liquides</b>	<b>26,5</b> 0,2	<b>29,5</b> 0,2	<b>28,2</b> 2,0	<b>27,4</b> 2,4	<b>30,6</b> 0,7	<b>33,7</b> 1,3	<b>30,9</b> 1,2
<b>Moyenne</b>	<b>25,6</b>	<b>29,7</b>	<b>27,2</b>	<b>26,8</b>	<b>29,9</b>	<b>32,8</b>	<b>29,5</b>

## E.4 Les logatomes CVCV

<b>Locuteur 1</b>	<b>/a/</b>							
	<b>F0<sub>C1</sub>-init</b>	<b>F0<sub>V1</sub>-init</b>	<b>F0<sub>V1</sub>-max</b>	<b>F0<sub>V1</sub>-finale</b>	<b>F0<sub>C2</sub>-min</b>	<b>F0<sub>V2</sub>-init</b>	<b>F0<sub>V2</sub>-max</b>	<b>F0<sub>V2</sub>-finale</b>
<b>Fricatives voisées</b>	15,5 1,2	22,5 2,4	28,0 1,1	26,0 1,4	24,3 1,4	28,0 1,1	32,5 0,5	30,8 0,9
<b>Fricatives non-voisées</b>		26,0 0,4	27,8 0,5	24,6 1,3		29,1 0,4	32,0 0,5	31,2 1,1
<b>Occlusives voisées</b>	15,5 2,4	22,8 2,5	26,0 0,6	24,4 1,7	22,2 0,6	26,9 0,1	32,1 0,6	30,4 0,5
<b>Occlusives non-voisées</b>		24,4 2,2	26,1 1,8	23,3 2,2		27,8 1,1	31,2 0,2	29,0 0,8
<b>Nasales</b>	18,6 0,1	25,5 1,0	28,2 0,7	28,0 0,5	28,0 0,4	29,8 1,3	33,1 0,9	30,9 1,3
<b>Liquides</b>	15,3 1,1	22,4 0,5	26,3 0,7	26,0 0,3	25,7 0,6	27,2 0,5	32,5 1,3	30,8 2,3
<b>Moyenne</b>	<b>16,2</b>	<b>23,9</b>	<b>27,1</b>	<b>25,4</b>	<b>25,1</b>	<b>28,1</b>	<b>32,2</b>	<b>30,5</b>
	<b>/i/</b>							
<b>Fricatives voisées</b>	16,7 2,2	24,2 0,6	29,4 0,7	26,0 0,4	24,3 1,3	29,0 0,8	33,8 0,6	31,1 0,2
<b>Fricatives non-voisées</b>		27,6 1,2	31,3 0,7	28,6 2,3		31,1 1,3	34,4 0,7	31,2 0,4
<b>Occlusives voisées</b>	18,1 1,4	24,5 0,7	29,7 1,0	27,1 1,4	25,3 0,5	27,6 1,8	33,5 0,7	30,2 0,3
<b>Occlusives non-voisées</b>		26,3 0,7	29,2 1,2	25,5 2,7		29,8 1,3	33,0 0,1	30,8 1,0
<b>Nasales</b>	19,1 1,0	26,1 1,2	29,5 1,0	29,3 1,2	28,9 1,3	30,7 1,0	33,7 0,3	31,0 0,4
<b>Liquides</b>	18,8 3,2	24,8 1,9	29,8 0,6	27,2 3,0	26,9 2,5	28,6 1,7	33,5 0,1	30,0 0,8
<b>Moyenne</b>	<b>18,2</b>	<b>25,6</b>	<b>29,8</b>	<b>27,3</b>	<b>26,3</b>	<b>29,5</b>	<b>33,7</b>	<b>30,7</b>
	<b>/u/</b>							
<b>Fricatives voisées</b>	17,0 2,1	26,0 1,9	29,5 1,0	26,6 1,3	23,9 2,6	28,5 1,1	33,0 0,5	30,6 1,7
<b>Fricatives non-voisées</b>		28,0 1,3	30,6 0,8	26,6 1,6		30,3 0,9	34,2 0,2	31,8 0,4
<b>Occlusives voisées</b>	16,9 2,9	25,0 1,8	29,5 1,9	27,6 1,7	24,5 1,1	28,9 1,3	33,6 0,6	31,9 1,3
<b>Occlusives non-voisées</b>		28,3 1,4	31,0 1,4	26,6 1,9		30,4 0,8	33,2 0,6	31,3 1,0
<b>Nasales</b>	20,6 1,4	26,9 0,4	29,2 0,4	28,6 0,8	27,9 1,7	29,5 0,1	33,5 0,7	30,7 0,8
<b>Liquides</b>	20,7 2,4	27,7 0,3	30,1 0,8	28,5 0,8	26,8 1,8	28,8 0,0	33,6 0,0	30,9 0,5
<b>Moyenne</b>	<b>18,8</b>	<b>27,0</b>	<b>30,0</b>	<b>27,4</b>	<b>25,8</b>	<b>29,4</b>	<b>33,5</b>	<b>31,2</b>

<b>Locuteur 1</b>	<b>/a/</b>							
	<b>F0<sub>C1</sub>-init</b>	<b>F0<sub>V1</sub>-init</b>	<b>F0<sub>V1</sub>-max</b>	<b>F0<sub>V1</sub>-finale</b>	<b>F0<sub>C2</sub>-min</b>	<b>F0<sub>V2</sub>-init</b>	<b>F0<sub>V2</sub>-max</b>	<b>F0<sub>V2</sub>-finale</b>
<b>Fricatives voisées</b>	<b>16,7</b> <i>0,5</i>	<b>20,5</b> <i>0,4</i>	<b>27,4</b> <i>0,2</i>	<b>26,8</b> <i>0,0</i>	<b>25,3</b> <i>0,3</i>	<b>26,3</b> <i>0,7</i>	<b>28,4</b> <i>2,2</i>	<b>27,7</b> <i>2,9</i>
<b>Fricatives non-voisées</b>		<b>25,8</b> <i>1,2</i>	<b>28,1</b> <i>1,3</i>	<b>26,8</b> <i>1,6</i>		<b>28,0</b> <i>0,9</i>	<b>28,2</b> <i>1,0</i>	<b>26,7</b> <i>2,7</i>
<b>Occlusives voisées</b>	<b>16,6</b> <i>0,3</i>	<b>20,4</b> <i>1,0</i>	<b>26,2</b> <i>0,9</i>	<b>25,3</b> <i>0,3</i>	<b>23,8</b> <i>0,9</i>	<b>24,6</b> <i>1,8</i>	<b>27,5</b> <i>1,5</i>	<b>26,1</b> <i>2,5</i>
<b>Occlusives non-voisées</b>		<b>24,2</b> <i>2,3</i>	<b>27,0</b> <i>2,4</i>	<b>25,5</b> <i>2,4</i>		<b>27,4</b> <i>1,5</i>	<b>28,2</b> <i>1,9</i>	<b>26,8</b> <i>2,6</i>
<b>Nasales</b>	<b>18,0</b> <i>0,4</i>	<b>22,7</b> <i>2,3</i>	<b>27,5</b> <i>1,0</i>	<b>27,5</b> <i>1,0</i>	<b>27,0</b> <i>0,6</i>	<b>27,5</b> <i>0,5</i>	<b>28,0</b> <i>1,3</i>	<b>27,0</b> <i>1,9</i>
<b>Liquides</b>	<b>16,2</b> <i>1,1</i>	<b>22,3</b> <i>1,2</i>	<b>26,9</b> <i>0,5</i>	<b>26,9</b> <i>0,4</i>	<b>26,6</b> <i>0,0</i>	<b>27,2</b> <i>0,7</i>	<b>28,8</b> <i>2,4</i>	<b>28,8</b> <i>2,4</i>
<b>Moyenne</b>	<b>16,9</b>	<b>22,6</b>	<b>27,2</b>	<b>26,5</b>	<b>25,7</b>	<b>26,9</b>	<b>28,2</b>	<b>27,2</b>
	<b>/i/</b>							
<b>Fricatives voisées</b>	<b>17,9</b> <i>0,7</i>	<b>26,9</b> <i>1,2</i>	<b>30,2</b> <i>0,3</i>	<b>28,7</b> <i>0,7</i>	<b>28,1</b> <i>0,9</i>	<b>29,5</b> <i>0,5</i>	<b>30,3</b> <i>0,3</i>	<b>28,8</b> <i>0,5</i>
<b>Fricatives non-voisées</b>		<b>27,6</b> <i>0,6</i>	<b>29,6</b> <i>0,8</i>	<b>27,3</b> <i>0,4</i>		<b>30,4</b> <i>0,8</i>	<b>30,4</b> <i>0,8</i>	<b>28,9</b> <i>0,8</i>
<b>Occlusives voisées</b>	<b>17,6</b> <i>0,6</i>	<b>22,1</b> <i>1,3</i>	<b>30,2</b> <i>0,6</i>	<b>28,7</b> <i>1,0</i>	<b>28,6</b> <i>1,0</i>	<b>29,4</b> <i>1,2</i>	<b>30,5</b> <i>0,3</i>	<b>29,2</b> <i>0,9</i>
<b>Occlusives non-voisées</b>		<b>27,2</b> <i>0,9</i>	<b>29,7</b> <i>0,6</i>	<b>27,0</b> <i>1,1</i>		<b>29,9</b> <i>0,2</i>	<b>30,1</b> <i>0,3</i>	<b>28,5</b> <i>0,0</i>
<b>Nasales</b>	<b>19,0</b> <i>2,6</i>	<b>26,7</b> <i>2,8</i>	<b>30,1</b> <i>0,9</i>	<b>30,0</b> <i>0,7</i>	<b>29,8</b> <i>0,4</i>	<b>30,1</b> <i>0,3</i>	<b>30,2</b> <i>0,2</i>	<b>29,5</b> <i>0,5</i>
<b>Liquides</b>	<b>19,1</b> <i>1,3</i>	<b>24,0</b> <i>0,9</i>	<b>30,0</b> <i>0,8</i>	<b>29,4</b> <i>0,8</i>	<b>28,1</b> <i>0,1</i>	<b>28,4</b> <i>0,5</i>	<b>30,1</b> <i>0,4</i>	<b>29,3</b> <i>0,1</i>
<b>Moyenne</b>	<b>18,4</b>	<b>25,7</b>	<b>30,0</b>	<b>28,5</b>	<b>28,6</b>	<b>29,6</b>	<b>30,3</b>	<b>29,0</b>
	<b>/u/</b>							
<b>Fricatives voisées</b>	<b>16,9</b> <i>0,8</i>	<b>25,6</b> <i>1,9</i>	<b>29,8</b> <i>1,2</i>	<b>28,6</b> <i>0,9</i>	<b>27,4</b> <i>1,2</i>	<b>28,9</b> <i>0,9</i>	<b>29,9</b> <i>0,5</i>	<b>29,1</b> <i>1,2</i>
<b>Fricatives non-voisées</b>		<b>29,9</b> <i>1,0</i>	<b>31,0</b> <i>1,0</i>	<b>27,6</b> <i>1,5</i>		<b>30,0</b> <i>0,8</i>	<b>30,7</b> <i>0,2</i>	<b>29,5</b> <i>0,9</i>
<b>Occlusives voisées</b>	<b>17,9</b> <i>0,2</i>	<b>25,2</b> <i>4,4</i>	<b>31,2</b> <i>0,7</i>	<b>30,0</b> <i>0,6</i>	<b>29,3</b> <i>1,0</i>	<b>30,1</b> <i>0,7</i>	<b>31,1</b> <i>0,7</i>	<b>30,1</b> <i>0,8</i>
<b>Occlusives non-voisées</b>		<b>29,0</b> <i>1,1</i>	<b>31,0</b> <i>0,3</i>	<b>27,4</b> <i>1,8</i>		<b>30,7</b> <i>0,7</i>	<b>31,0</b> <i>0,4</i>	<b>28,8</b> <i>0,4</i>
<b>Nasales</b>	<b>20,2</b> <i>1,9</i>	<b>27,5</b> <i>2,6</i>	<b>30,6</b> <i>1,0</i>	<b>30,2</b> <i>0,6</i>	<b>29,8</b> <i>0,6</i>	<b>30,3</b> <i>0,4</i>	<b>30,6</b> <i>0,5</i>	<b>28,7</b> <i>1,1</i>
<b>Liquides</b>	<b>18,2</b> <i>0,0</i>	<b>24,6</b> <i>2,4</i>	<b>30,0</b> <i>0,0</i>	<b>29,0</b> <i>0,0</i>	<b>28,4</b> <i>0,5</i>	<b>29,8</b> <i>0,1</i>	<b>30,4</b> <i>0,1</i>	<b>29,0</b> <i>1,1</i>
<b>Moyenne</b>	<b>18,3</b>	<b>27,0</b>	<b>30,6</b>	<b>28,8</b>	<b>28,7</b>	<b>30,0</b>	<b>30,6</b>	<b>29,2</b>

<b>Locuteur 1</b>	<b>/a/</b>							
	<b>F0<sub>C1</sub>-init</b>	<b>F0<sub>V1</sub>-init</b>	<b>F0<sub>V1</sub>-max</b>	<b>F0<sub>V1</sub>-finale</b>	<b>F0<sub>C2</sub>-min</b>	<b>F0<sub>V2</sub>-init</b>	<b>F0<sub>V2</sub>-max</b>	<b>F0<sub>V2</sub>-finale</b>
<b>Fricatives voisées</b>	<b>19,1</b> <i>1,6</i>	<b>25,3</b> <i>0,6</i>	<b>31,1</b> <i>0,3</i>	<b>28,7</b> <i>1,1</i>	<b>26,8</b> <i>0,9</i>	<b>29,6</b> <i>0,1</i>	<b>33,2</b> <i>1,1</i>	<b>27,4</b> <i>2,7</i>
<b>Fricatives non-voisées</b>		<b>28,5</b> <i>1,2</i>	<b>32,1</b> <i>0,5</i>	<b>27,1</b> <i>0,7</i>		<b>31,5</b> <i>0,6</i>	<b>32,9</b> <i>0,5</i>	<b>27,3</b> <i>2,6</i>
<b>Occlusives voisées</b>	<b>20,8</b> <i>3,5</i>	<b>25,2</b> <i>2,3</i>	<b>31,3</b> <i>0,7</i>	<b>29,1</b> <i>0,8</i>	<b>27,4</b> <i>0,1</i>	<b>29,8</b> <i>0,5</i>	<b>33,2</b> <i>0,2</i>	<b>28,7</b> <i>3,1</i>
<b>Occlusives non-voisées</b>		<b>26,5</b> <i>1,1</i>	<b>30,0</b> <i>1,6</i>	<b>26,5</b> <i>1,1</i>		<b>31,3</b> <i>0,5</i>	<b>33,6</b> <i>0,9</i>	<b>28,9</b> <i>3,6</i>
<b>Nasales</b>	<b>20,3</b> <i>2,5</i>	<b>26,4</b> <i>2,6</i>	<b>30,9</b> <i>1,1</i>	<b>30,6</b> <i>0,7</i>	<b>29,9</b> <i>1,0</i>	<b>30,7</b> <i>1,5</i>	<b>33,0</b> <i>0,4</i>	<b>28,6</b> <i>1,6</i>
<b>Liquides</b>	<b>20,1</b> <i>1,9</i>	<b>26,0</b> <i>0,5</i>	<b>31,0</b> <i>0,7</i>	<b>29,7</b> <i>2,6</i>	<b>29,0</b> <i>3,5</i>	<b>30,7</b> <i>1,3</i>	<b>33,5</b> <i>1,4</i>	<b>30,6</b> <i>0,2</i>
<b>Moyenne</b>	<b>20,1</b>	<b>26,3</b>	<b>31,1</b>	<b>28,6</b>	<b>28,3</b>	<b>30,6</b>	<b>33,3</b>	<b>28,6</b>
	<b>/i/</b>							
<b>Fricatives voisées</b>	<b>19,7</b> <i>0,4</i>	<b>26,7</b> <i>0,4</i>	<b>31,8</b> <i>0,5</i>	<b>28,4</b> <i>1,6</i>	<b>27,2</b> <i>1,8</i>	<b>29,9</b> <i>1,2</i>	<b>34,1</b> <i>0,8</i>	<b>29,7</b> <i>2,3</i>
<b>Fricatives non-voisées</b>		<b>29,6</b> <i>0,1</i>	<b>32,3</b> <i>0,6</i>	<b>27,8</b> <i>0,7</i>		<b>31,5</b> <i>0,5</i>	<b>33,8</b> <i>0,5</i>	<b>29,3</b> <i>1,5</i>
<b>Occlusives voisées</b>	<b>20,1</b> <i>0,6</i>	<b>24,8</b> <i>1,0</i>	<b>31,8</b> <i>0,6</i>	<b>28,9</b> <i>2,9</i>	<b>28,1</b> <i>3,2</i>	<b>30,4</b> <i>2,4</i>	<b>33,5</b> <i>0,7</i>	<b>28,4</b> <i>2,9</i>
<b>Occlusives non-voisées</b>		<b>28,9</b> <i>0,7</i>	<b>31,7</b> <i>0,3</i>	<b>28,7</b> <i>0,9</i>		<b>31,2</b> <i>0,7</i>	<b>33,8</b> <i>0,5</i>	<b>27,7</b> <i>4,8</i>
<b>Nasales</b>	<b>21,8</b> <i>0,8</i>	<b>28,6</b> <i>1,0</i>	<b>32,9</b> <i>1,1</i>	<b>32,8</b> <i>1,2</i>	<b>32,8</b> <i>1,2</i>	<b>33,3</b> <i>1,2</i>	<b>34,3</b> <i>0,7</i>	<b>31,0</b> <i>0,5</i>
<b>Liquides</b>	<b>21,1</b> <i>3,2</i>	<b>26,5</b> <i>0,1</i>	<b>31,6</b> <i>1,2</i>	<b>30,9</b> <i>0,7</i>	<b>30,3</b> <i>1,5</i>	<b>31,8</b> <i>0,3</i>	<b>34,1</b> <i>0,1</i>	<b>26,3</b> <i>0,5</i>
<b>Moyenne</b>	<b>20,7</b>	<b>27,5</b>	<b>32,0</b>	<b>29,6</b>	<b>29,6</b>	<b>31,4</b>	<b>33,9</b>	<b>28,8</b>
	<b>/u/</b>							
<b>Fricatives voisées</b>	<b>20,8</b> <i>1,0</i>	<b>28,7</b> <i>1,3</i>	<b>32,9</b> <i>1,5</i>	<b>29,5</b> <i>2,4</i>	<b>28,1</b> <i>2,6</i>	<b>31,7</b> <i>1,2</i>	<b>35,2</b> <i>0,5</i>	<b>29,6</b> <i>2,9</i>
<b>Fricatives non-voisées</b>		<b>30,8</b> <i>0,6</i>	<b>33,2</b> <i>0,8</i>	<b>28,2</b> <i>0,4</i>		<b>32,0</b> <i>0,4</i>	<b>34,6</b> <i>0,5</i>	<b>29,1</b> <i>1,6</i>
<b>Occlusives voisées</b>	<b>21,2</b> <i>1,0</i>	<b>24,9</b> <i>1,3</i>	<b>33,0</b> <i>0,4</i>	<b>30,7</b> <i>1,3</i>	<b>29,4</b> <i>2,7</i>	<b>31,3</b> <i>1,7</i>	<b>35,5</b> <i>0,5</i>	<b>29,2</b> <i>4,1</i>
<b>Occlusives non-voisées</b>		<b>30,7</b> <i>1,3</i>	<b>33,5</b> <i>1,7</i>	<b>29,4</b> <i>2,6</i>		<b>32,1</b> <i>0,9</i>	<b>35,3</b> <i>0,6</i>	<b>32,1</b> <i>1,4</i>
<b>Nasales</b>	<b>23,6</b> <i>1,7</i>	<b>30,1</b> <i>2,9</i>	<b>33,3</b> <i>1,5</i>	<b>32,6</b> <i>1,2</i>	<b>32,3</b> <i>1,3</i>	<b>33,1</b> <i>0,9</i>	<b>34,7</b> <i>0,9</i>	<b>30,8</b> <i>1,7</i>
<b>Liquides</b>	<b>22,0</b> <i>2,9</i>	<b>30,3</b> <i>1,2</i>	<b>34,0</b> <i>0,2</i>	<b>31,5</b> <i>1,3</i>	<b>30,8</b> <i>1,7</i>	<b>32,8</b> <i>0,5</i>	<b>35,7</b> <i>1,4</i>	<b>30,7</b> <i>2,0</i>
<b>Moyenne</b>	<b>21,9</b>	<b>29,2</b>	<b>33,3</b>	<b>30,3</b>	<b>30,2</b>	<b>32,2</b>	<b>35,2</b>	<b>30,3</b>



# Annexe F

## Résultats des tests de perception de la distance pour chaque sujet

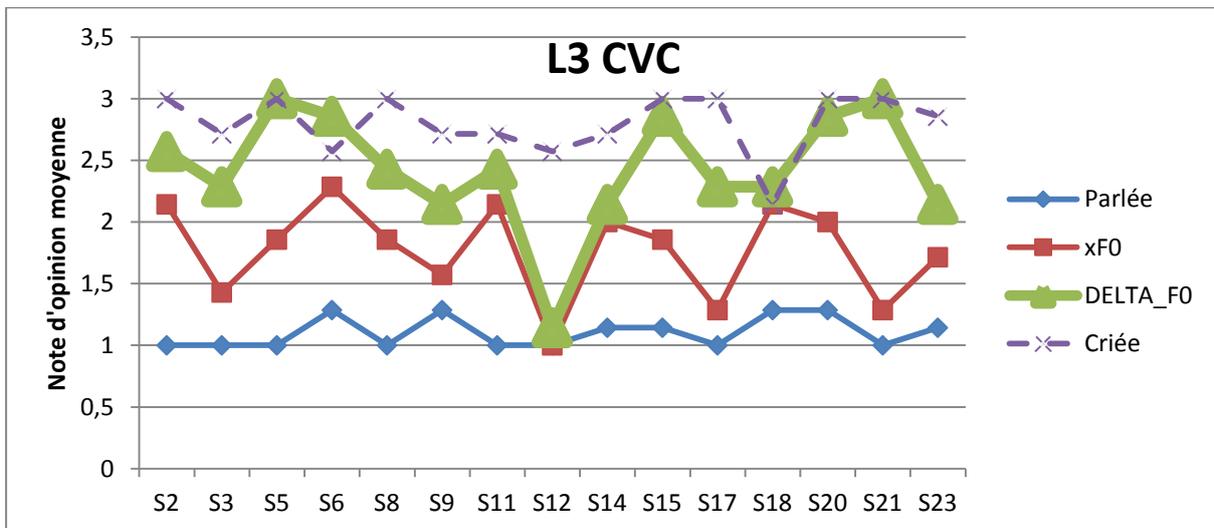
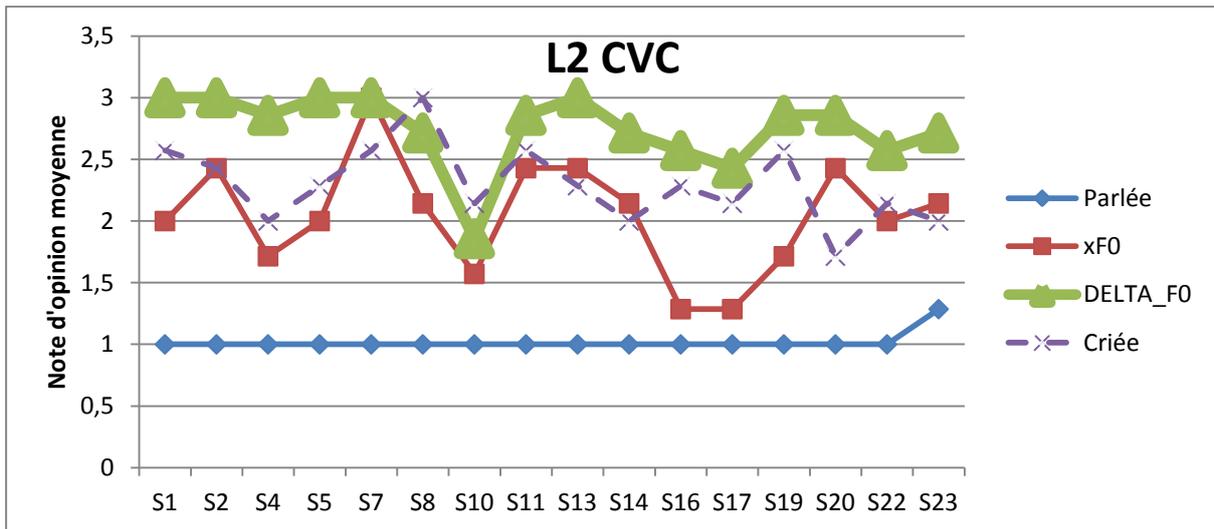
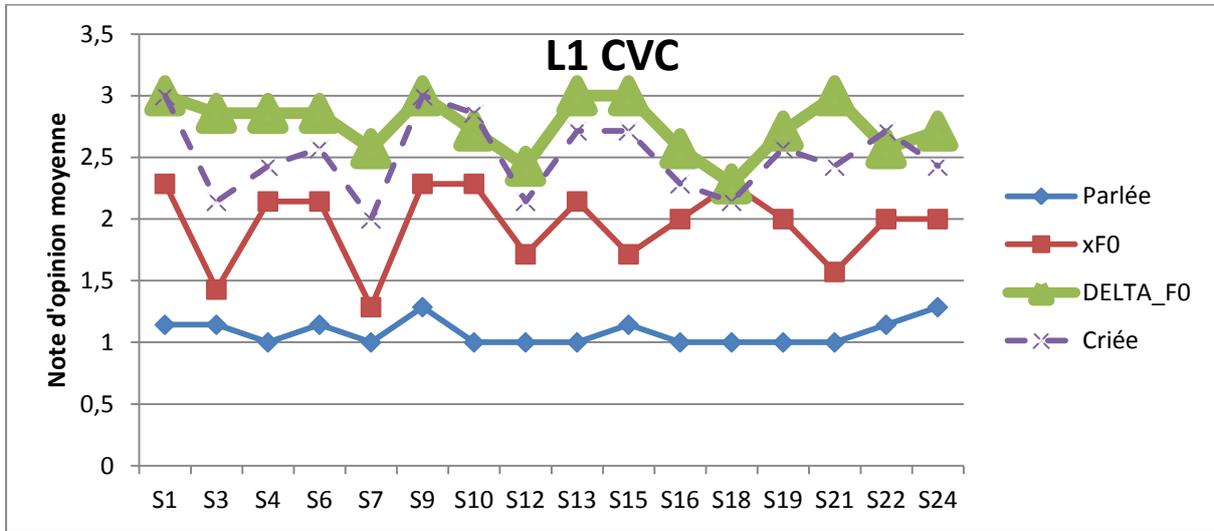
---

Cette annexe présente d'une part l'ordre de passage de chaque série dans les tests de perception de la distance présentés dans le chapitre 11 ainsi que les résultats obtenu en fonction des locuteurs et des sujets.

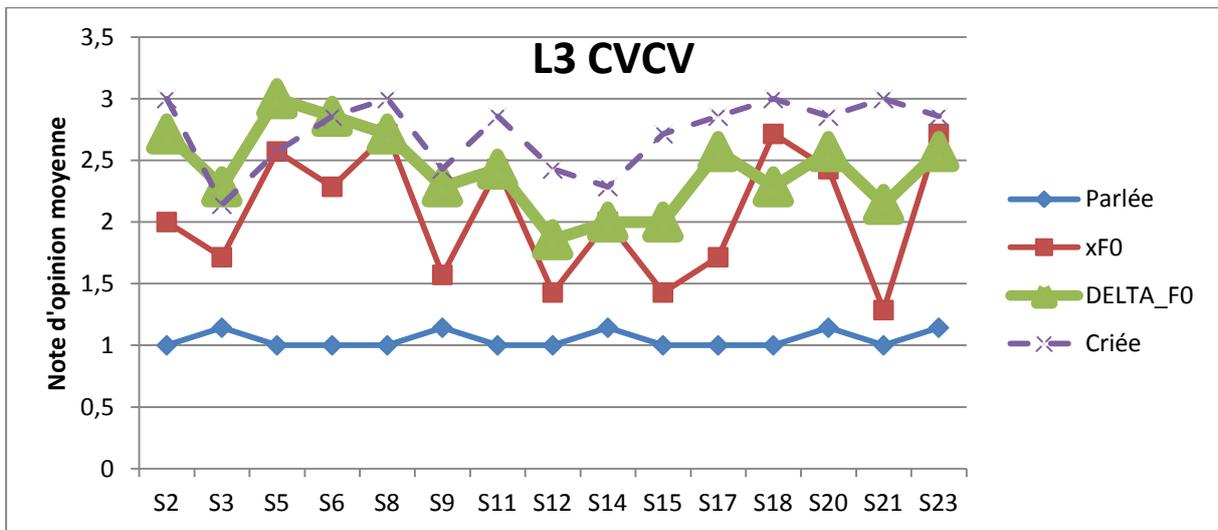
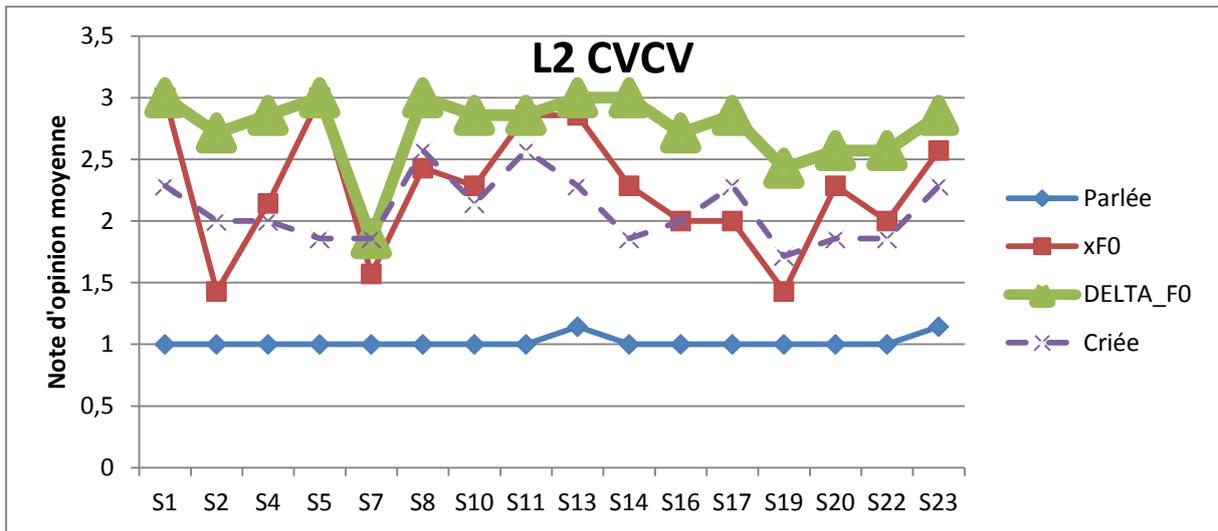
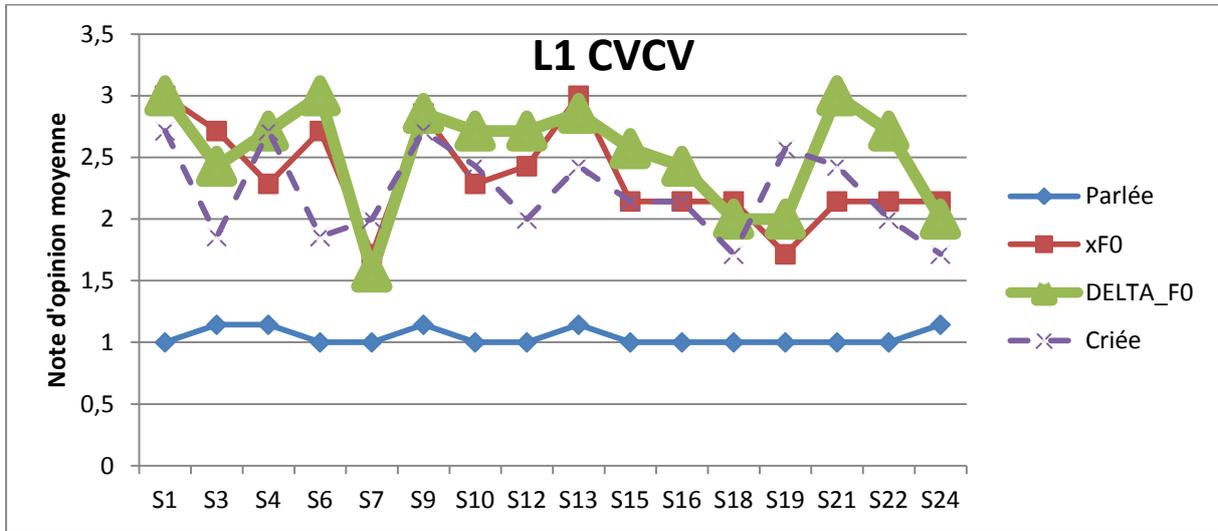
### F.1 Ordre de passage des séries en fonction des sujets

L1		L2		L3		Nom
CVC	CVCV	CVC	CVCV	CVC	CVCV	
1	2	4	3			S1
		2	1	3	4	S2
4	3			1	2	S3
2	1	3	4			S4
		1	2	4	3	S5
3	4			2	1	S6
1	2	4	3			S7
		2	1	3	4	S8
4	3			1	2	S9
2	1	3	4			S10
		1	2	4	3	S11
3	4			2	1	S12
1	2	4	3			S13
		2	1	3	4	S14
4	3			1	2	S15
2	1	3	4			S16
		1	2	4	3	S17
3	4			2	1	S18
1	2	4	3			S19
		2	1	3	4	S20
4	3			1	2	S21
2	1	3	4			S22
		1	2	4	3	S23
3	4			2	1	S24

### F.2 Résultats de chaque locuteur en fonction de la série pour les logatomes CVC



### F.3 Résultats de chaque locuteur en fonction de la série pour les logatomes CVCV



# RÉSUMÉ

Cette thèse porte sur la transformation de la voix d'un locuteur dans l'objectif d'indiquer la distance de celui-ci : une transformation en voix chuchotée pour indiquer une distance proche et une transformation en voix criée pour une distance plutôt éloignée. Nous effectuons dans un premier temps des analyses approfondies pour déterminer les paramètres les plus pertinentes dans une voix chuchotée et surtout dans une voix criée (beaucoup plus difficile). La contribution principale de cette partie est de montrer la pertinence des paramètres prosodiques dans la perception de l'effort vocal dans une voix criée. Nous proposons ensuite des descripteurs permettant de mieux caractériser les contours prosodiques. Pour la transformation proprement dite, nous proposons plusieurs nouvelles règles de transformation qui contrôlent de manière primordiale la qualité des voix transformées. Les résultats ont montré une très bonne qualité des voix chuchotées transformées ainsi que pour des voix criées pour des structures linguistiques relativement simples (CVC, CVCV, etc.).

# ABSTRACT

This thesis focuses on speaker voice transformation in the aim to indicate the distance of it: a spoken-to-whispered voice transformation to indicate a close distance and a spoken-to-shouted voice transformation for a rather far distance. We perform at first, in-depth analysis to determine most relevant features in whispered voices and especially in shouted voices (much harder). The main contribution of this part is to show the relevance of prosodic parameters in the perception of vocal effort in a shouted voice. Then, we propose some descriptors to better characterize the prosodic contours. For the actual transformation, we propose several new transformation rules which importantly control the quality of transformed voice. The results showed a very good quality of transformed whispered voices and transformed shouted voices for relatively simple linguistic structures (CVC, CVCV, etc.).