UNIVERSITÉ DU QUÉBEC À MONTRÉAL

DE LA PRISE EN COMPTE DE LA PROSODIE DU FRANÇAIS QUÉBÉCOIS : EFFETS SUR L'INTELLIGIBILITÉ ET LA QUALITÉ D'UN SYSTÈME DE SYNTHÈSE DE LA PAROLE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAITRISE EN LINGUISTIQUE

PAR

AMÉLIE PRÉMONT

JUIN 2015

Pour Mathieu.

REMERCIEMENTS

L'idée de cette étude nous est venue en discutant avec une collègue diplômée du laboratoire de phonétique employée dans l'industrie de la synthèse de la parole. Elle déplorait l'absence de travaux sur la prosodie du français québécois appliqués à ce domaine. Cette interrogation nous a amenée à lire sur le sujet et à découvrir les défis passionnants entourant la génération artificielle d'une voix. Ce mémoire est donc le fruit de cette recherche.

Nous tenons à remercier, d'abord, notre directrice de recherche Lucie Ménard. Elle est une directrice exceptionnelle, dotée à la fois de qualités humaines et intellectuelles remarquables. Merci de m'avoir permis de tester mes limites et de m'avoir poussée à entreprendre ce projet. Merci d'avoir été tour à tour compréhensive et exigeante, selon mes besoins. Merci de m'avoir accordé l'immense privilège de travailler au laboratoire – j'y ai plus appris que dans n'importe quel séminaire. Merci d'avoir créé une si belle ambiance de partage et de collaboration entre tes étudiants.

Nous souhaitons aussi remercier les professeurs du département de linguistique de l'UQAM qui ont accepté d'être membres du jury, Denis Foucambert et Elizabeth Smith.

Je tiens aussi à remercier mon mari, Mathieu. Il n'y a pas assez d'une page de remerciements pour te témoigner toute ma reconnaissance. Merci pour le support moral – de me prendre avec mes hauts et mes bas, dans les moments où rien ne

semble avancer comme dans ceux où la vaisselle s'empile parce que je travaille trop. Merci de m'avoir appris à programmer. Sans cela, ce mémoire n'aurait pas été possible. Merci aussi de m'avoir aidé dans les aspects les plus techniques du mémoire – dans les bogues informatiques, les maths et la multitude de données.

Un énorme merci à mes parents qui m'ont accompagnée quotidiennement dans les derniers moments de rédaction. Ma mère, Annie, merci pour la révision sans pitié et ton investissement en temps. Mon père, Jean-René, merci de nous avoir épaulées, maman et moi et de nous avoir ramenées les pieds sur terre quand il le fallait.

Je veux aussi remercier ma famille, notamment ma sœur Valérie – œil de lynx et la tribu Johnson (Louise, Robert, MC, Cédric, Sophie, Fred, Carou, Matt, Martin et Alexandra) toujours prête à participer aux études du laboratoire.

Merci aussi aux filles du laboratoire (Christine, Paméla, Marilène, Annie, Caroline, Dominique, Lucille, Marie, Mélanie, Marilyn et Suzie) pour avoir créé un environnement de travail si stimulant. J'ai énormément appris à vos côtés. Merci de m'avoir choisie pour régler vos petits problèmes informatiques – ils m'ont permis de trouver des solutions à des problèmes qu'aucun cours n'aurait abordés.

Finalement, un énorme merci à la trentaine de personnes qui ont accepté de participer à l'un ou l'autre de mes tests. Sans leur investissement en temps, cette recherche n'aurait pas vu le jour.

TABLE DES MATIÈRES

LISTE DES FIGURES ix
LISTE DES TABLEAUX xi
LISTE DES ABRÉVIATIONS SIGLES ET ACRONYMES xiii
RÉSUMÉ xv
INTRODUCTION 1
CHAPITRE I
PROBLÉMATIQUE ET CADRE THÉORIQUE 3
1.1 Synthèse de la parole 4
1.1.1 Les différents types de systèmes 5
1.1.2 Les modules des systèmes de synthèse 7
1.1.3 L'évaluation de la qualité des systèmes 8
1.2 Phonétique acoustique 13
1.3 Prosodie 16
1.3.1 Prosodie et paramètres acoustiques 17
1.3.2 Rythme et accentuation 18
1.3.3 Syllabe 22
1.3.4 Intonation 23
1.4 Prosodie en synthèse 25
1.5 Variation dialectale 26
1.6 Conclusion 27

CHAPITRE II
REVUE DE LA LITTÉRATURE 29
2.1 Prosodie et synthèse 29
2.1.1 Prosodie et intelligibilité 29
2.1.2 Prosodie et qualité 31
2.2 Synthèse et dialecte 32
2.2.1 De l'importance du FQ 34
2.2.2 Revue des voix de synthèse disponibles au Québec 35
2.3 Prosodie et français québécois 35
2.3.1 Revue des études réalisées avant 1995 36
2.3.2 Études comparatives entre le FF et le FQ 40
2.3.3 Études sur la prosodie du FQ 42
2.4 Question de recherche et objectifs 43
2.4.1 Premier objectif : Identifier certaines caractéristiques prosodiques principales du français québécois, dans sa perception et sa production 44
2.4.2 Deuxième objectif : Dans quelle mesure la prosodie du FQ affecte-t-elle l'intelligibilité ou la qualité d'une synthèse ? 45
2.4.3 Troisième objectif : Quels paramètres prosodiques ou combinaison de paramètres sont les plus pertinents dans l'amélioration d'une synthèse de la parole en FQ ? 46
CHAPITRE III
MÉTHODOLOGIE 47
3.1 Survol du protocole de tests 47
3.2 Étape 1 : Enregistrements 49
3.2.1 Corpus 49
3.2.2 Choix des sujets 51
3.2.3 Conditions d'enregistrement 51
3.3 Étape 2 : Dégagement des archétypes 52
3.3.1 Découpage syllabique avec Praat 52
3.3.2 Délexicalisation 53
3.3.3 Accord interjuges 58

3.4 Étape 3 : Synthèse et évaluation 61
3.4.1 Analyse et dégagement de règles 61
3.4.2 Régression logistique 63
3.4.3 Synthèse 69
3.4.4 Test de perception 78
3.5 Survol des objectifs 83
CHAPITRE IV
PRÉSENTATION DES RÉSULTATS 85
4.1 Accord interjuges 85
4.2 Analyse acoustique 86
4.3 Test de perception 90
4.3.1 Résultats généraux 90
4.3.2 Cohérence interne des séries 98
4.3.3 Matrice de corrélation 119
CHAPITRE V
DISCUSSION 124
5.1 Premier objectif : Identifier certaines caractéristiques prosodiques principales du français québécois, dans sa perception et sa production 124
5.1.1 Hypothèse 1 : Des différences prosodiques seront observées, à la perception comme à la production entre le FF et le FQ 125
5.1.2 Hypothèse 2 : Les différences prosodiques observées à la production se trouveront principalement au niveau de la durée 125
5.1.3 Hypothèse 3 : Les différences prosodiques observées à la perception se trouveront principalement au niveau de la F0 126
5.2 Deuxième objectif : Dans quelle mesure la prosodie du FQ affecte-t-elle l'intelligibilité ou la qualité d'une synthèse 128
5.2.1 Hypothèse 4 : Un modèle FF de la prosodie (F0 et durée) sur une production segmentale FQ entraîne une baisse d'intelligibilité et de qualité 129
5.3 Troisième objectif : Quels paramètres prosodiques ou combinaison de paramètres sont les plus pertinents dans l'amélioration d'une synthèse de la parole en FQ ? 131

5.3.1 Hypothèse 5 : D'une manière générale, c'est la F0 qui aura le plus d'impact, à la fois sur l'intelligibilité et sur la qualité..... 131

CONCLUSION 133

Appendices 136

APPENDICE A
EXEMPLE DE FORMULAIRE DE CONSENTEMENT 137

APPENDICE B
STIMULI DU TEST DE PERCEPTION 138

APPENDICE C
FICHIER .MLC TIRÉ DE EULER 144

APPENDICE D
GRAPHIQUES 147

APPENDICE E
MODIFICATION DES PARAMÈTRES PROSODIQUES 150

APPENDICE F
INTERFACE DU TEST DE PERCEPTION 158

APPENDICE G
SCRIPT PRAAT DE DÉLEXICALISATION 161

APPENDICE H
ANOVAS DES SÉRIES VS L'ENSEMBLE DES RÉPONSES 166

APPENDICE I
ANOVAS DES COHÉRENCES INTERNES DES SÉRIES 175

RÉFÉRENCES 184

LISTE DES FIGURES

Figure 1-1 Modules d'un système de synthèse 7
Figure 1-2 Exemples d'oscillogrammes 13
Figure 1-3 Amplitude, période et fréquence 14
Figure 1-4 Exemple de spectre 14
Figure 1-5 Exemple de spectrogramme représentant la suite [sa] 15
Figure 1-6 Représentation graphique des principes d'isochronie 19
Figure 1-7 Typologie des types d'accents 20
Figure 1-8 Organisation de la syllabe 22
Figure 1-9 Schéma du syntagme intonatif selon Jun et Fougeron (2000, p.214) 24
Figure 3-1 Survol du protocole de tests 49
Figure 3-2 Syllabation à l'aide de Praat 53
Figure 3-3 Phrase délexicalisée par filtrage (filtre passe-bas) 54
Figure 3-4 Phrase délexicalisée par *pulse train* 55
Figure 3-5 Phrase originale : « Le chat mange la souris » 57
Figure 3-6 Phrase resynthétisée 58
Figure 3-7 Interface de l'accord interjuges 59
Figure 3-8 Annotation des archétypes 61
Figure 3-9 Interface Mbrola 73
Figure 3-10 Étapes du test de perception 80
Figure 4-1 Arbre de classification pour l'origine perçue 88
Figure 4-2 Arbre de classification pour l'origine réelle 89

Figure 4-3 Distribution des données origine 93
Figure 4-4 Répartition des données MOS 94
Figure 4-5 Répartition des données pour le confort d'écoute 95
Figure 4-6 Répartition des données pour le naturel 96
Figure 4-7 Répartition des données pour l'intelligibilité 97
Figure 4-8 ANOVA cohérence interne de la série registre pour la question origine 102
Figure 4-9 ANOVA cohérence interne de la série registre pour la question MOS 103
Figure 4-10 Kruskal-Wallis cohérence interne de la série registre pour la question intelligibilité 104
Figure 4-11 ANOVA cohérence interne de la série F0 TH/TB pour la question origine 108
Figure 4-12 Anova Kruskal-Wallis pour la variable intelligibilité 112
Figure 4-13 ANOVA cohérence interne F0 pour la question origine 114
Figure 4-14 ANOVA cohérence interne série F0 : MOS 115
Figure 4-15 ANOVA cohérence interne série F0 : naturel 115
Figure 4-16 ANOVA Kruskal-Wallis pour la variable intelligibilité, série 7 116
Figure 4-17 Corrélation origine vs intel 121
Figure 4-18 Corrélation MOS vs origine 122
Figure 4-19 Corrélation naturel vs origine 123

LISTE DES TABLEAUX

Tableau 1-1 Exemples de phrases SUS (*Semantically Unpredictable Sentences*) 10

Tableau 1-2 Évaluation ACR de la qualité des voix de synthèse appliquée au français 12

Tableau 2-1 Perte d'intelligibilité selon le facteur prosodique manipulé 31

Tableau 2-2 Études comparatives des français québécois et hexagonaux avant 1995 36

Tableau 2-3 Marqueurs différenciant le FF du FQ 38

Tableau 3-1 Corpus initial 50

Tableau 3-2 Profil sociodémographique des locuteurs 51

Tableau 3-3 Données sociodémographiques des juges 59

Tableau 3-4 Données acoustiques des archétypes 62

Tableau 3-5 Variables acoustiques principales 66

Tableau 3-6 Variables statistiquement significatives pour le groupe *Origine perçue* 68

Tableau 3-7 Variables statistiquement significatives pour le groupe *Origine réelle* 68

Tableau 3-8 Cibles des cinq paramètres prosodiques retenus 70

Tableau 3-9 Cibles des paramètres prosodiques retenus 71

Tableau 3-10 Structures de phrase 72

Tableau 3-11 Ordre d'application des fonctions de modification 77

Tableau 3-12 Données sociodémographiques des participants 79

Tableau 4-1 Résultats de l'accord interjuges 86

Tableau 4-2 Résultats généraux du test de perception 91
Tableau 4-3 Cohérence interne des séries : F0 moyenne 99
Tableau 4-4 Cohérence interne des séries : Registre 101
Tableau 4-5 Cohérence interne de la série sylDur TH/TB 105
Tableau 4-6 Cohérence interne des séries : F0 TH/TB 107
Tableau 4-7 Cohérence interne des séries : durée des voyelles 110
Tableau 4-8 Cohérence interne des séries : durée 111
Tableau 4-9 Cohérence interne des séries : F0 113
Tableau 4-10 Seuils retenus pour la série 8 117
Tableau 4-11 Cohérence interne des séries : F0 et durée 117
Tableau 4-12 Seuils pour la série F0 et durée inversées 118
Tableau 4-13 Cohérence interne des séries : F0 et durée inversées 118
Tableau 4-14 Matrice de corrélation pour le test de perception 120
Tableau 5-1 Classification en fonction de l'origine réelle des locuteurs 126
Tableau 5-2 Classification en fonction de l'origine perçue des locuteurs 127
Tableau 5-3 Résultats pour le seuil FF 130

LISTE DES ABRÉVIATIONS SIGLES ET ACRONYMES

A	Amplitude
ACR	*Absolute Category Rating*
ANOVA	*Analysis of Variance* (analyse de variance)
dB	Décibel
F	Fréquence
FF	Français de France
FQ	Français du Québec
F0	Fréquence fondamentale
GR	Groupe rythmique
HTML	*HyperText Mark-up Language*
Hz	Hertz
MOS	*Mean Opinion Score*
ms	Milliseconde
MRT	*Modified Rhyme Test*
PC	*Personnal computer* (Ordinateur personnel)
PSOLA	*Pitch Synchronous Overlapp and Add*
SA	Syntagme accentuel
SI	Syntagmes intonatifs

SUS	*Semantically Unpredictable Sentences*
SPIN	*Speech Perception in Noise*
T	Période
TB	Ton bas
TH	Ton haut
TNM	Ton non marqué
TTS	*Text-to-Speech*
UQAM	Université du Québec à Montréal

RÉSUMÉ

Ce mémoire vise à déterminer l'impact de la prise en compte de la prosodie du français québécois sur l'intelligibilité et la qualité d'un système de synthèse. Un accord interjuges, effectué sur la base d'énoncés français de France (FF) et français du Québec (FQ) délexicalisés, a d'abord permis d'isoler cinq paramètres prosodiques permettant de différencier ces deux variétés, et ce, sur la base de la prosodie uniquement. Ces cinq paramètres prosodiques ont été implémentés en synthèse à l'aide du système Mbrola. Quinze participants ont jugé ces énoncés sur la base de l'origine géographique perçue, l'intelligibilité et la qualité des énoncés. Il ressort de ce test que des différences prosodiques existent bel et bien, à la perception comme à la production entre FF et FQ. De plus, l'utilisation d'une prosodie linguistiquement erronée – par exemple une prosodie FF sur une production segmentale FQ – entraîne une baisse d'intelligibilité et de qualité des énoncés synthétisés.

Mots clés : Prosodie, français québécois, synthèse de la parole

INTRODUCTION

La communication humaine consiste en l'interaction de deux processus : la production et la perception. En production, le locuteur articule des idées en mots par la parole. La perception, quant à elle, se charge de comprendre ce qui est dit par un interlocuteur, elle permet de décoder les mots et de les transformer en idées.

Les technologies vocales cherchent à reproduire ces deux mécanismes. La synthèse de la parole tente de reproduire artificiellement le volet production, alors que la reconnaissance de la parole se charge du volet perception.

La synthèse et la reconnaissance de la parole ont évolué de façon indépendante depuis cinquante ans (Dutoit et coll. 2002). Les équipes travaillant sur ces deux domaines le font généralement en vase clos, même au sein d'une même entreprise. En raison de la nature du problème – la reconnaissance devant considérer la variabilité, la synthèse cherchant plutôt à créer un seul locuteur « parfait » – les deux disciplines ont évolué en parallèle, trouvant souvent des stratégies différentes à des problèmes similaires.

Dans les faits, la recherche dans le domaine des technologies vocales vise souvent à permettre le développement d'un nouveau produit, d'une nouvelle fonctionnalité. Comme on a pu le constater ces dernières années, les progrès sont gigantesques. Les technologies vocales sont mises à partie dans de nombreux systèmes téléphoniques commerciaux, dans les voitures et sur les ordinateurs, entre autres applications.

Dans le cadre de ce mémoire, nous nous intéresserons à la synthèse vocale. Si la synthèse de la parole est très avancée pour certaines langues, particulièrement pour l'anglais, des ajustements linguistiques sont à apporter pour d'autres langues et un travail important doit être fait lorsqu'il est question de dialectes.

En effet, la justesse d'une synthèse ne dépend pas seulement de la qualité de son ingénierie, mais aussi du modèle linguistique qui la sous-tend. C'est ce modèle qui sera exploré ici, dans une perspective de synthèse vocale adaptée au français québécois.

Ce mémoire est divisé en cinq chapitres. Le premier présente la problématique de la synthèse vocale au Québec et son cadre théorique. Dans le second chapitre, une revue de la littérature cherche à établir ce qui est connu sur le sujet et présente la question de recherche, les objectifs et les hypothèses. Le troisième chapitre élabore la méthodologie utilisée afin de répondre aux objectifs. Dans le quatrième chapitre, les résultats sont présentés. Le dernier chapitre, quant à lui, offre une discussion et tente de répondre à la question de recherche.

CHAPITRE I

PROBLÉMATIQUE ET CADRE THÉORIQUE

Ce mémoire vise à offrir une contribution linguistique à un système de synthèse de la parole en français québécois. Nous nous retrouvons donc aux limites de trois disciplines qui seront abordées dans ce chapitre : la synthèse de la parole, la phonétique et la sociolinguistique. Dans le cadre de cette exploration, les défis particuliers auxquels fait face la synthèse au Québec seront identifiés. De plus, les concepts et notions nécessaires à la compréhension du problème seront définis et expliqués.

Au Québec, de nombreuses clientèles sont susceptibles de faire usage d'une voix de synthèse. Les 136 000 non-voyants québécois, d'abord, s'en servent pour accéder à l'informatique. Les outils d'aide à la communication[1] pour les personnes atteintes d'un trouble moteur ou langagier limitant la parole comprennent bien souvent une synthèse vocale (Beukelman et Mirenda 2005). De plus, au Québec, des milliers de personnes utilisent les voix de synthèse dans un cadre moins formel, que ce soit pour

[1] Outils visant à aider ou remplacer la parole humaine. Les utilisateurs de ces outils les transportent avec eux en tout temps. La voix de synthèse choisie devient alors la voix de l'usager.

le système mains libres de leur voiture ou pour communiquer avec leur téléphone intelligent.

Dans les deux cas, il importe que la voix québécoise disponible soit compréhensible et agréable à écouter. Or, l'étude de Côté-Giroux (2011) – dont il sera question plus en détail dans le chapitre 2 – montre que les synthèses québécoises sont généralement moins appréciées que les voix originaires de France.

1.1 Synthèse de la parole

Les recherches entourant la génération automatique de la parole ont débuté dans les années 1950 et ont amené une grande variété de systèmes de synthèse. En effet, selon l'objectif recherché, diverses stratégies ont été préférées et certains modules, ou sections, ont été mieux développés que d'autres. Cette section vise à décrire les différents types de systèmes et les modules typiques qui les composent, ainsi que leurs avantages et inconvénients.

De nombreuses méthodes sont utilisées pour générer des voix artificielles, allant de la plus simple à la plus complexe. Considérant que toute forme de manipulation de la parole pourrait être considérée comme de la synthèse, il convient de distinguer des niveaux de complexité et ce qu'on entend généralement par synthèse (O'Shaughnessy 2007).

La plupart des systèmes de parole automatisée utilisés aujourd'hui ne sont pas, en fait, des systèmes de synthèse. Généralement, il s'agit simplement de phrases préenregistrées. La plupart des systèmes dans les voitures, par exemple, ne nécessitent qu'un inventaire d'une cinquantaine de phrases-clés. Il en va de même pour plusieurs systèmes téléphoniques. Ils ont un domaine d'application très limité, ce qui réduit l'étendue du vocabulaire nécessaire. Dans ces cas, on ne considère pas qu'il s'agisse de véritable synthèse de la parole (au sens strict) (O'Shaughnessy 2007).

Toutefois, on remarque que d'autres situations exigent un système plus flexible. Un système téléphonique, qui, par exemple, devrait pouvoir nommer les noms dans l'annuaire, doit pouvoir le faire à partir du texte. Il serait contre-productif d'employer un professionnel pour faire la lecture de tous les noms puisque le système résultant serait beaucoup trop lourd. Il s'agit donc de créer un système qui puisse prendre un texte écrit et le convertir en parole. On parle aussi, dans le cas de l'annuaire téléphonique, d'un vocabulaire ouvert[2], ce qui signifie que l'on ne veut pas stocker des phrases entières ou des noms, mais plutôt de petites unités concaténables – petites unités de parole pouvant être jointes les unes aux autres. C'est à un système de ce type auquel on fait généralement référence en parlant de synthèse vocale : « un processus informatique de composition sonore permettant la transformation d'un texte en voix artificielle » (Côté-Giroux et coll., 2011).

1.1.1 Les différents types de systèmes

On peut diviser les différents systèmes de synthèse en deux grandes catégories, selon le type d'unités de base qu'ils sélectionnent (O'Shaughnessy 2007) : les systèmes paramétriques et les systèmes de synthèse par concaténation. Le premier type de systèmes – appelé *systèmes paramétriques*, ou synthèse par règle – cherche à reproduire la parole humaine telle que produite par le conduit vocal (O'Shaughnessy 2007). Ainsi, cet ensemble de systèmes comprend un module qui crée un son de source, par la suite modulé grâce à une série de paramètres correspondant généralement aux traits articulatoires. De plus en plus, les modèles employant cette méthode utilisent une approche purement statistique dans la modulation du son de source. Ils obtiennent des résultats acceptables et, dans certains cas, de meilleure qualité que la synthèse par sélection d'unités – sous-type de synthèse par

[2] Un vocabulaire fermé consiste en une banque de mots qui ne pourra pas être modifiée. Un vocabulaire ouvert, quant à lui, contient une quantité potentiellement infinie de mots – ce qui est le cas, entre autres, des noms propres.

concaténation (King 2010; Zen, Tokuda et Black 2009). Cependant, lorsqu'il est question d'émotions ou d'intonation et de rythme, les méthodes par concaténation sont souvent meilleures (Barra-Chicote et coll. 2010).

Plusieurs auteurs ont exploré la synthèse paramétrique, notamment en japonais (Zen, Tokuda et Black 2009), en anglais (King 2010; Klatt 1980), et en français européen (Lanchantin 2010). Ces systèmes sont employés principalement pour tester des hypothèses de recherches. Néanmoins, à notre connaissance, il n'existe aucun système de ce type appliqué au français québécois (désormais FQ).

Le deuxième type de système s'appelle *synthèse par concaténation*. Il s'agit, en fait, d'enregistrer un locuteur professionnel (*voice talent* [3]) et de découper le signal de parole en petites unités (phonèmes, diphones, triphones, etc.) qu'un algorithme vient ensuite concaténer (O'Shaughnessy 2007). Ce type de système est celui qui est généralement utilisé au-delà d'un cadre scientifique. La voix, puisqu'elle provient d'un être humain, est généralement intelligible et relativement naturelle (van Santen et coll. 1997; O'Shaughnessy 2007; Tamura et coll. 2010)[4]. Les difficultés surviennent surtout lors de l'assemblage des unités. En effet, superposer deux unités dont les limites ne coïncident pas nécessairement engendre un son qui semble interrompu et, d'une manière générale, robotisé (Tamura et coll. 2010). Néanmoins, des avancées récentes en traitement du signal ainsi que l'amélioration des systèmes de stockage offrant des bases de données plus larges d'unités de base permettent aujourd'hui à ces types de systèmes d'être les plus performants et donc les plus utilisés commercialement (King 2010).

---

[3] Professionnel dont le travail consiste à produire de la parole sans émotion à des fins d'enregistrements.

[4] L'adjectif *naturel* sera utilisé afin de distinguer le côté humain d'une voix de synthèse, par opposition à une voix robotisée.

1.1.2 Les modules des systèmes de synthèse

Un système de synthèse comporte deux grands modules : le traitement linguistique et le traitement du signal. Le traitement linguistique se compose de différentes étapes : l'analyse syntaxique, la phonétisation et la génération de la prosodie. Le traitement du signal, quant à lui, sélectionne les fichiers sonores dans la base de données et effectue les opérations de concaténation et de lissage entre les unités. Ce dernier module est généralement conçu et implémenté par des ingénieurs. On peut voir une représentation des grandes étapes de la synthèse dans la figure 1-2 inspirée de la classification faite par Goldman en 2001.

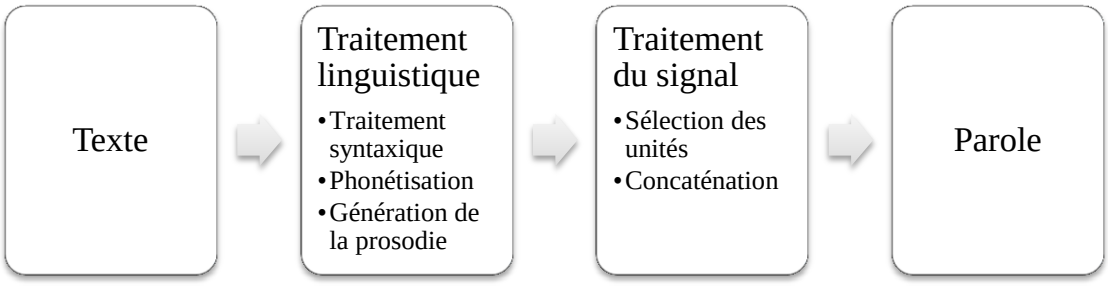


Figure 1-1 Modules d'un système de synthèse

L'*analyse syntaxique* est une étape qui a longtemps été jugée facultative, mais qui devient de plus en plus présente grâce à l'amélioration des analyseurs syntaxiques (Obin et coll. 2010). Il s'agit d'effectuer un premier découpage syntaxique pour permettre de rendre compte ou d'améliorer plusieurs phénomènes, tels que la liaison ou l'intonation.

La *phonétisation*, aussi appelée « conversion graphème-phonème », représente généralement le cœur du traitement linguistique dans un système de synthèse. Il s'agit de l'étape où il faut convertir les mots – ou graphèmes – en phonèmes. Pour réaliser cette conversion, il faut tenir compte des phénomènes phonologiques de la langue

(liaison, élision) et du fait que de nombreux mots inconnus puissent apparaître (noms propres, mots étrangers, sigles).

Le *module prosodique* d'un système de synthèse est essentiel puisque c'est celui-ci qui génère, entre autres, la durée de chaque son et des pauses, l'intonation et la force de chaque élément. Il s'agit, dans un premier temps, de faire une représentation phonologique intermédiaire de la prosodie, soit de regrouper les différentes unités. Par exemple, les différents sons sont d'abord regroupés en syllabes, puis en petits groupes de rythme. Par la suite, le module doit quantifier la réalisation de chaque phénomène. Autrement dit, il faut assigner des valeurs de durée et d'intonation pour chaque voyelle, chaque syllabe, par exemple. Une description plus détaillée de la prosodie sera présentée plus loin, tout comme les différentes manières de l'implémenter en synthèse.

1.1.3 L'évaluation de la qualité des systèmes

La performance d'un système de synthèse se mesure par une combinaison de facteurs : la qualité générale (appréciation et intelligibilité), le coût (complexité et lourdeur) et le délai (instantané ou en différé) (O'Shaughnessy 2007). Le seul de ces trois facteurs pouvant être influencé par une contribution linguistique, par opposition à un travail d'ingénierie, est la qualité générale. Ainsi, ce sera d'intelligibilité et d'appréciation dont il sera question dans la présente étude.

L'*intelligibilité* correspond à la facilité avec laquelle la parole de synthèse est perçue par l'auditeur – autrement dit la capacité à identifier correctement les phonèmes. Plusieurs éléments peuvent influencer l'intelligibilité d'une voix : le bruit ambiant, la longueur et la complexité des énoncés, la prévisibilité des phrases ou le débit des stimuli (Côté-Giroux et coll., 2011). Les études de voix de synthèse dégagent généralement cinq critères principaux pouvant influencer une tâche d'intelligibilité effectuée en laboratoire : la qualité du signal, la taille et la complexité du message, les capacités de mémoire à court terme du participant, la complexité de la tâche ou la

présence de tâches concurrentes et finalement l'expérience du participant avec le système (Lai, Wood et Considine 2000).

L'intelligibilité se mesure de plusieurs façons. On peut simplement demander aux auditeurs de transcrire ce qu'ils entendent dans un formulaire, de répéter le stimulus entendu ou de répondre à des questions précises concernant le message entendu. Les synthèses actuelles sont très performantes au point de vue de l'intelligibilité (Mayo, Clark et King 2011) et d'excellents résultats sont généralement obtenus. Compte tenu de cette situation, si on cherche à vérifier l'impact d'une modification particulière de la voix de synthèse, il peut être pertinent d'augmenter la difficulté de la tâche (ajouter du bruit ou changer la longueur des phrases, par exemple) pour éviter un effet de plafond.

Dans le cadre de la recherche en synthèse, plusieurs tests standardisés existent afin de mettre les nouveaux produits à l'épreuve. Deux d'entre eux sont retenus ici. Ce sont ceux utilisés lors de la compétition annuelle, le *Blizzard Challenge*, visant à tester les avancées en synthèse (Black et Tokuda 2005).

D'abord, dans le but d'évaluer l'aspect segmental de la synthèse, ou l'intelligibilité lexicale, le MRT (*Modified Rhyme Test*) (Kreul et coll. 1968) – ou une de ses variantes – est généralement employé, notamment dans le *Blizzard Challenge*. Il s'agit d'un test utilisant une phrase porteuse présentant des paires minimales[5]. Le phonème testé peut se trouver en début, milieu ou fin de mot, selon la variante de test utilisée.

Dans le but d'évaluer l'impact des phrases ou de la prosodie sur l'intelligibilité, le test SUS (*Semantically Unpredictable Sentences*) est généralement employé. Ce test présente des phrases utilisant cinq structures syntaxiques de base et des mots fréquents. Ce test est appliqué à plusieurs langues et, lorsque contrôlé, permet de faire

[5] Paires de mots ne variant que par un seul phonème, comme mère [m] et père [p].

des comparaisons efficaces entre différents synthétiseurs (Benoît, Grice et Hazan 1996). Dans le cadre du français, une banque de phrases SUS a été générée et contrôlée dans l'objectif d'effectuer des recherches en synthèse de la parole. Elles ont également été testées afin d'être aptes à être présentées dans le bruit. Le tableau suivant, tiré de Mareüil, D'Alessandro et Raake (2006), offre un exemple de ce type d'énoncés.

Tableau 1-1 Exemples de phrases SUS (*Semantically Unpredictable Sentences*)

La loi brille par la chance creuse.
La classe gaie montre le frein.
Quand le lien signe-t-il l'onde pleine ?
Le test clair mange la haine.
L'or jaune porte le dôme.
Comment la soif lance-t-elle le bol proche ?
Le mur siffle la buée qui vole.

Dans certains cas, notamment lorsqu'il est question de variation dialectale (ce sujet sera abordé plus loin dans ce chapitre), il peut être pertinent d'augmenter la difficulté de la tâche en ajoutant un bruit de fond venant masquer partiellement le stimulus. Dans ces cas, un corpus test est prévu spécifiquement pour cette tâche (*Speech Perception in Noise – SPIN)* (Kalikow, Stevens et Elliott 1977). Ce sont plusieurs séries de phrases classées selon leur prédictibilité. Les phrases choisies par l'expérimentateur sont ensuite présentées selon divers niveaux de bruit (Clopper et Bradlow 2008b).

L'*appréciation* d'une voix de synthèse est un objectif auquel de plus en plus de chercheurs en synthèse de la parole s'intéressent. En effet, on pourrait être porté à croire qu'il s'agit d'un objectif secondaire, l'objectif premier étant l'intelligibilité. Or, il appert que la qualité générale ou le naturel d'une voix a un impact très important, sinon plus que l'intelligibilité, dans l'adoption et l'utilisation des systèmes de synthèse dans la population en général (Nusbaum, Francis et Henly 1995). Il s'agit donc d'un facteur primordial sur lequel les chercheurs doivent se pencher.

Sa mesure, par ailleurs, est hautement subjective (Côté-Giroux et coll. 2011) puisque plusieurs facteurs viennent influencer son appréciation : la qualité segmentale de la voix, l'aspect naturel ou la qualité de la prosodie, entre autres. Plusieurs études démontrent que la prosodie joue un rôle crucial dans l'appréciation des voix de synthèse (Nusbaum, Francis et Henly, 1995; Paris, Thomas, Gilson et Kincaid, 2000).

L'évaluation de la qualité des voix de synthèse, en plus d'être subjective, pose de nombreux problèmes. En effet, plusieurs éléments indépendants viennent affecter les résultats. Par exemple, une mauvaise intelligibilité influencera forcément les scores d'appréciation. De plus, un élément particulier, le *naturel*, est régulièrement utilisé comme synonyme interchangeable avec le mot *appréciation* dans les études – plusieurs chercheurs se contentant de présenter une évaluation de l'intelligibilité suivie d'une évaluation du naturel – évaluation qui ne consiste en fait qu'en une appréciation générale de la voix.

Il n'existe pour le moment aucune définition objective et universellement reconnue du naturel, c'est une qualité subjective de la voix – la distinction entre une voix naturelle et une voix artificielle (Nusbaum, Francis et Henly 1995). Plusieurs aspects viennent influencer l'évaluation du naturel, notamment la structure prosodique, les caractéristiques du son de source ou la composition segmentale. Il est important dans la création des tests de prendre cet aspect en considération et d'utiliser au besoin des tests différents pour ces deux aspects.

Dans les tests standardisés en synthèse, notamment dans le *Blizzard Challenge* présenté dans la section plus haut, la mesure utilisée pour tester la qualité est le MOS (*Mean Opinion Scores*) sur une échelle de 1 à 5. Lorsque la quantité des stimuli le permet (s'il y en a peu), les chercheurs préfèrent faire une comparaison en paires mais cela se fait très rarement car il y a généralement de trop nombreux stimuli. Ce test permet d'obtenir une évaluation globale du système et de son acceptabilité auprès de la population générale. Cependant, il ne permet pas d'effectuer de diagnostic sur une composante précise du système, comme la prosodie (Viswanathan et Viswanathan

2005). En effet, étant très général, il ne permet que d'offrir un aperçu global de l'appréciation.

En français, l'évaluation standard a été développée par Boula de Mareüil et coll. (2006). Cette évaluation comprend une évaluation générale, le MOS, suivie d'une évaluation plus fine de divers éléments, soit un test ACR (*Absolute Category Rating*).

C'est un test en six catégories plus le MOS, donc sept catégories présentées sur une échelle continue ensuite divisée en cinq points. Cette échelle peut être vue dans le tableau 1-2, tiré de Boula de Mareüil et coll. (2006).

Lorsqu'il est question de l'évaluation spécifique de l'impact de la prosodie sur une synthèse de la parole il n'existe, à notre connaissance, aucun test standardisé universellement reconnu. En effet, de multiples tests existent visant à explorer divers aspects prosodiques. Ces divers tests seront vus plus en détail dans le chapitre 3.

Tableau 1-2 Évaluation ACR de la qualité des voix de synthèse appliquée au français

MOS
Comment appréciez-vous globalement ce que vous venez d'entendre ? Très mauvais – très bon
Compréhension
Comment décririez-vous la facilité à comprendre le message ? Très difficile – très facile
Confort d'écoute
Comment décririez-vous cette voix ? Très désagréable – très agréable
Non-monotonie
Évaluez le caractère monotone ou varié de ce que vous venez d'entendre. Très monotone – Très varié
Naturel
Comment apprécieriez-vous le naturel de ce que vous venez d'entendre ? Très artificiel – très naturel
Fluidité
Comment appréciez-vous le côté haché ou fluide de l'élocution ? Très haché – très fluide
Prononciation
Avez-vous remarqué des problèmes de prononciation ? Très gênant – aucun problème

1.2 Phonétique acoustique

Puisque le produit fini de la synthèse de la parole est une onde acoustique, il importe de bien en maîtriser les composantes. Pour ce faire, nous présentons dans cette section les rudiments de la phonétique acoustique.

Un *son* est en fait le déplacement dans l'air d'une fluctuation de pression à une vitesse approximative de 340 mètres par seconde (Martin, 1996, p.142). Une représentation de l'onde est faite à l'aide d'un graphique appelé oscillogramme, comme on peut le voir dans la figure 1-2.

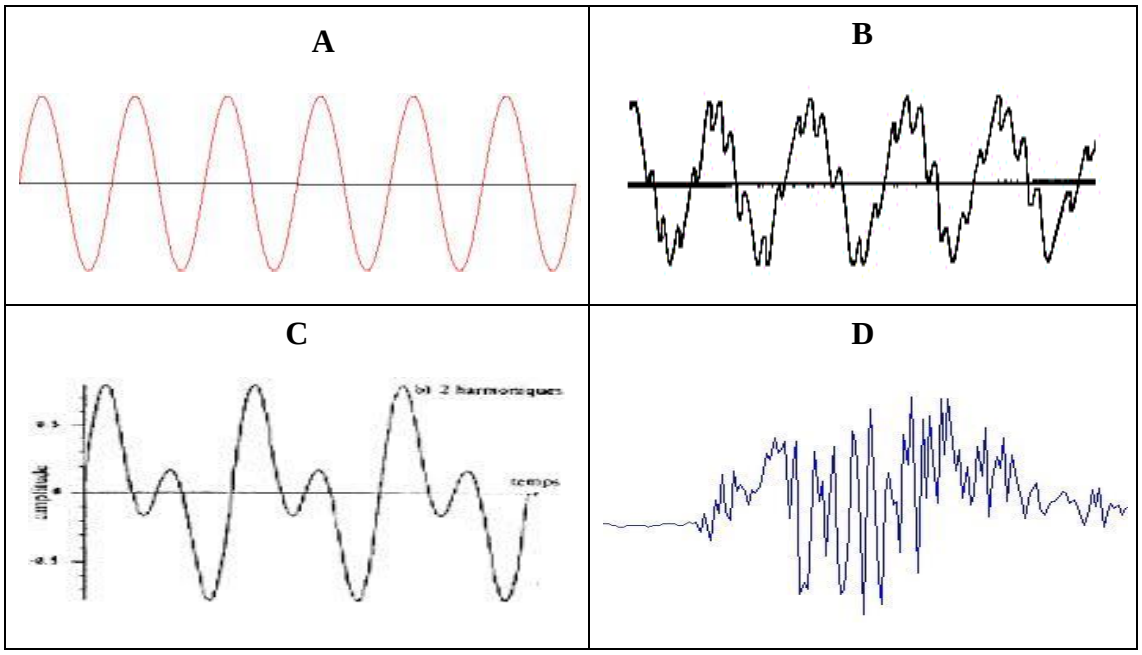


Figure 1-2 Exemples d'oscillogrammes

(A) – onde simple (B) onde complexe (C) onde périodique (D) onde apériodique

Sur un oscillogramme, l'axe des abscisses représente le temps et l'axe des ordonnées représente l'amplitude.

Selon sa forme, l'onde peut être simple ou complexe, périodique ou apériodique. Une *onde simple* n'est composée que d'une seule onde sinusoïdale (figure 1-2-A). Une

onde complexe est la somme de plusieurs ondes simples (figure 1-2-B). L'*onde périodique* (figure 1-2-C), quant à elle, est constituée d'un patron régulier qui se répète dans le temps, tandis qu'une *onde apériodique* (figure 1-2-D) est créée par des variations désordonnées. Trois éléments quantifiables permettent de décrire l'onde : la période (T), l'amplitude (A) et la fréquence (F). Ces éléments peuvent être visualisés à l'aide de la figure 1-3.

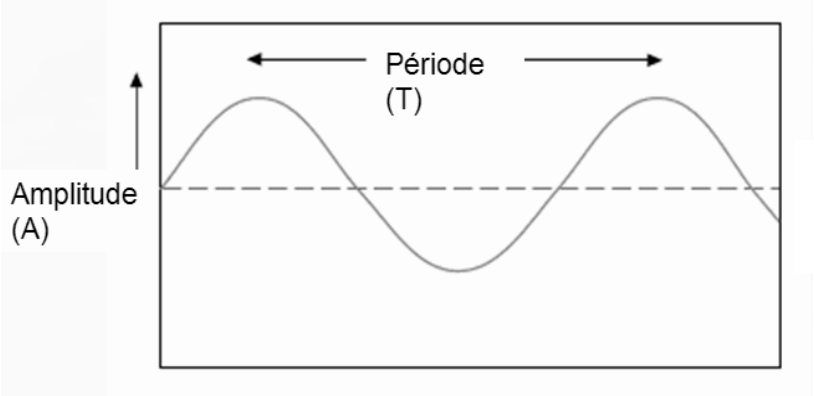


Figure 1-3 Amplitude, période et fréquence

Il peut être pertinent de décomposer l'onde complexe afin d'en étudier les diverses composantes. Un oscillogramme ne permet pas de visualiser adéquatement certain éléments pertinents; un spectre de raies est donc un autre graphique utilisé, représentant les données avec l'amplitude sur l'axe des ordonnées et la fréquence sur l'axe des abscisses, comme on peut voir dans la figure 1-4.

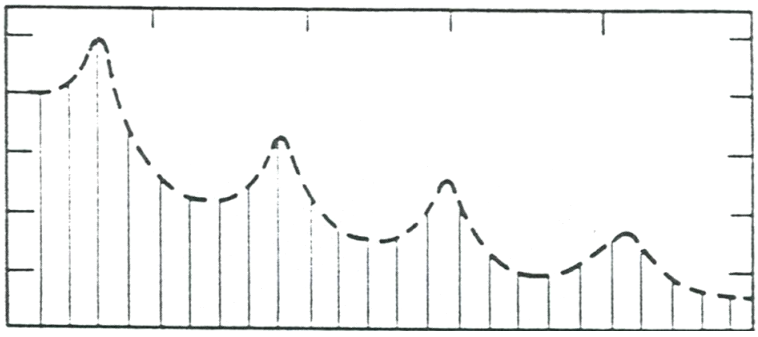


Figure 1-4 Exemple de spectre

Lorsqu'on décompose une onde complexe, on obtient plusieurs ondes simples : la fréquence fondamentale (l'onde ayant la plus basse fréquence) et un grand nombre d'harmoniques qui sont en fait des multiples entiers de la fréquence fondamentale. Sur le spectre, la première raie (première ligne verticale) représente la fréquence fondamentale et les raies suivantes sont les harmoniques. Avec la voix humaine, on remarque une caractéristique supplémentaire : le conduit vocal, par sa configuration particulière au moment de l'articulation, module l'onde acoustique et amplifie certains harmoniques. Ces harmoniques amplifiés sont les *formants*. Ce sont, sur le spectre de raies, les pics observés.

La parole, par sa nature même, varie énormément en fonction du temps. Il est donc très utile d'avoir une représentation graphique qui tient compte non seulement du temps, mais aussi des fréquences et de l'amplitude. Cette représentation graphique s'appelle le spectrogramme (figure 1-5).

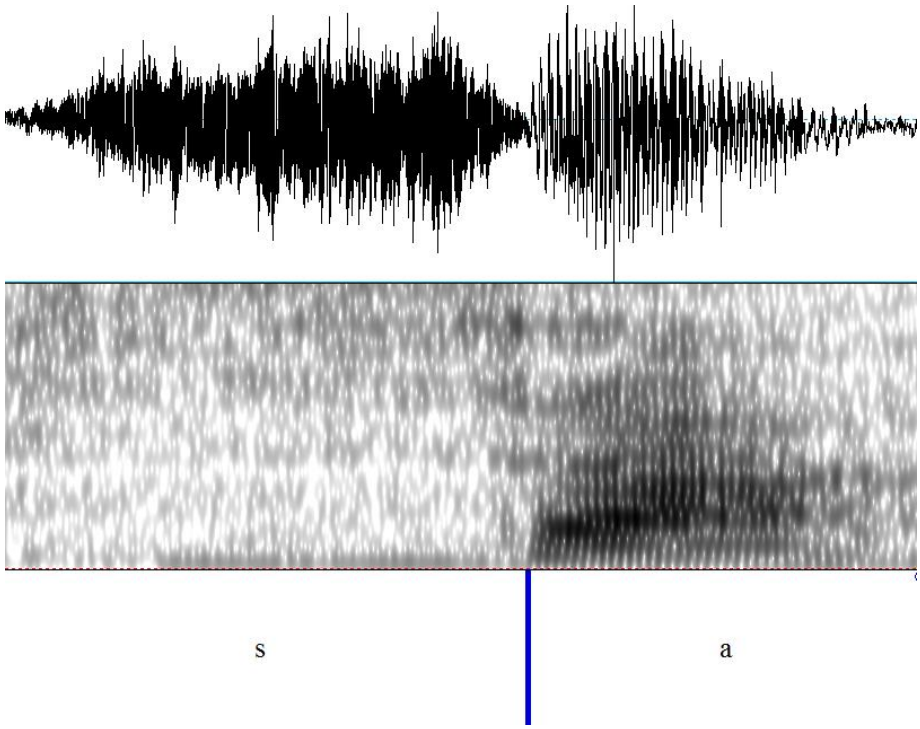


Figure 1-5 Exemple de spectrogramme représentant la suite [sa]

Sur cette figure, l'axe des ordonnées représente la fréquence, l'axe des abscisses, le temps alors que l'amplitude est représentée selon les niveaux de gris.

Comme on peut le voir, la parole humaine est composée d'ondes complexes, périodiques ou apériodiques. La consonne [s] illustrée dans le spectrogramme de la figure 1-5 est un exemple d'onde apériodique. La voyelle [a], par contre, est un exemple d'onde périodique complexe.

1.3 Prosodie

La parole est généralement divisée en deux grandes composantes : la composante segmentale (les sons qui sont produits) et la composante suprasegmentale (ou prosodie) (Martin, 1996, p.165). Dans le cadre de ce mémoire, nous nous concentrerons sur la description de ce deuxième volet d'un point de vue phonétique.

Tel qu'indiqué précédemment, la prosodie est un élément essentiel pour l'amélioration de la synthèse de la parole puisqu'elle a un impact à la fois sur l'intelligibilité et l'appréciation des systèmes (Monaghan 2002). Elle est, toutefois, complexe et, de surcroît, sa description ne fait pas l'unanimité chez les linguistes qui l'étudient. Deux visions principales s'opposent dans la définition de la prosodie (Fujisaki 1995). La première repose principalement sur les mesures acoustiques de la prosodie. La seconde se préoccupe davantage des niveaux de représentation et de l'organisation du système prosodique.

Dans une perspective de synthèse, la prosodie se définit généralement de la manière suivante :

> La prosodie est l'organisation systématique d'unités linguistiques variées en énoncés ou en groupes cohérents d'énoncés dans le processus de production de la parole. Sa réalisation comprend des traits à la fois segmentaux et

> suprasegmentaux et sert à transmettre des informations linguistiques, paralinguistiques et non-linguistiques.[6] (Fujisaki 1995)

Cette définition permet de rendre compte des deux grandes dimensions de la prosodie : son organisation et sa réalisation. De plus, elle met en lumière les multiples fonctions qu'elle remplit dans le langage.

1.3.1 Prosodie et paramètres acoustiques

Au niveau acoustique, la prosodie se manifeste de trois façons : par la fréquence fondamentale (F0), la durée et l'intensité.

La F0 se mesure en Hertz (Hz) et correspond à la *fréquence fondamentale*, c'est-à-dire la fréquence de l'onde sonore produite à la source. Elle est liée à la vitesse de vibration des cordes vocales. Elle donne à chaque personne sa hauteur de voix personnelle, qu'elle soit aigüe ou grave. Elle est traditionnellement associée à l'intonation ou à la mélodie de la voix. Ainsi, un locuteur dont la voix est perçue comme « chantante » produit généralement des variations de F0 plus grandes que la moyenne. La *durée*, quant à elle, se mesure en millisecondes (ms) et correspond à la durée relative de certains segments par rapport à d'autres. À titre d'exemple, dans la phrase « C'est LE petit garçon, pas UN petit garçon », la voyelle /e/ dans le déterminant est beaucoup plus longue que la même voyelle réalisée dans l'adjectif « petit ». Ces différences de durée s'observent dans tous les types d'items, que ce soient les voyelles, consonnes, pauses, syllabes ou énoncés. Finalement, l'*intensité* se mesure en décibels (dB) et correspond à l'amplitude du signal sonore. Dans la communication de tous les jours, on identifie l'intensité par le volume de la voix d'un locuteur selon qu'il parle plus ou moins fort.

---

[6] Traduction libre de : « Prosody is the systematic organization of various linguistic units into an utterance or a coherent group of utterances in the process of speech production. Its realization involves both segmental and suprasegmental features of speech, and serves to convey not only linguistic information, but also paralinguistic and non-linguistic information. »

Ces composantes acoustiques fondamentales donnent lieu à deux grands phénomènes prosodiques : le rythme et l'intonation. Ces manifestations ne sont pas associées à une composante acoustique précise, elles sont plutôt le résultat d'une combinaison des trois éléments, c'est-à-dire la F0, la durée et l'intensité, dont l'importance relative peut être différente selon la langue, la variante dialectale ou le registre.

1.3.2 Rythme et accentuation

La définition du *rythme* ne fait pas consensus chez les chercheurs qui s'intéressent à la prosodie. Néanmoins, plusieurs sont d'accord pour dire que les notions qui inspirent le rythme sont celles d'équilibre, de symétrie et de cadence (Lacheret-Dujour et Beaugendre, 1999, chapitre 2, pour une revue). On peut dès lors dégager deux caractéristiques principales de toute description du rythme : la périodicité et la forme.

Nous retiendrons tout de même une définition tirée de Landercy et Renard (1977) :

> Le rythme est une organisation de la parole dans son déroulement temporel. Cette organisation peut se décrire en termes de découpage en unités rythmiques, de succession de syllabes accentuées et inaccentuées de distribution syllabique dans les unités rythmiques, de durée relative des syllabes, de régularité des temps forts, etc.

Le *rythme* est donc l'étude de l'aspect temporel de la prosodie, ce qui comprend, entre autres, l'étude de l'accentuation et des pauses.

L'étude du rythme passe inévitablement par la notion d'*isochronie*, ou l'étude de l'alternance entre les temps forts et les temps faibles. On distingue traditionnellement deux principes rythmiques organisateurs dans les langues : l'isochronie accentuelle et l'isochronie syllabique[7] (Lacheret-Dujour et Beaugendre 1999).

---

[7] Respectivement en anglais, *stress-timed* et *syllable-timed*.

La distinction entre ces deux principes se retrouve au niveau de l'unité fondamentale autour de laquelle s'organise le rythme. Dans le cas de l'isochronie accentuelle (comme dans la langue anglaise), on considère que l'équilibre temporel repose sur la syllabe accentuée puisque les intervalles de temps entre les syllabes accentuées se compriment pour être isochrones, soit de même durée. Inversement, dans l'isochronie syllabique (comme en français), les syllabes non-accentuées seraient d'une durée uniforme. Ainsi, ce sont les syllabes accentuées qui changeraient de durée. Cette typologie a été remise en question par plusieurs auteurs. Citons notamment les travaux de Pasdeloup (1990) qui parlent plutôt d'un principe de progression. En effet, l'auteure constate que les syllabes non-accentuées en français ne sont pas toutes de la même longueur, mais qu'elles sont plutôt en progression : plus on s'approche de la syllabe accentuée et plus elles s'allongent. Ces différents modèles de durée relative peuvent être visualisés dans la figure 1-6.

	← Temps → Chaque case représente des espaces de temps de durées relatives. Les syllabes en caractères gras indiquent les syllabes accentuées.
Isochronie accentuelle Anglais	The **boat'** sails on the **ri'**ver. The | **boat** | sails on the | **ri** | ver En anglais, les intervalles entre deux syllabes accentuées se compriment pour être de même durée.
Isochronie syllabique Français	Le ba**teau'** arpente la ri**vière'**. Le | ba | **teau** | ar | pente | la | ri | **vière** Dans cette conception de l'accentuation en français, les syllabes non-accentuées sont d'une durée fixe et ce sont les syllabes accentuées qui s'allongent.
Principe de progression Français	Le ba**teau'** arpente la ri**vière'**. Le | ba | **teau** | ar | pente | la | ri | **vière** Dans cette conception de l'accentuation en français, les syllabes non-accentuées augmentent graduellement de durée pour atteindre leur point le plus long lors des syllabes accentuées.

Figure 1-6 Représentation graphique des principes d'isochronie[8].

[8] Il s'agit ici d'une représentation schématique ne prétendant pas reproduire les durées à l'échelle.

De pair avec le rythme, mentionnons l'élément essentiel qu'est l'accentuation. L'*accent*, dans sa définition la plus élémentaire, correspond à une syllabe plus proéminente par rapport aux syllabes environnantes. Cette proéminence est subjective puisque ce sont les auditeurs qui la remarquent. Selon le cas, elle peut se manifester par une montée ou une baisse de la F0, par une hausse de l'intensité, par une durée plus longue du segment ou par une combinaison de plusieurs de ces trois facteurs.

Il existe plusieurs types d'accents. Nous adoptons le classement fonctionnel effectué par Di Cristo (2000). On peut voir une représentation graphique de ce classement dans la figure 1-7.

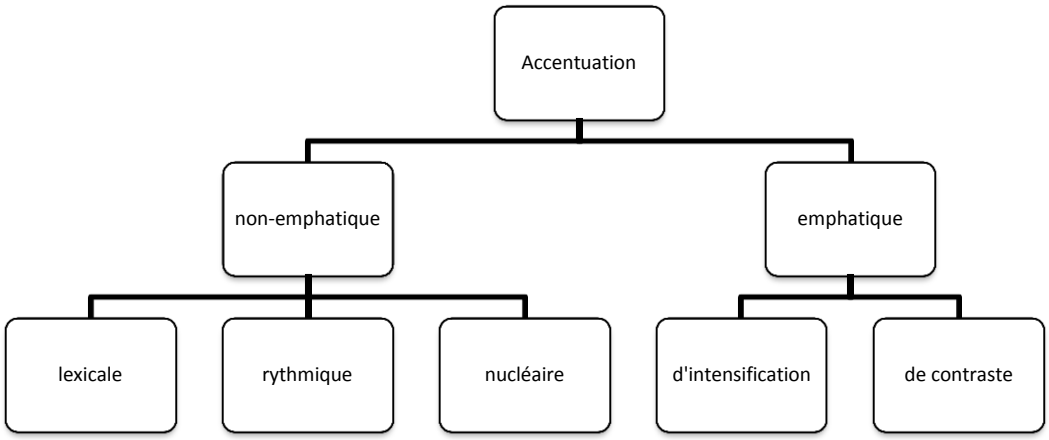


Figure 1-7 Typologie des types d'accents

Comme on peut le voir dans la figure 1-7, l'accentuation est divisée en deux classes : l'accent non-emphatique et emphatique, la première ayant donc des fonctions plus grammaticales, l'autre plus émotives.

La première catégorie d'accent non-emphatique est l'*accent lexical*, absent en français. Il est plus communément appelé accent de mot. C'est l'accent qui sert à

démarquer des mots dont les segments sont identiques. En effet, seule la position de l'accent permet de distinguer ces mots sémantiquement. On le retrouve dans des langues comme l'anglais, le portugais ou l'espagnol.

L'*accent rythmique*, quant à lui, sert à organiser le discours en fonction des règles métriques de la langue. En français, on distingue deux types d'accents rythmiques : l'accent primaire (A1) et l'accent secondaire (A2). L'accent primaire se positionne généralement en fin de mot et permet de démarquer des groupes dans le discours. L'accent secondaire, quant à lui, se positionne dans des mots de plus de trois syllabes. On peut voir une position possible des accents primaires et secondaires dans la phrase suivante :

(1) Le petit carnassier mange la grosse salamandre.

A2 A1 A3 A4

L'*accent nucléaire*, finalement, est le dernier des accents non-emphatique. Il est plus communément appelé accent de phrase et en français on y fait aussi référence en tant qu'accent final. Il désigne la dernière syllabe accentuée de l'unité intonative. Dans la phrase mise en exemple, on voit donc que l'accent nucléaire (A4) pourrait se positionner sur la dernière syllabe de la phrase.

Les *accents emphatiques* (A3), quant à eux, au niveau fonctionnel, servent à véhiculer les attitudes du locuteur. L'*accent d'identification* sert à renforcer, appuyer l'information. On peut en voir un exemple dans une phrase comme « C'est le plus beau que j'ai jamais vu ». L'*accent de contraste*, quant à lui, sert plutôt à mettre deux éléments en relief – cet accent peut aussi être qualifié de mise en relief sémantique. On peut en voir un exemple dans une phrase comme « Ce n'est pas lui, c'est elle ».

1.3.3 Syllabe

La *syllabe* est une unité fondamentale lors de l'analyse des phénomènes suprasegmentaux. Ceux-ci, puisqu'ils ne peuvent se manifester à travers les segments, ont besoin d'une unité porteuse. Cette unité est la syllabe. En effet, ces phénomènes sont considérés comme indépendants de la chaine phonémique depuis l'avènement de la phonologie non linéaire (pour une revue, voir Meynadier, 2001).

La syllabe est, contrairement aux autres unités linguistiques abordées dans ce chapitre, connue de tous les locuteurs d'une langue. En effet, elle est régulièrement utilisée, dès le plus jeune âge, dans toute une série de comptines et petits poèmes. Or, arriver à une définition phonétiquement et phonologiquement valable de la syllabe s'avère scientifiquement ardu. Meynadier (2001) offre un excellent survol des différents problèmes entourant la description et la définition de la syllabe.

Malgré tout, une description fonctionnelle de la syllabe est tout de même largement adoptée (Meynadier, 2001). Selon cette conception, la syllabe est une structure hiérarchique organisée en sous-constituants : l'attaque et la rime. La rime est elle-même organisée en sous-constituants, le noyau et la coda. La figure 1-8 représente graphiquement cette organisation.

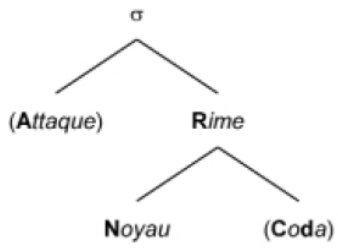


Figure 1-8 Organisation de la syllabe

Selon cette organisation, le noyau de la syllabe est toujours occupé par une composante vocalique. Dans l'exemple d'un mot comme balle, la voyelle [a] occupe la position noyau. La consonne initiale [b], quant à elle, occupe la position d'attaque alors que la consonne finale [l] occupe la position de coda.

Ce mode d'organisation permet de mettre en opposition deux types de syllabes : les syllabes ouvertes et fermées. Une syllabe ouverte est une syllabe où la position coda est vide. Un mot comme « pas » [pa] en est un bon exemple. Par ailleurs, une syllabe fermée est une syllabe où la position coda est remplie, comme dans l'exemple de « balle » [bal].

Par ailleurs, de nombreuses difficultés peuvent être rencontrées lorsqu'il est question de la syllabe, notamment lorsqu'il est question de syllabation. Quelques-unes de ces difficultés seront abordées et expliquées plus en détail dans le chapitre 3.

1.3.4 Intonation

L'*intonation* représente généralement la modulation de F0 dans des domaines d'étendues diverses. On y réfère également par l'expression « contour mélodique ». Bien que l'intonation soit d'abord caractérisée par la modulation de la fréquence fondamentale, les paramètres de durée et d'intensité sont aussi présents dans sa réalisation phonétique.

Une des premières considérations liées à l'intonation repose sur la congruence à la syntaxe. En effet, une étude de l'intonation permet de constater que, dans de nombreux cas, la structure intonative est similaire à la structure syntaxique. C'est ce qu'on observe dans des phrases comme « Mais oui mon cher, réellement » et « Mais oui mon cher Rey, elle ment ». Dans ces cas, on dit que l'intonation remplit des fonctions démarcatives, d'identification de catégories et de hiérarchisation. Bref, elle organise le discours, rend audible la structure syntaxique (Lacheret-Dujour et Beaugendre 1999).

Cependant, dans d'autres cas, force est de constater que la structure intonative n'est pas congruente à la syntaxe. Dans ces cas, on observe plutôt une fonction sémantique à l'intonation, elle permet de structurer le thème de l'énoncé (Lacheret-Dujour et Beaugendre 1999).

Finalement, au carrefour de la syntaxe et de la sémantique, l'intonation permet de distinguer la modalité de la phrase. En effet, en français, un changement dans la courbe mélodique est associé aux différentes modalités. On retient ici quatre configurations intonatives principales, la phrase assertive, la phrase impérative, la question totale et l'interrogation. Chacune de ces modalités est associée à un contour intonatif particulier. La phrase assertive, par exemple, est marquée par un contour mélodique descendant, alors que la question totale est marquée par un contour montant (Di Cristo 1998).

Le modèle de l'intonation française retenu pour ce mémoire est celui de Jun et Fougeron (2000). Il a été retenu puisqu'il emploie la méthode métrique autosegmentale, méthode largement utilisée en synthèse, notamment pour les systèmes disponibles en français retenus dans la méthodologie (Euler (Bagein et coll. 2000)). La figure 1-9 permet de voir une représentation schématique de la phrase prosodique dans ce modèle : le syntagme intonatif (SI).

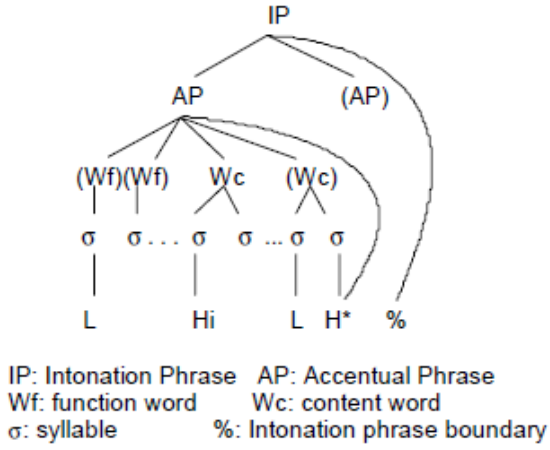


Figure 1-9 Schéma du syntagme intonatif selon Jun et Fougeron (2000, p.214)[9]

[9] En français, IP se lit SI (Syntagme intonatif), AP se lit SA (Syntagme accentuel), Wf se lit Mg (mot grammatical), Wc se lit Mp (mot plein) et *Intonation phrase boundary* se traduit par frontière de syntagme intonatif.

Selon cette représentation, le SI est composé d'un ou de plusieurs syntagmes accentuels (SA). Les SA, quant à eux, sont divisés en syllabes marquées par cinq types de tons : tons hauts (Hi ou H*), tons bas (L) ou tons de frontière de SI (L% ou H%).

1.4 Prosodie en synthèse

Comme il a été vu dans la première section, le travail d'un système de synthèse est de transformer un texte écrit en parole. Ce processus est divisé en de multiples composantes appelées modules. Ces modules fonctionnent indépendamment les uns des autres et transforment l'information petit à petit. Le traitement prosodique est abordé dans deux de ces modules.

Le premier module où on traite de prosodie est celui de l'analyse du texte. Les phrases sont analysées et on leur assigne une série d'étiquettes. Ces étiquettes peuvent être aussi variées que le type de phrase (affirmative, interrogative), la place des accents dans la phrase ou encore la présence d'un accent d'emphase.

Le deuxième module, quant à lui convertit ces étiquettes en valeurs numériques (durée, valeur de F0, etc.). Ce module emploie des modèles mathématiques pour faire ces opérations. Il existe des modèles mathématiquement distincts pour l'intonation et le rythme; Taylor (2009) en offre une excellente revue. Ces modèles ne seront pas expliqués en détail ici. En effet, étant principalement le fruit d'une réflexion mathématique, ils ne seront pas traités dans le cadre de ce mémoire.

Bien évidemment, compte tenu des différences prosodiques importantes qui existent entre les différentes langues du monde, les modèles prosodiques doivent être adaptés afin d'en tenir compte.

1.5 Variation dialectale

Cette étude vise à offrir une contribution linguistique à un système de synthèse de la parole dans un contexte québécois. Il est donc important, afin d'atteindre cet objectif, d'avoir en tête les concepts clés liés à l'étude de la variation dialectale.

Dans un premier temps, il nous apparaît important de définir le terme *dialecte*. Nous retenons la définition suivante, tirée de Brown et Levinson p. 300 (1979).

> Une variété de langues se distinguant des autres variétés par des facteurs traversant la grammaire, comprenant les aspects phonologiques, syntaxiques, lexicaux et prosodiques, ce qui peut être spécifié comme un sous-ensemble distinctif des règles linguistiques de la langue.[10]

On entend par *accent* le sous-ensemble phonétique d'un dialecte. Dans le cadre de ce mémoire, entendu qu'il n'est ici question que de phonétique, les termes dialecte, accent et variété seront utilisés comme des synonymes.

Il est important, dans l'étude des dialectes, de tenir compte de l'ambiguïté existant entre les variations d'origine géographique et les variations d'origine sociale. En effet, il est largement reconnu que les facteurs sociaux comme l'âge, la classe sociale, le niveau d'éducation ou le sexe peuvent venir affecter la façon de parler d'un locuteur. Il est donc important de s'assurer que les locuteurs testés proviennent de groupes sociaux similaires en ce qui a trait au sexe, à l'âge, à l'éducation ou à la classe sociale.

Dans le cas spécifique du FQ, il convient, de plus, de tenir compte des attitudes face aux différentes variétés linguistiques. En effet, il est reconnu que les auditeurs portent des jugements assez forts envers certains traits de leur langue – certains traits étant fortement stigmatisés (Paradis, Brousseau et Dolbec 1993).

---

[10] Traduction libre de : «a variety of a language distinguished from other varieties by features cross-cutting the grammar, including phonological, syntactic, lexical and prosodic features, which can be specified as a distinctive subset of the linguistic rules of a language.»

En socio-phonétique, donc, les recherches visent à atteindre deux objectifs principaux. D'abord, il s'agit de déterminer l'aisance avec laquelle les locuteurs identifient les différents dialectes de leur langue. En deuxième lieu, il s'agit de trouver sur quels critères phonétiques ils se basent pour effectuer cette identification.

Parmi plusieurs facteurs phonétiques permettant d'identifier l'origine géographique d'un locuteur, la prosodie occupe une place particulière. En effet, alors qu'il est relativement aisé d'isoler les autres composantes linguistiques liées au dialecte, la prosodie repose sur le contenu segmental pour se manifester, elle ne peut être prise en isolation. Ainsi, les études sur l'importance de la prosodie reposent sur divers mécanismes de *délexicalisation* ou techniques tentant d'isoler la prosodie du contenu segmental[11].

La première de ces études date des années 1960, au début de la démocratisation des manipulations acoustiques du signal sonore (Bush 1967). Il ressort de cette étude effectuée sur l'anglais (dialectes américains, britanniques et indiens) que les locuteurs peuvent, au-dessus du seuil de la chance, identifier l'origine géographique des locuteurs à l'aide des seuls facteurs prosodiques. Les études subséquentes, utilisant les mêmes procédés de délexicalisation, obtiennent des résultats similaires : il est possible d'identifier le dialecte, mais la tâche est ardue (Peters et coll. 2002).

1.6 Conclusion

Les différents concepts présentés dans ce chapitre visent à éclairer le lecteur dans la compréhension des chapitres qui suivent. Il aura été question, d'abord, de synthèse de la parole. Dans le cadre de ce survol, le lecteur aura pu prendre conscience de l'importance de la prosodie sur deux aspects importants de la synthèse : l'intelligibilité et la qualité. Par la suite, une exploration de l'onde acoustique aura été

[11] Ces techniques seront explorées avec plus d'attention dans le chapitre 3.

effectuée et une définition de la prosodie et de ses composantes principales, le rythme et l'intonation, de même que de ses paramètres acoustiques, la F0, la durée et l'intensité ont été présentés. Finalement, puisque cette recherche se fait dans un cadre québécois, un survol des questions entourant la variation dialectale a été effectué.

CHAPITRE II

REVUE DE LA LITTÉRATURE

Le chapitre précédent a permis de dresser un portrait des différents concepts abordés lors de ce projet. Les notions clés ont été définies, de même que les approches théoriques qui seront employées dans cette étude. Ce chapitre vise plutôt à présenter les études principales qui mènent à la question de recherche. En effet, plusieurs études ont été faites, dans un ou l'autre des domaines présentés plus haut, qui peuvent approfondir notre compréhension du sujet.

2.1 Prosodie et synthèse

2.1.1 Prosodie et intelligibilité

Plusieurs études se sont penchées sur l'impact de la prosodie sur l'intelligibilité. Nous en présentons quelques-unes ici, qui présentent les impacts de la F0 et de la durée des voyelles sur l'intelligibilité dans des conditions idéales et dans le bruit.

La première étude présentée ici est celle de Laures et Weismer (1999). Il s'agit d'une étude préliminaire où les auteurs tentaient de déterminer l'influence d'un contour intonatif plat – ou absence de variation de F0 – sur l'intelligibilité d'une phrase isolée. Ils concluent que ce type de contours entraîne une baisse importante de l'intelligibilité – une tâche de transcription était utilisée – de même qu'une baisse dans la qualité – les auditeurs jugeant négativement les contours plats. Ceci met en

lumière l'importance de F0 dans le décodage des énoncés. En effet, les changements de fréquence fondamentale dans la phrase permettent d'attirer l'attention des auditeurs sur les mots de contenu dans une phrase. Ainsi, toute modification dans le signal de parole visant à réduire les contrastes entre unités adjacentes a le potentiel d'en réduire la saillance et donc, de réduire l'intelligibilité.

La seconde étude que nous avons retenue est celle de Miller, Schlauch et Watson (2010). Elle va plus loin que la première étude puisque les auteurs ont effectué diverses manipulations de F0 qu'ils ont ensuite présentées dans le bruit. En effet, en plus d'avoir créé des contours plats, ils ont aussi exagéré les contours existants, les ont inversés et modifiés sinusodoïdalement à des valeurs de 2.5 et 5 Hz (manipulation qui a déjà été utilisée pour rendre les voyelles plus saillantes). Les auteurs ont trouvé que les contours plats ou exagérés diminuent l'intelligibilité par un facteur d'environ 13% alors que les autres manipulations la diminuent encore plus, par un facteur de 23 %. Il ressort de cette étude que toute manipulation non naturelle de F0 entraîne des pertes d'intelligibilité. Mais, en plus, les manipulations qui sont linguistiquement erronées ont un effet encore plus grave que les manipulations dites linguistiquement neutres.

Une troisième étude s'est penchée sur l'étude des contours plats et a voulu vérifier leur impact dans divers contextes de bruit (Laures et Bunton 2003). Les auteurs ont testé de la parole non modifiée et des contours plats avec divers types de bruit : bruit blanc et bruit de voix. Il ressort de leur étude que les contours plats ont un effet négatif équivalent sur l'intelligibilité, et ce, peu importe le type de bruit.

La dernière étude présentée dans cette section est celle de Spitzer, Liss et Mattys (2007). Ils ont utilisé la méthode de la resynthèse (méthode qui sera abordée dans le chapitre 3) ce qui leur a permis d'étudier, en plus des contours plats de F0, l'effet de l'égalisation de la durée des voyelles sur l'intelligibilité de phrases dans le bruit. Leurs résultats sont présentés dans le tableau 2-1.

Tableau 2-1 Perte d'intelligibilité selon le facteur prosodique manipulé[1]

Condition	Contrôle	F0	Durée	F0 et Durée
Intelligibilité (%)	92,93	59,02	53,46	28,21
Écart type	2,24	5,22	5,28	5,14

On observe d'abord la perte importante d'intelligibilité dès qu'il y a manipulation d'un facteur prosodique. Cette perte est encore plus grave lorsque les deux paramètres prosodiques sont mis ensemble. Cette étude montre clairement l'importance capitale qu'ont la fréquence fondamentale et la durée sur l'intelligibilité d'un système de synthèse.

2.1.2 Prosodie et qualité

Dans cette section, nous présenterons deux études qui se sont penchées sur les facteurs venant influencer le naturel des voix de synthèse. La première a été réalisée par Ratcliff, Coughlin et Lehman (2002). Les auteurs de cette étude se sont concentrés sur les différences de qualité perçues entre les voix de synthèse et la voix humaine. Suite à une première étude où les participants avaient suggéré que le débit, la hauteur de la voix et les pauses étaient ce qu'ils avaient perçu de plus problématique, les auteurs ont choisi de tester l'impact de ces trois éléments. Ils ont donc modifié, pour une phrase perçue comme très naturelle, le débit (plus rapide et plus lent), la F0 moyenne (plus aigüe ou plus grave) et le nombre de pauses. Les auteurs concluent que les énoncés ayant un débit plus rapide et moins de pauses sont mieux jugés que les autres. Les auteurs ne trouvent aucun effet de F0. Il est important de noter que les auteurs n'ont changé que la hauteur moyenne de la voix, sans changer la forme de la courbe de F0. On note tout de même que des facteurs de durée – dans ce cas-ci le débit – ont un impact sur la qualité des voix.

[1] Tableau tiré de Spitzer, Liss et Mattys (2007) p.3663

La seconde étude présentée ici est celle de Paris et coll. (2000) et se penche à la fois sur l'intelligibilité et la qualité. Les auteurs ont testé l'impact de deux facteurs linguistiques sur la parole de synthèse : la prosodie et la prévisibilité sémantique des énoncés. Ils ont effectué deux tâches d'intelligibilité : le MRT et une tâche de rappel immédiat. De plus, les participants devaient effectuer deux évaluations subjectives sur l'intelligibilité perçue et le naturel. Dans le cas de la perte de la prosodie, les chercheurs n'indiquent pas quelle méthode ils ont utilisée, mais nous supposons qu'ils ont dû aplanir la F0 ou égaliser les voyelles.

Les chercheurs observent d'abord que la perte de la prosodie, de même que la perte de contexte sémantique entraînent une diminution de l'intelligibilité de la parole humaine. Or, la même opération effectuée sur de la parole de synthèse n'entraîne pas une perte d'intelligibilité aussi importante, particulièrement dans le cas de la prosodie. Les auteurs concluent que la prosodie, telle qu'implémentée dans les systèmes utilisés, n'offrait déjà pas d'indices linguistiques permettant de favoriser l'intelligibilité.

Au point de vue des jugements subjectifs des auditeurs, la perte de la prosodie a un impact dévastateur sur les résultats. En effet, les résultats sont particulièrement bas pour le naturel lorsque les indices prosodiques sont enlevés. Cette étude montre, encore une fois, toute l'importance qu'il convient d'accorder à la prosodie en synthèse, que ce soit pour améliorer l'intelligibilité ou la qualité.

2.2 Synthèse et dialecte

La création de systèmes de synthèse appropriés repose souvent sur la création de systèmes propres aux différents dialectes d'une langue. En anglais, par exemple, il existe des systèmes distincts pour l'anglais américain et pour l'anglais britannique, entre autres. La création de tels systèmes repose sur une bonne connaissance linguistique des divers phénomènes qui distinguent ces variétés. De plus, il importe d'accorder une attention particulière aux phénomènes qui sont les mieux perçus chez

les auditeurs, autrement dit ceux qui ont le plus d'impact sur l'intelligibilité ou la qualité des systèmes.

Les études se penchant sur les phénomènes linguistiques distinguant différents dialectes d'une même langue sont extrêmement nombreuses et variées. Dans le cadre de cette revue, nous porterons une attention particulière à celles qui évaluent les différences prosodiques.

La première étude que nous présentons a été effectuée par Bush (1967). Il s'agissait alors d'identifier les caractéristiques acoustiques distinguant les anglais américain, britannique et indien. Les énoncés enregistrés ont été par la suite filtrés de manière à conserver différentes informations acoustiques : un filtre passe-bas et un filtre passe-haut. Le filtre passe-bas est une méthode de traitement qui ne conserve que les plus basses fréquences. Il est alors possible de distinguer les variations de F0 sans percevoir le contenu segmental de l'énoncé. L'auteur observe que les participants ayant à identifier l'origine géographique des locuteurs avaient un taux de succès d'environ 70 % dans la condition de filtre passe-bas. Ceci indique que des informations prosodiques liées à la F0 contribuent à l'identification de l'origine géographique.

Une étude plus récente visant à étudier l'influence de l'intonation sur la distinction entre divers dialectes de l'allemand (Peters et coll. 2002) est présentée ici. Les auteurs ont manipulé la F0 dans des phrases segmentalement identiques. La F0 a été manipulée à l'aide d'une resynthèse PSOLA[2]. D'abord, les chercheurs ont confirmé l'hypothèse selon laquelle les locuteurs allemands étaient capables d'identifier l'origine géographique d'un locuteur sur la base de l'intonation uniquement. Les auteurs ont ensuite observé que les performances des participants dans l'identification des origines géographiques variaient selon leur expérience linguistique. En effet, les

[2] PSOLA (Pitch Synchronous Overlap and Add) est une méthode de traitement du signal permettant de modifier la durée et/ou la F0 d'un signal acoustique.

participants qui étaient familiers avec la variété locale en plus d'une autre variété, par contact personnel et non par les médias, performaient mieux que ceux qui n'étaient en contact qu'avec une seule variété. Cette étude indique non seulement que des différences dialectales peuvent être véhiculées à travers la prosodie mais, en plus, elle jette un éclairage particulier sur l'importance de l'expérience linguistique des participants.

La dernière étude présentée dans cette section est celle de (Clopper et Bradlow 2008a). Cette étude se penche sur l'identification des dialectes de l'anglais américain dans diverses conditions de bruit. Elle confirme d'abord que les participants sont capables d'identifier l'origine géographique des locuteurs dans des conditions d'écoute idéales. Autrement dit, les Américains sont assez performants dans la reconnaissance de leurs divers accents. Cependant, dans des conditions de grands bruits, la reconnaissance de plusieurs dialectes devient sévèrement perturbée, et ce, peu importe le dialecte parlé par le participant. En effet, c'est la variété standard (*General American*) qui est la plus reconnue et la plus intelligible dans le bruit par tous les participants. Cette étude est pertinente puisqu'elle suggère que la familiarité avec une variété n'est pas le seul élément qui peut affecter l'intelligibilité, en effet, une variété moins marquée ou standard pourrait faciliter le décodage.

Les études présentées dans cette section font un bon survol des éléments importants dont il faut tenir compte dans l'étude des dialectes en synthèse. Il est important de tenir compte de deux grands facteurs, les caractéristiques phonétiques du dialecte et la capacité des auditeurs de les percevoir de même que leur attitude par rapport à ceux-ci.

2.2.1 De l'importance du FQ

Dans cette section, nous aborderons l'étude de Brousseau (1992) qui s'est penchée sur l'effet de l'utilisation de la variété québécoise sur l'intelligibilité de la parole de synthèse. Il a créé un test de reconnaissance de mots monosyllabiques dans le bruit.

Certains de ces mots comprenaient des traits segmentaux du FQ, tels que l'affrication de /t/ et /d/ ou encore l'ouverture des voyelles hautes /i/ et /y/.

Les résultats qu'il obtient lui permettent de conclure que l'utilisation de la variété québécoise, auprès d'auditeurs l'ayant comme langue maternelle, augmente significativement l'intelligibilité de la parole de synthèse.

2.2.2 Revue des voix de synthèse disponibles au Québec

Dans cette section, nous présenterons une étude de Côté-Giroux et coll. (2011) faisant une revue des voix de synthèse disponibles auprès des Québécois et, surtout, de leur appréciation. Le corpus comprenait à la fois des voix FF et des voix FQ, en plus d'une voix humaine, québécoise.

Les résultats apportent des conclusions intéressantes. Au niveau de l'intelligibilité, ce sont la voix humaine et une des voix de synthèse québécoises qui obtiennent les meilleurs résultats. Or, au plan de l'appréciation, c'est la voix humaine qui a encore le meilleur score, mais ce sont deux voix de synthèse françaises qui devancent la première voix de synthèse québécoise. C'est donc dire que, même si elle est plus intelligible, la voix de synthèse québécoise pose encore problème auprès des auditeurs qui en jugent la qualité.

2.3 Prosodie et français québécois

Dans le cadre de ce mémoire, l'étude de la prosodie en synthèse se fait dans un contexte de français québécois. Ce contexte particulier nous pousse à considérer les différences prosodiques qui existent entre les français québécois et hexagonaux. La plupart des travaux comparatifs ont été réalisés avant 1995 et la section 2.1.1 en offre un survol. Trois études comparatives seront abordées plus en détail, celles de Ménard (1998), de Bissonnette (2000) et de Kaminskaïa (2005). De plus, une étude faite uniquement sur la prosodie du FQ sera présentée, celle de Thibault (1999).

2.3.1 Revue des études réalisées avant 1995

Plusieurs études ont tenté de comparer les différences prosodiques entre les deux dialectes pour un élément prosodique précis. Ménard (1998) expose une dizaine d'études où sont abordés des phénomènes suprasegmentaux différents. Presque toutes concluent que des différences existent sur le plan prosodique, à la perception comme à la production, entre les deux variétés de français.

Les études retenues pour cette synthèse sont difficilement comparables puisqu'elles utilisent des méthodologies très différentes. Les composantes prosodiques à l'étude diffèrent, allant de l'étude d'un seul paramètre acoustique à des études en incluant plusieurs. De plus, l'étendue de l'unité à l'étude n'est pas la même, certains étudiant l'effet de la prosodie sur les voyelles, d'autres sur la phrase. Enfin, le style de parole à l'étude varie grandement aussi, allant de lectures de listes de mots à des corpus de parole spontanée. Le tableau 2-2 permet de comparer les différentes méthodes utilisées dans ces études.

Tableau 2-2 Études comparatives des français québécois et hexagonaux avant 1995[3]

Auteur	**Composante**	**Domaine**	**Nombre de sujets**	**Corpus**
Vinay (1955)	Durée	Voyelle et syllabe	N/D	500 phrases
Gendron (1966)	Durée, Énergie, F0	Voyelle, syllabe et énoncé	16 (Québec), 2 (Paris)	Listes de mots et phrases
Boudreault (1967)	F0, Rythme	Phrase	7 H de Beaupré (banlieue de Québec) 7 H de Paris	Environ 1000 phrases lues, d'une longueur de 2 à 12 syllabes
Artaud & Martin (1968)	Durée	Mot : voyelle et syllabe	5 H canadiens	12 mots trisyllabiques par

[3] Tableau tiré de Ménard (1998) p. 58

	Énergie		5 H français	sujet, tirés d'entrevues de deux émissions radiophoniques : *Gens du Sud* (Québec) et *Impromptus de Paris* (France)
Holder (1968)	Fréquence fondamentale (registre)	Énoncé et voyelle	10 H français 10 H canadiens du sud de l'Ontario	Extraits d'entrevues de deux émissions radiophoniques : *Gens du Sud* (Québec) et *Impromptus de Paris* (France)
Robinson (1968)	Accentuation (place)	Groupe rythmique	5 H français 5 H canadiens du sud de l'Ontario	100 groupes rythmiques tirés d'entrevues radiophoniques
	Durée	Syllabe et groupe rythmique		
Szmidt (1968)	Fréquence fondamentale	Énoncé et finale d'énoncé	3 H d'origine française résidant à Toronto 3 H d'origine canadienne résidant à Toronto	384 phrases interrogatives suscitées
Ouellet (1992)	Durée	Voyelle, consonne et syllabe	1 H (Québec) 1H (France)	300 séquences de type CVC (syllabe entravée)

À partir de ces observations, on remarque d'abord le nombre relativement peu élevé de participants – allant d'aussi peu que deux individus à vingt personnes. Il est aussi intéressant de se pencher sur l'unité étudiée, celle-ci pouvant être aussi petite que la voyelle, allant jusqu'à la phrase entière. De plus, aucune des études n'effectue d'analyse dans toutes les dimensions de l'étendue – de la voyelle jusqu'à la phrase. Finalement, le contexte de production varie aussi. On retrouve d'un côté une production plus spontanée, tirée d'entrevues radiophoniques, et de l'autre, de la parole de laboratoire – de la lecture suscitée. En somme, les différentes méthodologies employées sont très diversifiées.

Malgré tout, les chercheurs ont relevé des différences prosodiques entre les deux variétés. Il est normal, compte tenu des méthodologiques diverses, que des phénomènes distincts soient soulevés et que des conclusions variées soient obtenues. Il ressort de ces études que des différences existent au niveau des trois paramètres acoustiques : durée, intensité et fréquence fondamentale. Le tableau 2-3 permet de visualiser la grande variété de marqueurs observables entre les deux dialectes.

Tableau 2-3 Marqueurs différenciant le FF du FQ [4]

Auteur	**Marqueurs**
Vinay (1955)	- Durée accrue des syllabes accentuées du français du nord de la France, par rapport à FQ; - Voyelles inaccentuées plus longues en FQ, particulièrement en position prétonique.
Gendron (1966)	- Durée similaire, en FQ et en FF, des voyelles accentuées en syllabes ouvertes et fermées; - Voyelles inaccentuées, formes dérivées ou non, plus longues en FQ qu'en FF; - Patrons de durée distincts : Disyllabes FQ : _ _ FF : Y _ Trisyllabes FQ : _ Y _ FF : Y Y _ Quadrisyllabes FQ : _ Y Y _ FF : Y Y Y _ - Énergie articulatoire plus importante en FQ qu'en FF, pour les voyelles, moins importante pour les consonnes; - Frontières syllabiques moins nettes en FQ; - Impression de débit plus lent en FQ; - Étendue de fréquence et variation tonale (intravocalique et intervocalique) plus importante en FF qu'en FQ.
Boudreault (1967)	- Durée importante des syllabes prétoniques en FQ; - Importance des durées vocalique et consonantique inaccentuées en FQ par rapport au FF; - Variation interindividuelle plus importante pour les sujets français, en ce qui concerne les variations mélodiques; - Variation intraindividuelle plus importante pour les sujets canadiens.

[4] Tableau tiré de Ménard (1998) p. 59

Artaud & Martin (1968)	- Voyelles souvent plus longues en FF qu'en FQ; - Énergie vocalique plus importante en FQ qu'en FF; - Répartition différente de l'énergie articulatoire en FQ et FF; - Énergie accrue des syllabes finale et initiale d'énoncés en FQ, par rapport à la syllabe finale en FF; - Augmentation de la durée de la première à la troisième syllabe en FF, durée de la deuxième syllabe légèrement plus importante que la première syllabe en FQ; - Valeur d'énergie et de durée constantes dans les deux premières syllabes, en FQ; - En FF, valeur d'énergie plus grande sur la première syllabe et durée plus importante sur la seconde syllabe.
Holder (1968)	- Registre (étendue de fréquence fondamentale) plus étendu en FF qu'en FQ
Robinson (1968)	- Patron stable des groupes rythmiques en FF : une ou plusieurs syllabes inaccentuées suivies d'une syllabe accentuée beaucoup plus longue que les autres syllabes; - Aucun patron stable de groupes rythmiques en FQ; - Groupes de deux à six syllabes : en FF, durée de la syllabe finale beaucoup plus importante que les autres syllabes, alors qu'en FQ, durée équivalente des syllabes selon les positions.
Szmidt (1968)	- Dans les phrases interrogatives à inversion du sujet, montée de l'intonation à l'inversion plus fréquent en FF qu'en FQ
Ouellet (1992)	- Distinctions de la répartition temporelle dans les syllabes, en FQ et en FF : attraction de la première consonne et de la voyelle, en FF, tandis qu'en FQ, la voyelle et la consonne entravante montrent un maximum de cohésion.

De nombreuses informations peuvent être tirées de ces études. On remarque d'abord que la plupart des auteurs notent des différences au niveau de la durée (6 études) plus qu'aux niveaux de la F0 (3 études) ou de l'intensité (2 études).

Au niveau de la durée, les auteurs notent, par exemple, que les syllabes non-accentuées sont plus longues en FQ qu'en FF. Certains notent aussi que les voyelles sont plus longues en FF qu'en FQ. Au niveau de la F0, l'observation la plus populaire est celle voulant que l'étendue du registre soit plus grande en FF. Finalement, au niveau de l'intensité, certains auteurs notent une intensité plus grande en FQ, pour les voyelles ou en syllabe finale.

En somme, il est possible d'affirmer que les études comparatives effectuées avant 1995 notent principalement des différences au niveau de la durée, et que les différences de fréquence fondamentale se limitent à l'observation selon laquelle le registre est plus étendu en FF.

2.3.2 Études comparatives entre le FF et le FQ

Depuis 1995, quelques études comparatives entre la prosodie du FF et du FQ ont été réalisées. La plus importante est sans aucun doute celle de Ménard (1998). Elle sera présentée ici en détail, tout comme les études subséquentes de Bissonnette (2000) et de Kaminskaïa (2004).

2.3.2.1 Comparaison entre FF et FQ

L'étude de Ménard (1998) est une des plus importantes lorsqu'il est question de comparaison entre le FF et le FQ. En effet, l'originalité de l'étude aura été de centrer la recherche sur l'auditeur plutôt que sur le locuteur. Il est intéressant de connaître les différences existant à la production, mais il l'est encore plus de savoir quelles sont les particularités utilisées par les auditeurs lorsque vient le moment d'identifier l'origine d'un locuteur.

Ménard a d'abord cherché à savoir si les auditeurs québécois pouvaient identifier l'origine géographique d'un locuteur uniquement à partir de la prosodie. Pour ce faire, elle a utilisé une méthode de délexicalisation par filtrage[5]. Cette technique permet de dépouiller une production de tout contenu segmental en ne conservant que les plus basses fréquences, ou l'intonation. Cela donne l'effet d'entendre quelqu'un à travers un mur, ou encore d'entendre de la parole alors qu'on a la tête sous l'eau.

Elle a donc délexicalisé plusieurs énoncés tirés d'un corpus de bulletins de nouvelles québécois et français. Un groupe de québécois a ensuite tenté d'identifier l'origine

[5] Cette méthode sera expliquée plus en détail dans le chapitre 3.

des locuteurs à partir de ces énoncés dépourvus de contenu segmental. Cette démarche aura montré que des auditeurs québécois peuvent reconnaître, à partir des seuls indices prosodiques, l'origine géographique des locuteurs.

Elle va plus loin dans l'analyse en tentant de découvrir ce qui permet aux phrases fortement identifiées comme québécoise ou française de se démarquer. Elle conclut que seul le registre (l'étendue de F0) permet de servir de véritable marqueur[6] prosodique du français québécois.

2.3.2.2 Questions de registre

L'étude de Bissonnette (2000) visait, suite à l'étude de Ménard (1998), à explorer plus en détail les différences qui existent, à la production, entre les registres québécois et français. Elle emploie un corpus d'énoncés produits par des lecteurs de nouvelles provenant des deux dialectes.

Elle effectue une analyse exhaustive du registre. D'abord, elle utilise deux unités d'étude, l'Hertz (Hz) qui est l'unité acoustique et le demi-ton qui est l'unité perceptive. De plus, elle élargit son étude pour couvrir plusieurs domaines d'étendue : le discours, l'énoncé et le syntagme intonatif. Finalement, bien que le registre soit une manifestation de F0, elle multiplie les paramètres acoustiques à l'étude : la hauteur simple, l'étendue, l'écart type et la distribution des fréquences.

Elle conclut que les Québécois utilisent un registre plus bas et plus étendu que les Français, et ce, peu importe l'unité utilisée ou le domaine exploré. Cette conclusion vient à l'encontre de celle des études d'avant 1995 qui avaient observé un registre plus étendu chez les Français, plutôt que chez les Québécois.

---

[6] Dans ce cas précis, Ménard définit un *marqueur* comme étant une particularité linguistique permettant aux auditeurs de faire un jugement quant à l'origine géographique du locuteur.

2.3.2.3 Questions d'intonation

L'étude de Kaminskaïa (2004) cherchait à mettre en évidence les disparités existant, au niveau intonatif, entre les locuteurs québécois et français. En effet, puisque des particularités au niveau du registre ont été trouvées par Ménard (1998) et Bissonnette (2000), on peut déduire que c'est parce qu'elles mettent en lumière des réalisations différentes de la grammaire tonale.

L'auteure emploie le texte lu comme corpus. L'hypothèse de départ avancée est que les grammaires tonales sont identiques entre les deux variétés de français. C'est le contour intonatif selon le modèle prosodique de Jun et Fougeron (utilisation de tons, hauts ou bas) qui est à la base de l'étude. L'analyse porte principalement sur les différences de réalisation entre les deux dialectes.

De nombreuses occurrences de celles-ci sont identifiées, particulièrement lors du passage d'un ton à l'autre. En effet, ces passages sont plus abrupts chez les Québécois, particulièrement lors du passage d'un SA à un autre. Ce phénomène produit une intonation plus modulée chez les Québécois. Les tests statistiques effectués démontrent en outre que ces écarts sont très significatifs et ne peuvent être expliqués que par le facteur du dialecte.

2.3.3 Études sur la prosodie du FQ

L'étude présentée dans cette section ne cherche pas à comparer les systèmes prosodiques français et québécois mais plutôt à documenter la prosodie de ce dernier dialecte. Il nous apparaît pertinent de la présenter ici.

2.3.3.1 Texte lu et parole spontanée

L'étude de Thibault (1999) se penche sur les différences prosodiques existant entre les français québécois lus et spontanés. Elle démarre son étude avec l'objectif clair d'obtenir plus de données sur la prosodie du FQ dans le but d'améliorer les systèmes de synthèse de la parole. L'auteure effectue des enregistrements de plusieurs styles de

parole (lecture, monologue de direction, conversation spontanée). Ensuite, elle analyse phonétiquement les énoncés en les découpant en syntagmes intonatifs (SI) et en groupes rythmiques (GR). Les paramètres qu'elle étudie sont le débit, la longueur en syllabe des SI et des GR et la durée des SI, le registre et le nombre de proéminences par SI. Finalement, elle effectue des tests perceptifs afin de déterminer la réalité perceptuelle de ces paramètres.

Ces analyses lui permettent de conclure que les paramètres phonétiques les plus pertinents dans la distinction entre parole spontanée et parole lue sont le registre, plus bas et plus étroit en parole spontanée. Elle note aussi que le débit est un facteur important dans la distinction entre le monologue d'instruction et la conversation (le débit était plus bas dans le premier cas). Les analyses perceptives montrent que les auditeurs ne sont en mesure de bien identifier que deux styles : le monologue d'instruction et la conversation. Finalement, une analyse tonale révèle qu'un nombre limité de patrons intonatifs sont employés et que les tons bas sont plus typiques de la parole spontanée.

2.4 Question de recherche et objectifs

La problématique présentée jusqu'ici nous amène poser les questions de recherche. En effet, sachant que des différences prosodiques sont perceptibles entre les français de France et du Québec et que l'utilisation de la variété de langue régionale améliore l'intelligibilité et la qualité d'un système de synthèse, est-ce que l'emploi de la prosodie du français québécois dans un système de synthèse peut en améliorer l'appréciation et l'intelligibilité ? De plus, nous souhaitons atteindre quelques objectifs, présentés ci-dessous.

2.4.1 Premier objectif : Identifier certaines caractéristiques prosodiques principales du français québécois, dans sa perception et sa production

Cet objectif est le premier pas vers l'implémentation d'une synthèse québécoise adéquate. En effet, comme le lecteur l'aura remarqué dans ce chapitre, même si de nombreuses études ont été faites sur la prosodie du FQ, les résultats sont plutôt disparates et, parfois même, se contredisent. Il sera fort important, dans le cadre de ce mémoire, de bien caractériser la prosodie du FQ.

Les études antérieures sur le sujet ont utilisé deux grandes approches : une étude de la production et une étude centrée sur l'auditeur. C'est pourquoi nous jugeons important de combiner ces deux approches dans notre analyse afin, nous l'espérons, de trouver des résultats permettant de faire une synthèse des études antérieures. Nous formulons, de plus, trois hypothèses liées à cet objectif.

2.4.1.1 Hypothèse 1 : Des différences prosodiques seront observées, à la perception comme à la production entre le FF et le FQ.

Puisque la quasi-totalité des études présentées dans ce chapitre démontrent des différences prosodiques entre le FF et le FQ, il est prévisible que nous trouvions nous aussi des disparités entre ces deux dialectes, et ce, à la fois à la production et à la perception.

2.4.1.2 Hypothèse 2 : Les différences prosodiques observées à la **production** se trouveront principalement au niveau de la **durée.**

Les études faites avant 1995 (section 2.1.1) ont comme point commun qu'elles évaluent les différences entre les deux dialectes à l'aide de la production uniquement, sans tenir compte de l'aspect perceptif. Ces études isolent de nombreuses différences entre les deux dialectes, mais le paramètre acoustique de la durée est présent dans la plupart d'entre elles. Nous faisons donc l'hypothèse qu'au niveau de la production uniquement, nous observerons, nous aussi, des différences de durée entre les deux dialectes.

2.4.1.3 Hypothèse 3 : Les différences prosodiques observées à la **perception** se trouveront principalement au niveau de la **F0.**

Les études subséquentes adoptent une approche centrée sur l'auditeur où l'aspect perceptif revêt beaucoup d'importance. Il s'agissait alors de déterminer les éléments acoustiques utilisés pour découvrir l'origine géographique d'un locuteur. Or, ces études mettent l'accent sur l'aspect prédominant de la F0, dans le cadre du registre, notamment. Il nous est permis de croire que nous observerons nous aussi des différences de F0 dans l'aspect perceptif de notre étude.

2.4.2 Deuxième objectif : Dans quelle mesure la prosodie du FQ affecte-t-elle l'intelligibilité ou la qualité d'une synthèse ?

Comme nous avons pu l'observer dans ce chapitre, la prosodie a un impact clair à la fois sur la qualité et l'intelligibilité des systèmes de synthèse. De plus, la familiarité avec la variété de langue utilisée augmente elle aussi la qualité des systèmes. Finalement, les réalisations prosodiques linguistiquement inappropriées ont un impact négatif à la fois sur l'intelligibilité et la qualité des systèmes. Il s'agit ici de déterminer quel impact aura le cas spécifique de l'utilisation de la prosodie du FQ en synthèse.

2.4.2.1 Hypothèse 4 : Un modèle FF de la prosodie (F0 et durée) sur une production segmentale FQ entraîne une baisse d'intelligibilité et de qualité.

Nous formulons l'hypothèse qu'un modèle inapproprié de la prosodie, soit l'utilisation d'un modèle FF sur une production segmentale FQ, entraînera une baisse d'intelligibilité et de qualité.

2.4.3 Troisième objectif : Quels paramètres prosodiques ou combinaison de paramètres sont les plus pertinents dans l'amélioration d'une synthèse de la parole en FQ ?

Tel que le lecteur aura pu l'observer dans ce chapitre, certains paramètres prosodiques ont plus d'impact que d'autres. Il s'agit ici de déterminer quels sont les paramètres qui ont le plus d'impact sur une synthèse dans les cas spécifiques du FQ.

2.4.3.1 Hypothèse 5 : D'une manière générale, c'est la F0 qui aura le plus d'impact, à la fois sur l'intelligibilité et sur la qualité.

Dans la plupart des études présentées dans ce chapitre, c'est la F0 et non pas la durée qui a généralement le plus d'impact. De plus, dans les cas spécifiques du FQ, les études de Ménard et Bissonnette semblent démontrer que la F0 est le paramètre le plus perçu par les auditeurs. Il semble donc réaliste qu'il s'agira du paramètre ayant le plus d'impact.

CHAPITRE III

MÉTHODOLOGIE

Afin de répondre aux objectifs de cette recherche, une méthodologie en trois temps a été développée. Dans ce chapitre, un survol du protocole de test sera d'abord effectué. Par la suite, chacune des trois étapes méthodologiques sera vue en détail et chaque choix sera justifié. Finalement, les objectifs seront revus afin d'expliquer comment cette méthodologie permettra d'y répondre.

3.1 Survol du protocole de tests

Les diverses étapes méthodologiques sont illustrées dans la figure 3-1. Elles seront décrites plus en détail dans les sections suivantes. Cette section vise plutôt à présenter brièvement les diverses étapes et les liens qui existent entre elles. De plus, nous aborderons la terminologie spécifique retenue pour cette étude. En effet, vu le nombre d'étapes, les différents énoncés utilisés de même que les divers participants se verront assigner des noms différents.

La première étape vise à assembler un corpus comparable d'énoncés produits en FF et en FQ. Il s'agit en fait d'enregistrer des français (FF) et des québécois (FQ) et de leur faire lire le même corpus de phrases. Les sujets se verront attribuer le nom de *locuteurs originaux* et les énoncés retenus correspondent au *corpus initial*.

La deuxième étape a pour objectif d'identifier, parmi tous les énoncés produits dans le corpus, ceux qui sont le plus facilement reconnus comme étant FF ou FQ, et ce, par le biais de la prosodie uniquement. Dans cet objectif, plusieurs méthodes de délexicalisation ont été évaluées et la resynthèse a été retenue. Les phrases ainsi délexicalisées ont été jugées lors d'un accord interjuges – les participants à ce test portent le nom de *juges*. Les phrases les mieux identifiées ont été conservées pour la suite de l'analyse – il s'agit des *archétypes*.

La troisième et dernière étape cherche à implémenter ces données dans un système de synthèse. Il s'agit alors d'identifier les corrélats prosodiques permettant de déterminer l'origine des locuteurs et de les implémenter dans le système de synthèse Mbrola[1]. Une batterie de tests statistiques a d'abord été utilisée afin de déterminer quels énoncés parmi les archétypes pouvaient être identifiés comme étant les *prototypes* FF et FQ. De plus, ces tests ont permis de dégager les paramètres prosodiques les plus pertinents dans l'identification de l'origine géographique. Par la suite, ces prototypes et paramètres prosodiques seront implémentés à l'aide du système de synthèse Mbrola. Finalement, un test de perception sera utilisé pour mesurer l'efficacité de ces changements. Ce test comprend une tâche d'intelligibilité, une tâche d'identification de l'origine géographique des locuteurs et une tâche d'appréciation. Les sujets de ce test sont appelés les *participants*.

[1] Ce logiciel sera présenté plus en détail dans la section 3.4.3.3

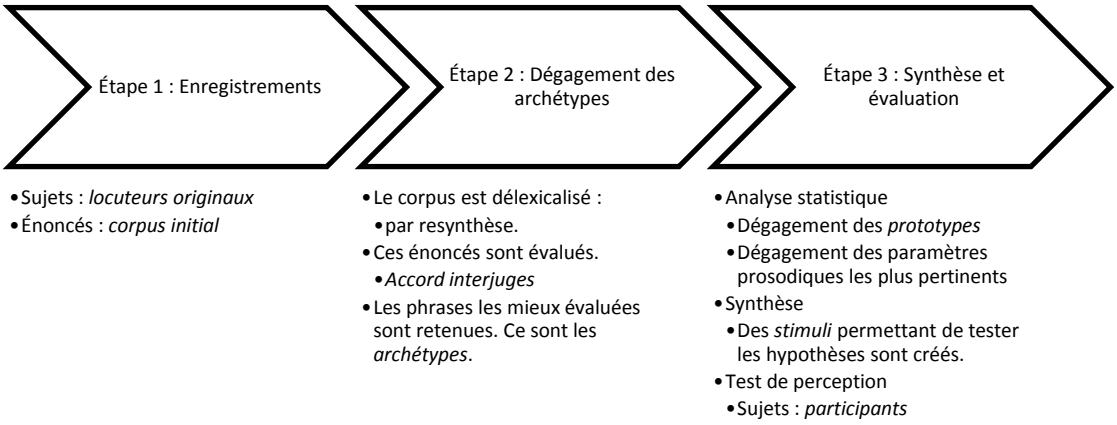


Figure 3-1 Survol du protocole de tests

3.2 Étape 1 : Enregistrements

La toute première étape de cette étude aura été d'enregistrer des locuteurs français et québécois d'une manière contrôlée.

3.2.1 Corpus

Un corpus original a été développé pour cette étude. Il est disponible dans le tableau 3-1. Ce corpus, en plus de varier les structures syntaxiques, intègre deux éléments prosodiques régulièrement employés en synthèse, l'interrogation et l'emphase. Le corpus est constitué de 38 phrases comprenant de 3 à 16 syllabes. De plus, une série de phrases comprenant des noms de trois syllabes ou plus a été créée afin de susciter des accents secondaires – variant ainsi les réalisations prosodiques.

Tableau 3-1 Corpus initial

Mots de moins de trois syllabes		
Neutres	**Interrogatives**	**Emphase**
- Le chat mange la souris. - Il la mange. - Le petit chat mange la grosse souris. - Le chat qui se cache derrière la table mange la souris. - Le chat mange la souris qui se cache derrière la table.	- Le chat mange la souris ? - Il la mange ? - Le petit chat mange la grosse souris ? - Le chat qui se cache derrière la table mange la souris ? - Le chat mange la souris qui se cache derrière la table ? - Qui mange la souris ? - Le chat mange quoi ? - Qu'est-ce que le chat mange ? - Où est le chat qui mange la souris ? - Où est la souris que le chat mange ?	- C'est le chat qui mange la souris, pas le chien. - Le chat mange la souris, pas le fromage. - Le chat MANGE la souris. - Le chat mange une GROSSE souris.
Mots de plus de trois syllabes		
Neutres	**Interrogatives**	**Emphase**
- Le carnassier mange la salamandre. - Il la mange. - Le petit carnassier mange la grosse salamandre. - Le carnassier qui se cache derrière la table mange la salamandre. - Le carnassier mange la salamandre qui se cache derrière la table.	- Le carnassier mange la salamandre ? - Il la mange ? - Le petit carnassier mange la grosse salamandre? - Le carnassier qui se cache derrière la table mange la salamandre ? - Le carnassier mange la salamandre qui se cache derrière la table ? - Qui mange la salamandre ?	- C'est le carnassier qui mange la salamandre, pas le chien. - Le carnassier mange la salamandre, pas le fromage. - Le carnassier MANGE la salamandre. - Le carnassier mange une GROSSE salamandre.

	- Le carnassier mange quoi ? - Qu'est-ce que le carnassier mange ? - Où est le carnassier qui mange la salamandre ? - Où est la salamandre que le carnassier mange ?

3.2.2 Choix des sujets

Les locuteurs recrutés sont au nombre de six : trois locuteurs natifs du FF et trois locuteurs natifs du FQ. Tous étaient de sexe masculin, âgés entre 23 et 28 ans et étudiants universitaires. Les locuteurs du FQ étaient originaires de la région de Montréal et y avaient passé la majorité de leur vie. Les locuteurs du FF étaient originaires de la région parisienne ou du nord de la France et étaient en échange au Québec depuis moins d'un an – la durée moyenne étant de sept mois. Ces données peuvent être visualisées dans le tableau 3-2.

Tableau 3-2 Profil sociodémographique des locuteurs

	ÂGE	ORIGINE	DUREE DE SEJOUR
Q1	23	Montréal	
Q2	24	Laval	
Q3	27	Longueuil	
Moyenne	24.6		
F1	25	Paris	7 mois
F2	23	Tours	7 mois
F3	28	Versailles	8 mois
Moyenne	25.3		7.3

3.2.3 Conditions d'enregistrement

Les enregistrements ont tous été réalisés dans la chambre sourde du laboratoire de phonétique de l'UQAM. Les données ont été captées à l'aide d'une enregistreuse Olympus LS-10S. Dès leur arrivée, les participants ont lu et signé le formulaire de consentement (appendice A). L'expérimentatrice était présente s'ils avaient des questions. Il leur a également été mentionné qu'ils pouvaient abandonner l'étude à n'importe quel moment. Ils ont été rémunérés à ce moment. Ils pouvaient donc quitter

s'ils en ressentaient le besoin, et ce, sans préjudice. Par la suite, les locuteurs ont eu à remplir un formulaire sociodémographique, visant à vérifier les lieux où ils ont vécu, leur âge et la langue maternelle de leurs parents. Finalement, ils ont tous effectué un dépistage auditif[2] visant à vérifier que leur audition était normale.

Ils ont eu ensuite à lire le corpus de phrases. Chaque phrase devait être lue deux fois en présence de l'expérimentatrice – la même pour tous les locuteurs. Ceux-ci avaient pour consigne de parler le plus naturellement possible tout en produisant les phénomènes à l'étude (interrogation, emphase). Ils étaient avertis à l'avance qu'ils pouvaient avoir à répéter si l'expérimentatrice le leur demandait. Des répétitions ont été demandées si les locuteurs hésitaient, avaient des bris de voix (éternuement, toux, raclement de gorge, voix craquante) ou s'ils riaient durant leur production.

3.3 Étape 2 : Dégagement des archétypes

Suite à ces enregistrements, il est apparu important de dégager les énoncés qui étaient les mieux perçus comme étant FF ou FQ, et ce, sur la base de la prosodie uniquement.

3.3.1 Découpage syllabique avec Praat

La première étape de ce traitement a consisté en une analyse préliminaire et un découpage syllabique à l'aide du logiciel de traitement phonétique Praat (Boersma et Weenink 2012).

[2] Ce dépistage auditif a été effectué selon le protocole établi au laboratoire de phonétique de l'UQAM.

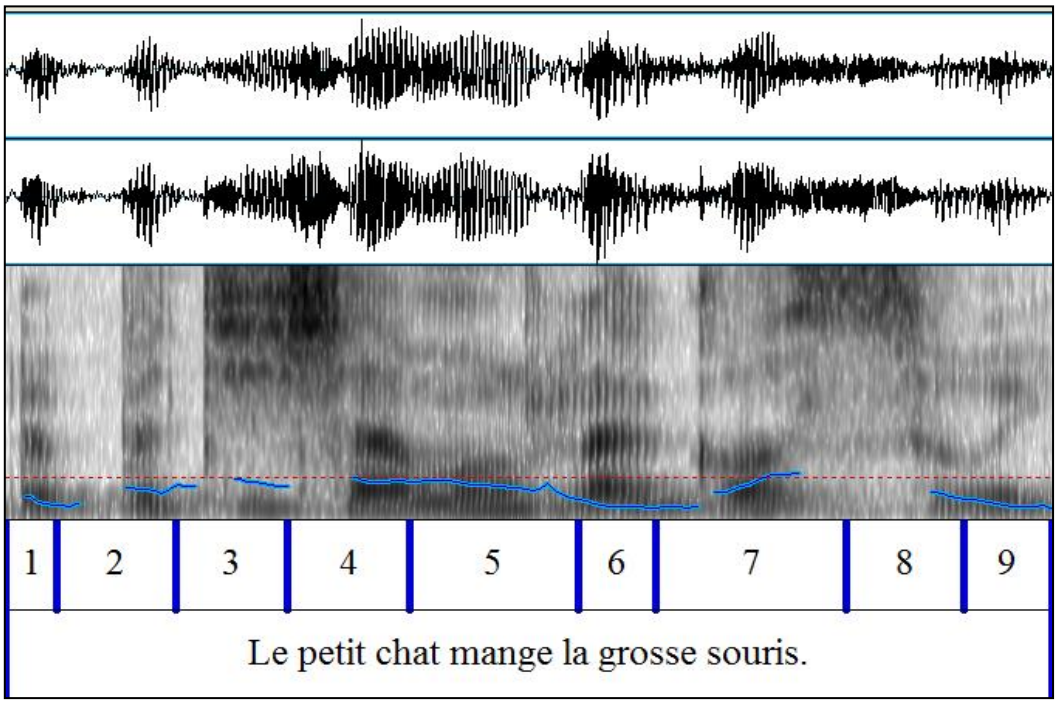


Figure 3-2 Syllabation à l'aide de Praat

3.3.2 Délexicalisation

Suite à cette analyse préliminaire vient l'étape de la *délexicalisation*. Cette étape est une des étapes les plus délicates du processus. Il s'agit en fait de tenter de retirer le contenu segmental des énoncés afin de ne conserver que les aspects prosodiques : durée, F0 et intensité.

Plusieurs méthodes de délexicalisation ont été utilisées dans la recherche en prosodie et nous les présenterons ici, de même que la méthode qui a été retenue pour l'étude.

3.3.2.1 Délexicalisation par filtrage

La *délexicalisation par filtrage* est l'une des plus employées (voir notamment Ménard (1998), Trofimovich et Baker (2006), Huang et Jun (2011)). Il s'agit d'utiliser un filtre passe-bas afin de ne conserver que les basses fréquences – celles porteuses de la fréquence fondamentale. Les fréquences plus hautes sont celles permettant de véhiculer les autres informations, notamment les formants, autrement

dit le contenu segmental. Cette méthode a l'avantage d'être très simple à réaliser et de permettre une étude très juste de la F0. Cependant, la perte des informations segmentales entraîne aussi la perte des informations liées à la durée. En effet, puisque les frontières entre segments sont floues, il n'est plus possible pour l'auditeur de percevoir adéquatement des données comme les longueurs relatives des voyelles ou des syllabes.

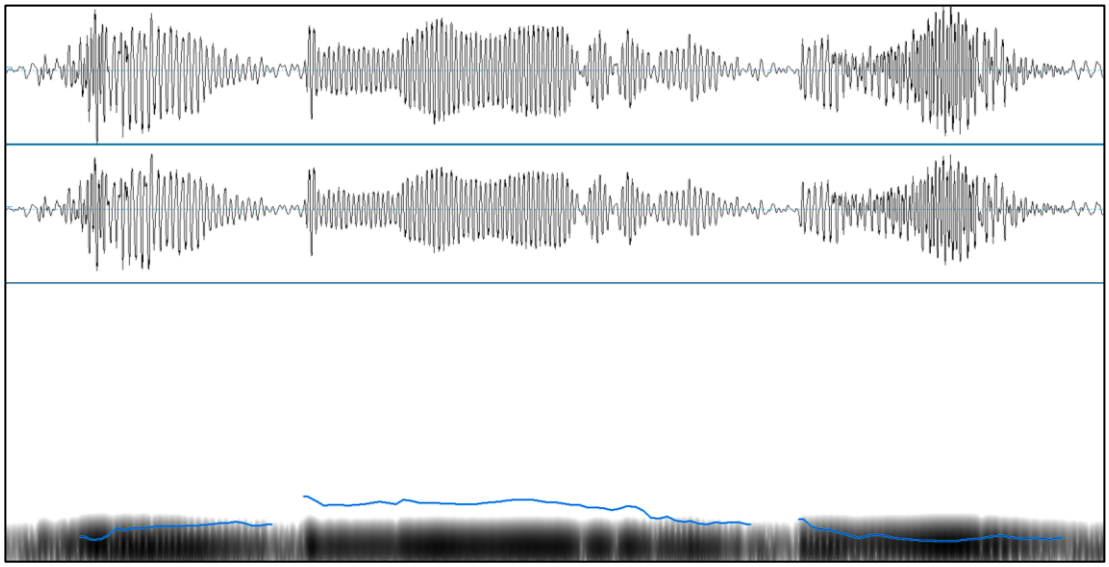


Figure 3-3 Phrase délexicalisée par filtrage (filtre passe-bas)

3.3.2.2 Délexicalisation par pulse train

La *délexicalisation par pulse train* a été utilisée notamment par Ohala et Gilbert (1981). Il s'agit de convertir la courbe prosodique en une série de points de fréquence simples. Par exemple, si la F0 atteint à un moment 150 Hz, le *pulse train* générera une onde sonore de 150 Hz. Cette opération a pour but de simplifier l'onde sonore en enlevant la complexité amenée par le conduit vocal. On perd ainsi toutes les informations qui pourraient permettre d'identifier l'individu ou son sexe. Il ne reste alors que la courbe de F0, dans sa forme la plus simple. Malgré ces caractéristiques, cette méthode a, comme la méthode précédente, le désavantage d'entraîner la perte des informations liées à la durée des divers segments.

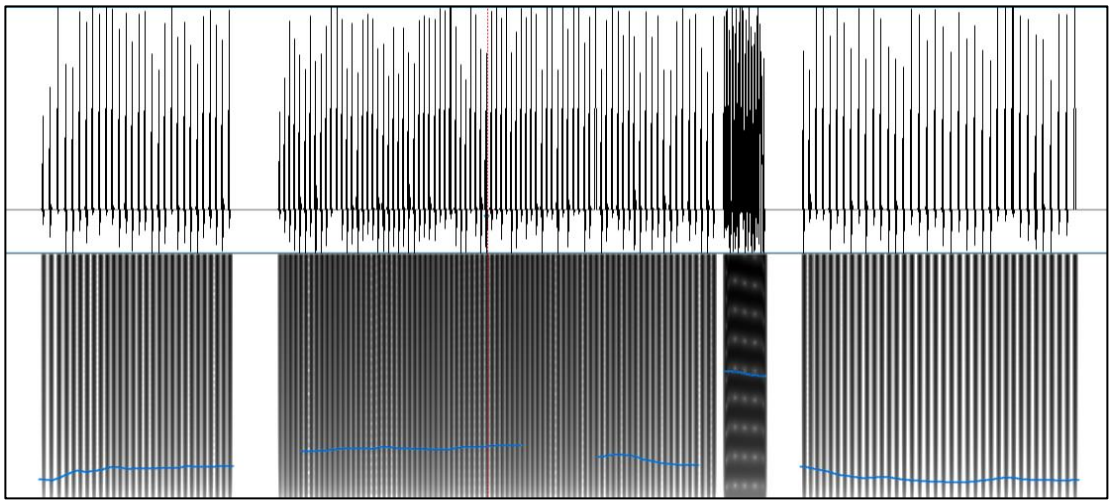


Figure 3-4 Phrase délexicalisée par *pulse train*

3.3.2.3 Délexicalisation par contrôle segmental

Avec la méthode de la *délexicalisation par contrôle segmental*, il s'agit de contrôler finement le contenu segmental des phrases à produire. Autrement dit, les auteurs utilisant cette méthode écrivent à l'avance des phrases qu'ils savent ne pas varier segmentalement entre les deux dialectes. Puis, ils font lire ces phrases à des locuteurs originaires de différentes régions. Ces phrases présentent alors des différences dialectales uniquement au niveau de la prosodie.

Cette méthode a l'avantage de conserver toutes les données prosodiques produites. Cependant, elle ne peut être utilisée pour tous les dialectes. En effet, dans certains cas, les différences phonologiques sont si importantes qu'il est très difficile voire impossible de créer des phrases qui seraient segmentalement identiques entre les deux variétés. Ce serait par exemple impossible à réaliser entre les anglais américain et britannique. Ce serait aussi très ardu voire impossible entre le FF et le FQ. Elle a néanmoins été utilisée dans le cadre de l'italien (Petrone 2010) et de l'allemand (Peters et coll. 2002).

3.3.2.4 Délexicalisation par resynthèse

La dernière méthode présentée dans cette revue, la *délexicalisation par resynthèse*, est de plus en plus utilisée (voir notamment Spitzer, Liss et Mattys (2007) pour une revue). En effet, la démocratisation des outils informatiques de synthèse et de traitement du signal la rend de plus en plus accessible aux chercheurs. Il s'agit d'enregistrer les participants puis d'extraire les données prosodiques qu'ils ont produites – durée de chaque segment, courbe de F0, courbe d'intensité – puis d'appliquer ces données sur une phrase porteuse. Cette phrase porteuse peut être une phrase sémantiquement appropriée ou être composée de mots sans liens entre eux ou encore n'être composée que de syllabes vides de contenu sémantique, comme des *lala*, au choix du chercheur.

Cette méthode est celle qui a été retenue pour cette étude. En effet, bien qu'elle demande un volume de travail considérable, elle permet de conserver une donnée que nous jugeons essentielle : la durée.

Un module de délexicalisation a été créé informatiquement pour cette étude à l'aide du logiciel Praat. Le script peut être consulté en appendice H. Une méthode originale a dû être créée. En effet, nous n'avons pas trouvé de logiciel permettant d'effectuer cette manipulation en français. Nous avons donc créé cette méthode de toutes pièces.

D'abord, nous avons choisi, pour la phrase porteuse, d'utiliser la répétition d'une syllabe simple, *la*, qui se répète et s'adapte à la phrase à délexicaliser. Deux types de syllabes porteuses ont été créés et testés. Dans un premier temps, nous avons enregistré une voix humaine masculine québécoise. Cet enregistrement n'a pas donné les résultats attendus. En effet, n'ayant pas traité les limites de l'enregistrement, la superposition de plusieurs syllabes entraînait un effet auditif hachuré – dû au fait que l'onde acoustique ne se rejoignait pas exactement aux limites.

Après plusieurs tentatives infructueuses afin de corriger ce problème, une autre avenue a été suivie : l'utilisation d'une voix de synthèse. En effet, les voix de synthèse sont enregistrées avec des locuteurs professionnels et sont ensuite traitées

par des ingénieurs en traitement du signal précisément pour faciliter les joints entre les différents segments. L'utilisation de la syllabe *la* générée par Mbrola a permis la création de stimuli d'une qualité suffisante.

Par la suite, les données prosodiques extraites des énoncés du corpus ont été resynthétisées sur les phrases créées par des suites de la syllabes *la*. Plus précisément, la durée de chaque syllabe a été ajustée, les pauses ont été intégrées, la courbe de F0 a été accolée à la phrase de même que la courbe d'intensité. La figure suivante montre le spectrogramme d'une phrase de base suivie du spectrogramme de la phrase délexicalisée. Le lecteur remarquera que la F0 est reproduite fidèlement, de même que la durée de chaque syllabe.

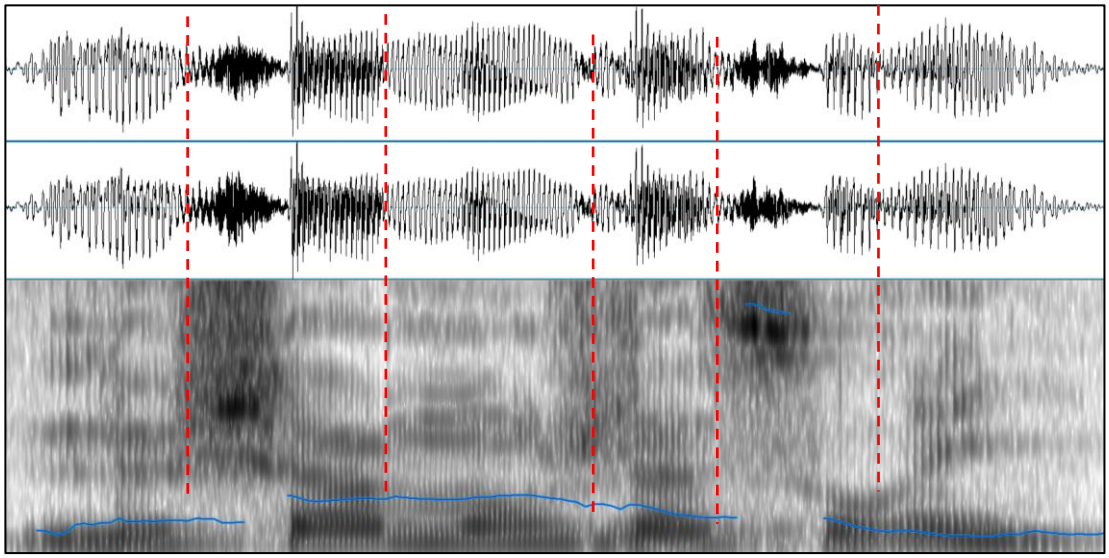


Figure 3-5 Phrase originale : « Le chat mange la souris »
Les lignes pointillées indiquent les limites syllabiques

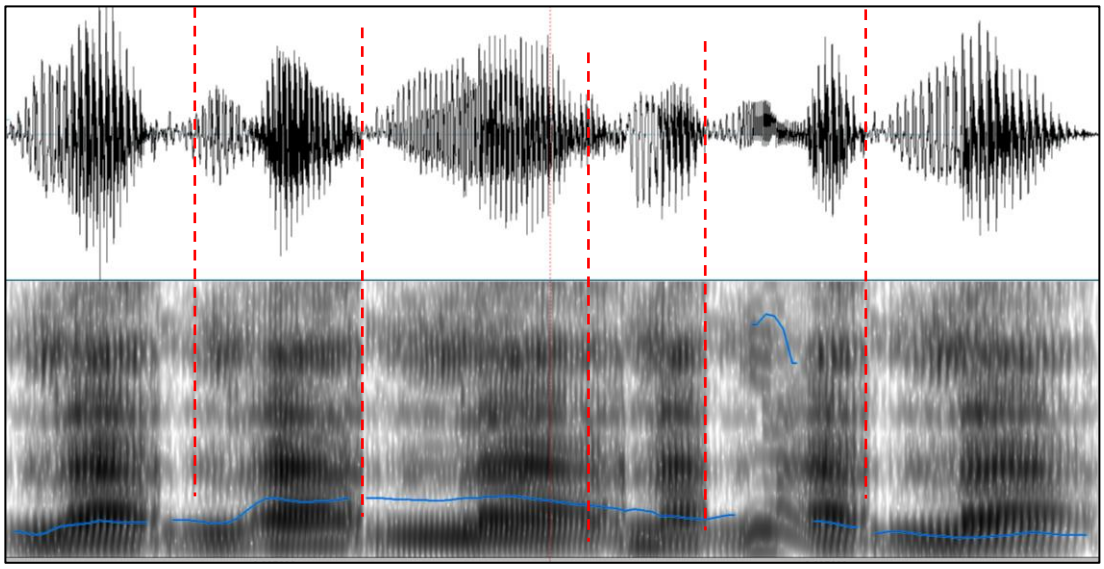


Figure 3-6 Phrase resynthétisée

Bien entendu, cette méthode comporte aussi des limites. Dans un premier temps, les diverses manipulations acoustiques, sans traitement du signal approprié, entraînent une perte importante au niveau du naturel. Cette perte est accentuée par l'utilisation d'une voix de synthèse. Finalement, une erreur s'est introduite à notre insu dans l'algorithme que nous n'avons pu corriger avant la passation du test. Il aurait fallu contrôler la durée des voyelles en plus de la durée des syllabes. Malheureusement, ceci entraîna une petite perte d'information. Il a été jugé que cette perte était minimale, surtout considérant que la plupart des études choisissent le filtrage comme outil de délexicalisation et qu'elles perdent de toute façon la totalité des données liées à la durée.

3.3.3 Accord interjuges

Les phrases délexicalisées ont par la suite été soumises à des juges qui avaient pour tâche de déterminer l'origine géographique des locuteurs. Ce test a été créé et administré à l'aide du logiciel Praat. Pour chaque stimulus, les juges devaient répondre à l'aide de l'interface présentée dans la figure 3-7. Seule la réponse était

retenue, pas le temps de réaction. En effet, la fatigue étant importante, les temps de réaction étaient inutilisables.

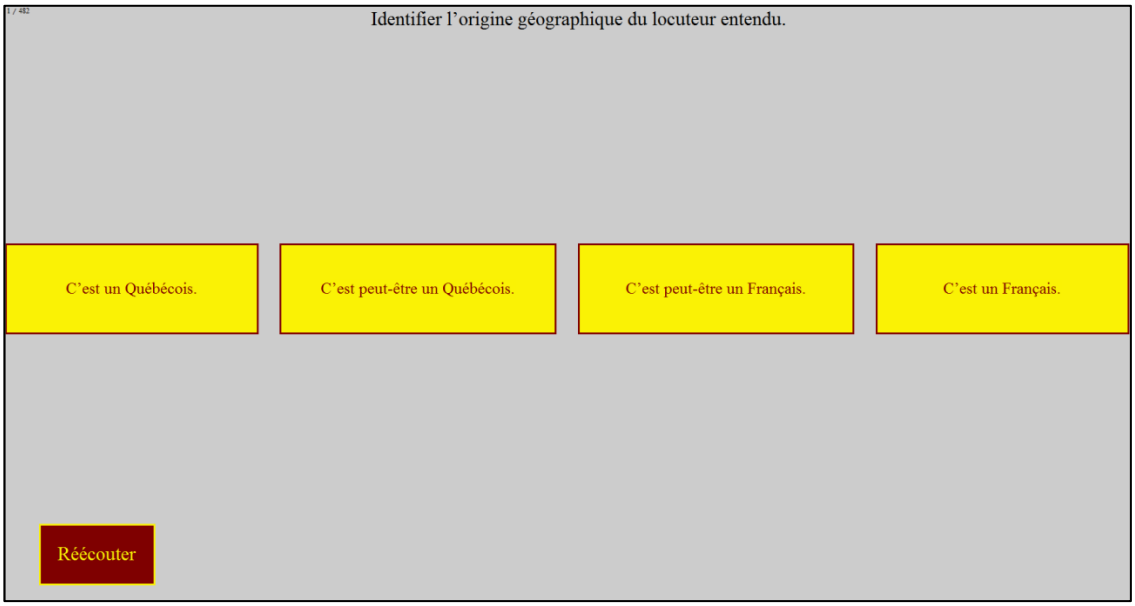


Figure 3-7 Interface de l'accord interjuges

Tel qu'on peut le voir dans la figure, les participants devaient répondre, pour chaque phrase, à une seule question : « Identifier l'origine géographique du locuteur entendu ». Ils avaient quatre choix de réponse : « C'est un Québécois », « C'est peut-être un Québécois », « C'est peut-être un Français », « C'est un Français ».

3.3.3.1 Choix des juges

Les juges étaient des étudiants du département de linguistique habitués à participer à ce type d'études. Il y avait trois femmes et deux hommes. Leurs données sociodémographiques pertinentes sont présentées dans le tableau suivant.

Tableau 3-3 Données sociodémographiques des juges

	ÂGE	Sexe	ORIGINE	Occupation
J1	30	F	Québec	Membres du laboratoire de phonétique
J2	22	F	Montréal	
J3	32	F	Laval	
J4	25	M	Longueuil	Étudiants en linguistique
J5	28	M	Montréal	

3.3.3.2 Déroulement du test

Les juges ont évalué 482 stimuli. Ils ont pris environ une heure pour réaliser ce test. Puisqu'un test de cette longueur et de cette complexité perceptive est cognitivement demandant, les sujets pouvaient prendre des pauses de la longueur de leur choix à tous les 20 stimuli. Chaque phrase pouvait être écoutée jusqu'à trois fois. Ce format de test correspond, en outre, au standard de l'industrie. Par exemple, dans le dernier *Blizzard Challenge* (King et Karaiskos 2013), les juges qui complétaient le test y consacraient de 45 minutes à une heure et chaque session d'écoute comprenait de 10 à 15 stimuli.

Les juges retenus pour notre test, familiers avec ce type de procédure, ont reçu comme instruction de répondre selon leur première impression. La possibilité de réécouter le stimulus a été ajoutée puisque la fatigue s'installait rapidement et que le participant du pré-test avait rapporté répondre au hasard lorsque son attention se portait ailleurs, ce qui arrivait souvent. Le test s'est déroulé dans la chambre sourde du laboratoire de phonétique de l'UQAM, avec une interface PC.

3.3.3.3 Dégagement des archétypes

Suite à ce test, les énoncés délexicalisés ayant reçu un accord interjuges de 4 ou plus sur 5 ont été retenus aux fins d'analyse[3]. Ces énoncés sont dorénavant nommés archétypes. Il est à noter que plusieurs phrases ont été identifiées comme étant françaises alors qu'elles avaient été produites par des québécois et vice-versa. Nous parlerons donc d'archétypes concordants (ou correctement identifiés) et non-concordants.

[3] Les résultats de l'accord interjuges sont présentés dans le chapitre 4, section 4.1.

3.4 Étape 3 : Synthèse et évaluation

L'étape 2 (section 3.3) présentée précédemment aura permis d'identifier les énoncés qui sont les mieux perçus comme étant FF ou FQ. Les analyses subséquentes visent à répondre plus spécifiquement aux objectifs de recherche.

3.4.1 Analyse et dégagement de règles

Cette étape vise à caractériser le plus finement possible les éléments prosodiques ayant guidé les juges dans leur identification. Une analyse linguistique fine a d'abord été réalisée pour chacun des énoncés retenus.

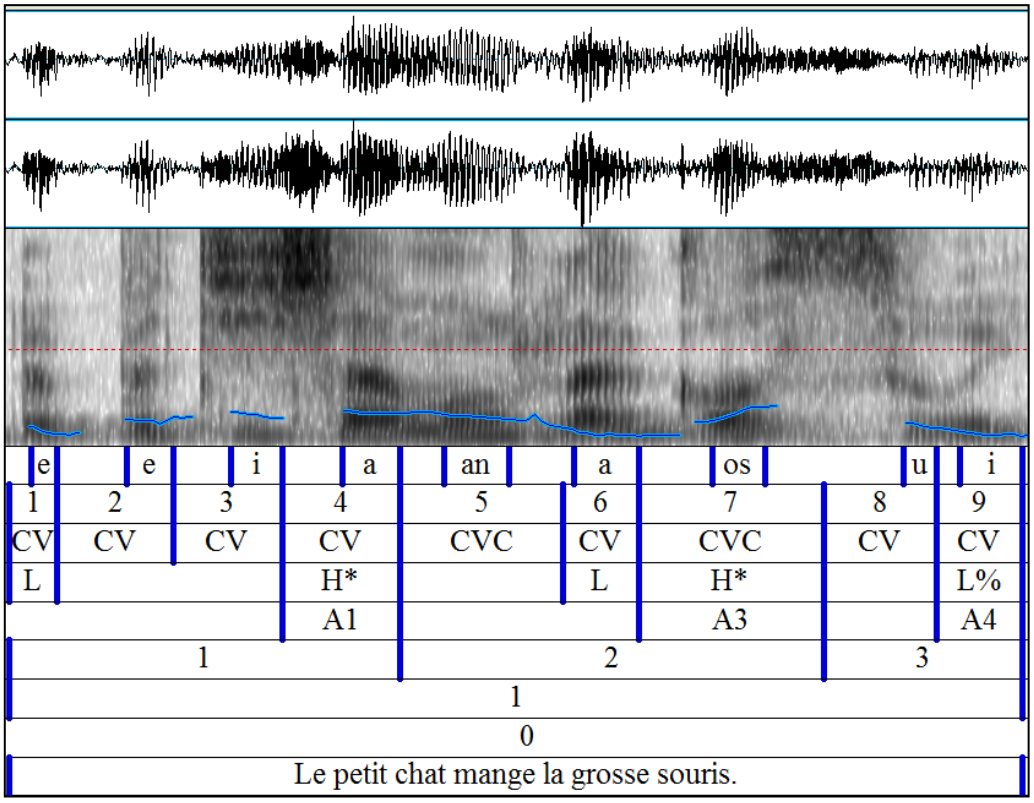


Figure 3-8 Annotation des archétypes

La figure 3-8 montre le travail d'annotation qui a été effectué pour chacun des 232 archétypes. D'abord, sur le premier palier d'annotation, les voyelles ont été

délimitées. Sur le second, on retrouve le découpage syllabique effectué précédemment. Sur le troisième, une étiquette indique le type de syllabe – *C* indiquant une consonne, *V* une voyelle. Sur le palier suivant a été effectué une annotation en tons selon les principes intonatifs de Jun et Fougeron (2000). Le cinquième palier, quant à lui, présente les syllabes accentuées selon le modèle de Di Cristo (2000). Le palier qui suit regroupe les syllabes en groupes rythmiques (GR), puis en syntagmes intonatifs (SI). Finalement, le huitième palier indique la présence de pauses alors que le dernier palier montre la transcription orthographique.

À partir de cette annotation, une grande variété de paramètres acoustiques peut être extraite. Ces paramètres sont représentés dans le tableau suivant. Pour chaque élément prosodique (voyelle, syllabe, GR, SI, A, ton, etc.) des paramètres précis de F0, durée et intensité ont été extraits.

Tableau 3-4 Données acoustiques des archétypes

Éléments prosodiques	Paramètres acoustiques

• Voyelle
• Syllabe
• Type de syllabe
  ○ Ouverte/fermée
• Groupe rythmique
• Syntagme intonatif
• Accents
  ○ A1, A2, A3, A4
• Tons
  ○ L, L%, H*, Hi, H%
• Pauses
	• F0
	  ○ Moyenne, à 25 %, 50 %, 75 %, minimum, maximum, registre (F0 max – F0 min)
	  ○ Ratios de F0 (ex : voyelles non-accentuées/ voyelles accentuées)
	• Durée
	  ○ Moyenne
	  ○ Ratios de durée
	• Intensité
	  ○ Moyenne, à 25 %, 50 %, 75 %, minimum, maximum
	  ○ Ratios d'intensité

Pour chaque élément prosodique de la colonne de gauche, tous les paramètres acoustiques pertinents de la colonne de droite ont été calculés.	

À partir de ces données, une série de tests statistiques ont été réalisés. Ils visaient à atteindre les objectifs suivants :

- déterminer quels paramètres prosodiques (ou combinaison de paramètres)

avaient le plus d'impact dans les décisions prises par les juges;
- déterminer quels énoncés pourraient être considérés comme prototypiques de chaque origine géographique.

3.4.2 Régression logistique

Le modèle statistique choisi afin de remplir ces objectifs est la *régression logistique*[4]. Il s'agit d'un modèle de régression où la variable dépendante est dichotomique (notée 0 ou 1) – ici FF ou FQ - et où les variables indépendantes peuvent être continues ou catégorielles – dans notre cas, elles sont toutes continues. Il s'agit d'un modèle où on mesure l'apport des variables indépendantes à la variable dépendante. L'hypothèse nulle, dans notre cas, se formulerait de la manière suivante : *L'ensemble des variables indépendantes (le modèle) ne permet pas de prédire la variable dépendante (FF ou FQ).* Ainsi, un résultat significatif indique qu'au moins une variable, dans le bassin de variables indépendantes, permet de prédire au moins en partie les résultats de la variable dépendante.

Dans un premier temps, il aura fallu identifier, parmi notre ensemble de près de 300 variables, un sous-ensemble permettant de créer un modèle statistiquement significatif, et ce, pour nos deux critères de sélection : l'identification (l'origine perçue) et l'origine (l'origine réelle). L'identification correspond aux catégories qui ont été assignées aux énoncés par les juges. L'origine, quant à elle, désigne l'origine réelle des locuteurs de chaque énoncé. Un sous-ensemble de 22 variables linguistiquement pertinentes a donc été sélectionné. Il peut être vu dans le tableau 3-5.

Ces variables ont été choisies parce qu'elles couvrent un ensemble large de réalisations prosodiques, les trois paramètres acoustiques (F0, durée et intensité) et parce qu'elles s'implémentent facilement en synthèse.

[4] Les informations théoriques sur la régression logistique présentées ici sont essentiellement tirées du site de statistiques de l'université de Sherbrooke (Yergeau et coll. 2013).

À partir de cet ensemble assez large de variables, un premier modèle statistique a été tenté. Celui-ci s'est avéré être non-significatif à la fois pour l'identification ($p = 0.3898$) et pour l'origine ($p = 0.3117$). Nous avons alors tenté de trouver un sous-ensemble de variables qui permettrait d'expliquer statistiquement les différences entre nos groupes.

Le choix de cet outil statistique nous force à effectuer un élagage important dans nos variables. En effet, seules les variables pertinentes doivent être incluses dans le modèle et des variables colinéaires ou ayant une corrélation trop grande peuvent venir affecter les résultats. Ainsi, plusieurs sous-ensembles de variables ont été testés en tenant compte des réalités linguistiques liées à chaque variable, des corrélations entre elles et des résultats de chaque itération du modèle (les variables approchant un niveau significatif ont été préservées, par exemple).

Le choix d'un modèle statistique de ce genre offre de nombreux avantages, notamment la possibilité de tenir compte d'un ensemble élevé de variables très différentes. Néanmoins, il est certain qu'il ne s'agit ici que d'un *modèle*. Il est possible que des variables acoustiques utilisées par les êtres humains dans leurs jugements linguistiques ne soient pas retenues par le modèle. Néanmoins, les variables retenues par le modèle permettent d'expliquer la variable dépendante, et ce, de manière significative : ce sont de véritables prédicteurs.

Pour le groupe *identification (origine perçue)*, un sous-ensemble de huit variables (tableau 3-7) a permis de générer un modèle statistiquement significatif ($p = 0.0298$, tableau 3-6). Parmi ces huit variables, cinq contribuaient significativement au modèle. Il s'agit des variables *F0 moy* (F0 moyenne), *Registre* (F0 max – F0 min), *syl_durTH-TB* (Durée des syllabes, ratio entre TH et TB), *fmoyTH-TB* (F0 moyenne, ratio entre TH et TB) et *dureeV* (Durée moyenne des voyelles).

Pour le groupe *origine*, une itération avec 18 variables a permis de produire un modèle statistiquement significatif (p = 0.0037, tableau 3-8). De ces 18 variables, seules quatre pouvaient être considérées comme des prédicteurs : *durée V*, *durée S*, *#Acc/#NACC* et *dureeAcc/Nacc* (tableau 3-7).

Tableau 3-5 Variables acoustiques principales

Duree V	Durée moyenne des voyelles
Intensite moy	Intensité moyenne
F0 moy	F0 moyenne
Registre	F0 max-F0 min
Duree S	Durée moyenne des syllabes
syl_durACC - NACC	Ratio entre la durée des syllabes accentuées et celle des syllabes non accentuées
syl_durTH - TB	Ratio entre la durée des syllabes marquées par un ton haut et celle des syllabes marquées par un ton bas
syl_durTH - TNM	Ratio entre la durée des syllabes marquées par un ton haut et celle des syllabes non marquées par un ton
syl_durTB - TNM	Ratio entre la durée des syllabes marquées par un ton bas et celles non marquées par un ton
#ACC / #syl	Nombre de syllabes accentuées sur le nombre total de syllabes dans un énoncé
dureeACC - NACC	Ratio entre la durée de la voyelle des syllabes accentuées et celle de la voyelle des syllabes non accentuées
dureeTH - TB	Ratio entre la durée de la voyelle des syllabes marquées par un ton haut et celle de la voyelle des syllabes non marquées par un ton
dureeTH - TNM	Ratio entre la durée de la voyelle des syllabes marquées par un ton haut et celle de la voyelle des syllabes non marquées par un ton
dureeTB - TNM	Ratio entre la durée de la voyelle des syllabes marquées par un ton bas et celle de la voyelle des syllabes non marquées par un ton

fmoyACC - NACC	Ratio de la F0 moyenne entre les syllabes accentuées et non accentuées
fmoyTH - TB	Ratio de la F0 moyenne entre les syllabes marquées par un ton haut et celle des syllabes marquées par un ton bas
fmoyTH - TNM	Ratio de la F0 moyenne entre les syllabes marquées par un ton haut et celle des syllabes non marquées par un ton
fmoyTB - TNM	Ratio de la F0 moyenne entre les syllabes marquées par un ton haut et celle des syllabes non marquées par un ton
imoyACC - NACC	Ratio de l'intensité moyenne entre les syllabes accentuées et non accentuées
imoyTH - TB	Ratio de l'intensité moyenne entre les syllabes marquées par un ton haut et celle des syllabes marquées par un ton
imoyTH - TNM	Ratio de l'intensité moyenne entre les syllabes marquées par un ton haut et celle des syllabes non marquées par un ton
imoyTB - TNM	Ratio de l'intensité moyenne entre les syllabes marquées par un ton haut et celle des syllabes non marquées par un ton

Tableau 3-6 Variables statistiquement significatives pour le groupe *Origine perçue*

IDENDIFICATION - Test of all effects Modeled probability that IDENDIFICATION = Q			
	Degr. of - Freedom	**Wald - Stat.**	**p**
Intercept	1	0.40490	0.524568
Intensite moy	1	2.29701	0.129623
F0 moy	**1**	**20.25056**	**0.000007**
Registre (F0 max-F0 min)	**1**	**11.65029**	**0.000642**
syl_durTH - TB	**1**	**11.21467**	**0.000812**
#ACC / #syl	1	1.40497	0.235894
fmoyTH - TB	**1**	**15.91635**	**0.000066**
Duree V	**1**	**10.62496**	**0.001116**
dureeTH - TB	1	0.41347	0.520213

Tableau 3-7 Variables statistiquement significatives pour le groupe *Origine réelle*

ORIGINE - Test of all effects Modeled probability that ORIGINE = F			
	Degr. of - Freedom	**Wald - Stat.**	**p**
Intercept	1	0.06274	0.802220
Duree V	**1**	**27.29925**	**0.000000**
Registre (F0 max-F0 min)	1	1.60189	0.205636
Duree S	**1**	**6.70208**	**0.009630**
syl_durACC - NACC	1	0.11973	0.729323
syl_durTH - TB	1	2.85592	0.091038
syl_durTH - TNM	1	1.52551	0.216788
syl_durTB - TNM	1	1.16173	0.281108
#ACC / #syl	**1**	**8.37100**	**0.003813**
dureeACC - NACC	**1**	**6.02978**	**0.014067**
dureeTH - TB	1	0.89247	0.344808
dureeTH - TNM	1	2.49735	0.114038
dureeTB - TNM	1	0.28279	0.594877

fmoyACC - NACC	1	1.69078	0.193498
fmoyTH - TB	1	0.20562	0.650222
fmoyTH - TNM	1	0.02611	0.871625
fmoyTB - TNM	1	0.00556	0.940583
imoyTH - TB	1	0.31448	0.574947
imoyTB - TNM	1	0.08271	0.773658

3.4.3 Synthèse

À partir de ces données sur les paramètres acoustiques les plus pertinents à l'identification de l'origine géographique des locuteurs ainsi qu'à leur origine réelle, des phrases ont été synthétisées.

Il s'agit ici d'implémenter à des niveaux divers les paramètres identifiés lors de l'analyse statistique. Les paramètres retenus pour cette analyse sont ceux obtenus à l'aide du groupe *Identification*. En effet, il est important, dans une perspective de synthèse, de ne pas se contenter de reproduire les productions des locuteurs, mais aussi de se concentrer sur les paramètres qui ont le plus d'impact sur les auditeurs. Ainsi, le choix des variables du groupe *Identification* permet de concentrer nos efforts sur des éléments prosodiques qui ont été bien perçus par les juges.

3.4.3.1 Choix des seuils acoustiques

Afin de tester ces paramètres prosodiques, un certain nombre de stimuli a été créé. Dans un premier temps, chacun des cinq paramètres acoustiques identifiés plus haut aura été étudié de manière indépendante. Pour ce faire, une série de paliers (ou seuils) a été définie allant d'une valeur menant à une identification québécoise à une indentification française.

Dans cette optique, les cinq paramètres acoustiques identifiés dans le tableau 3-7 ont fait l'objet d'une analyse plus poussée. Pour chaque variable, un graphique montrant les valeurs moyennes et les écarts types pour chaque groupe a été créé. Ces graphiques permettent d'identifier la directionalité des variables. Autrement dit, ils

permettent d'identifier plus clairement les valeurs auxquelles les participants assignent une origine géographique plus française ou plus québécoise. Par exemple, le graphique pour la F0 moyenne permet de voir qu'une F0 moyenne plus élevée est associée à une identification plus FF alors qu'une F0 moyenne plus grave est associée à une identification plus FQ. Ces cinq graphiques sont présentés dans l'appendice E. Dans un premier temps, quatre seuils ont été définis (tableau 3-8). Chaque seuil représente un point dans les graphiques : les moyennes pour les groupes FF et FQ ou encore les limites des écarts types. Ces choix ont aussi été guidés par nos connaissances linguistiques de chaque phénomène prosodique retenu. De plus, il nous a semblé pertinent de créer des échelles assez larges – raison pour laquelle les points limites des écarts types ont été retenus comme seuils dans plusieurs cas. Par exemple, le point le plus FF pour la F0 moyenne a été établi à 150 Hz – ou le point supérieur de l'échelle d'écart-type pour l'identification FF. Le point le plus FQ, quant à lui, a été établi à 120 Hz – ou la limite inférieure de l'échelle d'écart type pour l'identification FQ. Un processus de sélection similaire a été employé pour les seuils des tableaux 3-8 et 3-9.

Tableau 3-8 Cibles des cinq paramètres prosodiques retenus

Paramètre acoustique	1 Plutôt français	2 ←	3 →	4 Plutôt québécois
F0 moyenne	150 Hz	140 Hz	130 Hz	120 Hz
Registre	45 Hz	30 Hz	20 Hz	10 Hz
Durée V	60 ms	80 ms	95 ms	110 ms
F0 TH/TB	2	1.5	1	0.5
Durée S TH/TB	1.5	1.3	1	0.7

Cependant, comme il sera expliqué plus loin dans ce chapitre, le pré-test nous aura permis de constater que le nombre de phrases ainsi créé était beaucoup trop important pour être évalué en une seule séance. Ainsi, il a été plutôt décidé de choisir trois paliers, présentés dans le tableau 3-9.

Tableau 3-9 Cibles des paramètres prosodiques retenus

Paramètre acoustique	1 Plutôt français	2 ↔	3 Plutôt québécois
F0 moyenne	150 Hz	135 Hz	120 Hz
Registre	120 Hz	60 Hz	10 Hz
Durée V	60 ms	80 ms	95 ms
F0 TH/TB	2	1.5	1
Durée S TH/TB	1.5	1.3	1

Ces seuils permettent donc de tester des modifications apportées à chaque paramètre acoustique pris en isolation. Dans un second temps, des combinaisons de paramètres ont été effectuées. Ainsi, les deux paramètres liés à la durée sont appliqués à une série de phrases. De même, les trois paramètres liés à la F0 sont appliqués à une autre série de phrases. Finalement, deux séries permettent de rendre compte de l'ensemble des paramètres. Une série où les seuils pour la durée sont identifiés comme québécois alors que les seuils pour la F0 sont identifiés comme français. Une dernière série où les seuils pour la durée et la F0 sont concordants selon l'origine.

3.4.3.2 Création des phrases porteuses

Ces cibles doivent être présentées sur des phrases porteuses. Le choix méthodologique retenu pour ces phrases aura été d'utiliser des phrases SUS, tel que présenté dans le chapitre 1. Malheureusement, le corpus réalisé par Boula de Mareüil et coll. (2006) employait des structures syntaxiques différentes de celles utilisées dans le corpus initial. Il ne comprenait notamment pas de phrases permettant de réaliser des accents d'emphase. Il a donc été décidé de créer nos propres phrases.

Pour ce faire, un lexique a été créé à l'aide du dictionnaire *Le Petit Robert* (Robert, Rey et Rey-Debove 2009). Par la suite, un simple script Python permet de sélectionner les mots aléatoirement afin d'en faire des phrases, selon une structure préalablement choisie. Les phrases ont ensuite été sélectionnées afin d'éliminer celles

qui, par pur hasard, étaient sémantiquement prévisibles et celles qui étaient syntaxiquement incorrectes.

Six structures ont été réalisées. Elles sont présentées, à l'aide d'exemples, dans le tableau 3-10.

Tableau 3-10 Structures de phrase

Type de phrase	Structure	Exemple
Neutre (1)	Det, N, V, Det, N	Des fous redressent le nez.
Neutre (2)	Det, Adj, N, V, Det, Adj, N	Cette copie boulotte raffine le matou chanceux.
Emphase (1)	*C'est* Det N *qui* V Det N, *pas* Det N	Ce sont ces chats qui dérobent le nez, pas ce robot.
Emphase (2)	Det **N** V Det Adj N	Des **pies** dérobent les douces fées.
Question (1)	Det Adj N V *quoi ?*	Un chou dodu bouge quoi ?
Question (2)	Q Det N V Pro ?	Que ce bison répare-t-il ?

Dans un premier temps, un bassin de 216 phrases est ainsi obtenu. En effet, il y a neuf séries à étudier, pour lesquelles il y a quatre paliers et six types de phrases, donc 9*4*6 = 216. Or, comme il a été mentionné précédemment, ce volume d'énoncés était beaucoup trop grand. En effet, une première tentative de passation du test résultant a pris près de trois heures. Il a donc été décidé, en plus de réduire le nombre de paliers, de réduire le nombre de types de phrases. Ainsi, il a été choisi de n'utiliser qu'une version de phrase neutre, une phrase avec emphase et une phrase interrogative. La version de phrase sans l'adjectif a été retenue. Ces changements amènent le nombre d'énoncés à 81 (9*3*3 = 81). Ces 81 stimuli sont présentés dans l'appendice C. On y retrouve la transcription orthographique, la transcription phonétique utilisée, la série et les cibles.

3.4.3.3 Présentation de Mbrola

À partir de ces phrases, une synthèse a été créée. Avant d'aborder plus en détail le développement de cette synthèse, il convient ici de présenter le système retenu : Mbrola (Dutoit et coll. 1996).

Il s'agit d'un système de concaténation de diphones. Ce système a été choisi pour cette étude puisqu'il possède une voix masculine en français québécois (CAN 1) et qu'il permet un contrôle fin des paramètres prosodiques. La figure 3-9 présente un exemple de phrase dans l'interface. Chaque ligne présente les données pour un phonème. Le premier caractère est le phonème en tant que tel. Le chiffre qui suit est la durée de ce phonème. Si des paires de chiffres suivent, elles permettent de déterminer la F0 pour ce phonème. En effet, le premier chiffre de chaque paire présente le point (en pourcent) de la durée de la voyelle où la cible de F0, le second chiffre, doit être atteinte.

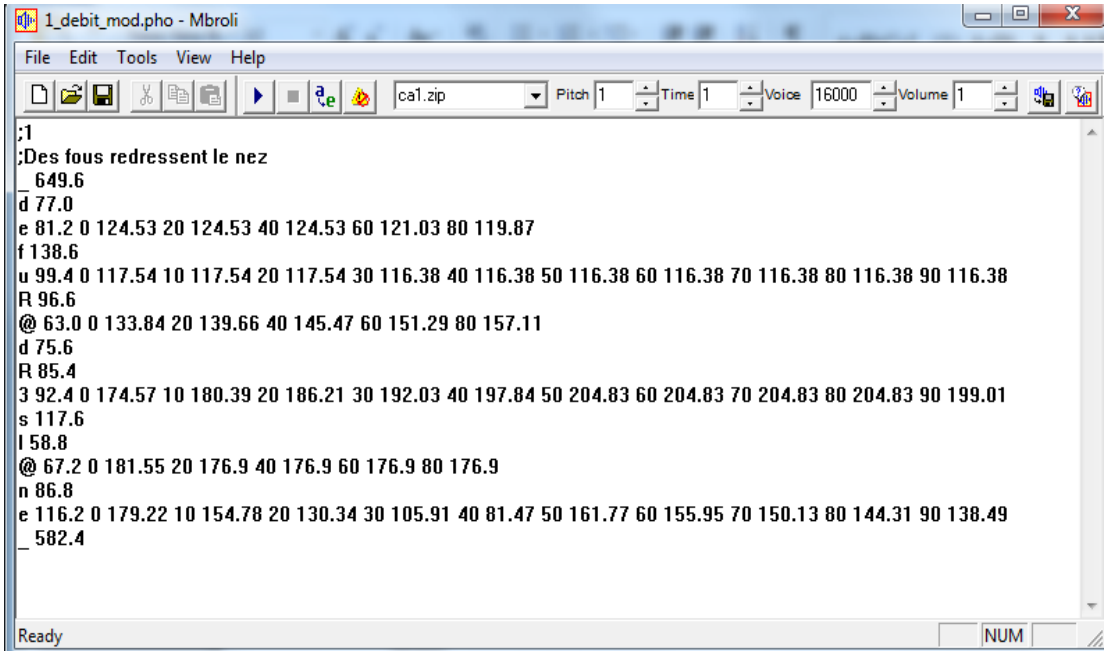


Figure 3-9 Interface Mbrola

3.4.3.4 Euler : fonctionnement du TTS

Mbrola n'est qu'une partie d'un système de synthèse – le tout dernier module, celui qui génère la parole. Un module supplémentaire est nécessaire afin de générer une courbe prosodique initiale. Ce module prend comme point de départ une phrase transcrite orthographiquement et génère une phonétisation – les phonèmes, leur durée et leur F0. Ici, le système Euler a été retenu (Bagein et coll. 2000). Il est disponible librement sur le Net, s'interface avec Mbrola et génère une phonétisation comprenant une analyse phonologique en tons hauts et bas. Puisque les paramètres acoustiques les plus pertinents devaient tenir compte des tons hauts et bas, c'est ce système qui a été retenu. Malheureusement, il ne permettait pas d'effectuer de phonétisation en français québécois. Cette étape a donc dû être ajustée manuellement. L'appendice D permet de montrer un fichier type produit par le système Euler.

À partir de cette phonétisation, des modifications seront apportées aux données prosodiques initiales afin que chaque phrase corresponde adéquatement à la cible qui lui a été attribuée. Pour ce faire, cinq modules de transformation ont été créés – un module pour chaque paramètre prosodique à l'étude. Ils sont présentés en détail ci-dessous. Chaque module a été implémenté à l'aide du langage de programmation Python. Le code est disponible en appendice F.

3.4.3.5 Modification de la F0 moyenne

L'objectif de la modification de F0 était de faire en sorte que la F0 moyenne des énoncés atteigne la cible retenue. Pour ce faire, chaque valeur de F0 aura été multipliée par le ratio obtenu par l'équation ci-dessous.

$$R = \frac{cible}{\bar{x}\, F0}$$

Ainsi, si la cible à atteindre est plus élevée que la moyenne actuelle, le ratio (R) sera plus grand que 1. Autrement, le ratio sera entre zéro et un. Chaque valeur est modifiée d'une manière proportionnelle – la courbe de F0 est ainsi préservée.

3.4.3.6 Modification du registre

La modification du registre vise à s'assurer que le registre – soit l'écart entre les valeurs maximales et minimales de F0, corresponde bien aux cibles établies. Pour ce faire, la fonction mathématique ci-dessous a été appliquée à chaque valeur de F0.

$$F0_{initiale} + R\left(\frac{F0_{initiale} - F0_{médiane}}{F0_{médiane} - F0\ min}\right)$$

$$o\text{ù } R = \frac{Cible - (F0\ max - F0\ min)}{2}$$

$$et\ o\text{ù } F0_{médiane} = \frac{F0\ \max - F0\ min}{2} + F0\ min$$

En effet, il fallait éviter de créer un effet plafond et plancher – ce qui aurait grandement affecté le naturel et la qualité de la phrase sans être représentatif d'une production naturelle avec un registre plus large ou plus étroit. L'utilisation d'un ratio, tel que présenté dans l'équation ci-dessus, permet de maintenir une proportion dans la courbe.

3.4.3.7 Modification de la durée des voyelles

La modification de la durée des voyelles, quant à elle, visait à changer la durée moyenne des voyelles. La durée moyenne de celles-ci devait alors correspondre à la cible retenue. La durée de chaque voyelle a ainsi été multipliée par un ratio calculé par la formule suivante.

$$R = \frac{cible}{\bar{x}\ durée\ V}$$

Il s'agit de la même formule mathématique que pour la modification de la F0 moyenne. En effet, il s'agissait ici de modifier les durées des voyelles, et ce, d'une manière uniforme afin que la durée moyenne corresponde à la cible sans que les différences de durée entre les différentes voyelles ne soient modifiées.

3.4.3.8 Modification de F0 entre les TH et TB

L'objectif de la modification de F0 entre les TH et TB était de changer le ratio créant les différences de F0 entre les tons hauts et bas. Pour ce faire, les valeurs de F0 des tons hauts ont été modifiées, multipliées par un ratio (R) calculé par la formule mathématique ci-dessous.

$$R = \frac{cible}{\bar{x}\,TH/\bar{x}\,TB}$$

Cette formule ne s'applique que sur les tons hauts. En effet, nous avons constaté que les tons bas générés par Euler étaient, à la fois pour la durée et la F0, dans les valeurs les plus basses possibles. Diminuer les valeurs de F0 ou de durée pour les TB affectait grandement la qualité des énoncés. Ainsi, la formule mathématique retenue ici permet de faire en sorte que le ratio moyen existant entre les TH et TB pour un énoncé corresponde à la cible tout en conservant des différences d'un ton à l'autre. Autrement dit, si la cible est, par exemple, un ratio de 1.5, ce ne sont pas toutes les frontières TH/TB qui seront différentes par un ratio de 1.5 – c'est plutôt le ratio entre la valeur moyenne des TH et la valeur moyenne des TB qui sera de 1.5. Les différents ratios individuels seront donc modifiés d'une manière proportionnelle – la courbe initiale sera relativement préservée.

3.4.3.9 Modification de la durée des syllabes entre les TH et TB

Finalement, cette dernière modification (durée des syllabes entre TH et TB) visait elle aussi les ratios entre tons hauts et bas, mais sur le paramètre acoustique de la durée des syllabes. La fonction mathématique utilisée est la même que pour la modification

précédente. Elle s'applique cependant à la durée plutôt qu'à la F0. Ainsi, les durées de chaque segment pour les syllabes étiquetées TH sont multipliées par le ratio R.

Pour les séries demandant des combinaisons de plusieurs paramètres, ces cinq modules auront été appliqués au fichier initial dans un ordre précis. Ces modifications sont présentées dans le tableau 3-11.

Tableau 3-11 Ordre d'application des fonctions de modification

Série	Ordre d'application des fonctions
F0	Registre, syl_f0, f0 moyen
Durée	Duree_v, syl_dur, duree_v
Série totale concordante et série totale inversée	Registre, syl_f0, f0 moyenne, duree_v, syl_dur, dureeV

Dans plusieurs cas, certaines fonctions sont appliquées à plus d'une reprise. Pour mieux comprendre les raisons derrière cette démarche, prenons l'exemple de modification de la durée. Dans un premier temps, les modifications sont apportées à la durée des voyelles. Puis, c'est la durée des syllabes qui est modifiée, les tons hauts par rapport aux tons bas. Cette modification entraîne, forcément, des changements au niveau de la durée des voyelles. Afin de maintenir une moyenne fiable de durée de voyelles, la première fonction a été appliquée une seconde fois.

Finalement, lors de la première écoute des fichiers résultats, il est apparu que le débit était beaucoup trop rapide. Il a donc été décidé d'augmenter la durée de chaque segment uniformément par un facteur de 1.4 [1], et ce, avant l'application des modifications.

[1] Ce facteur a été choisi après essais et erreurs.

3.4.4 Test de perception

Afin de tester les impacts des diverses manipulations ainsi effectuées, un test de perception a été créé. Dans un premier temps, il vise à déterminer si les changements prosodiques effectués ont eu un impact sur l'intelligibilité de la voix de synthèse. Dans un second temps, il cherche également à déterminer si ces modifications ont eu un impact sur la perception de l'origine géographique des voix. Finalement, il sera question de qualité – le test tentera de vérifier les impacts des modifications sur divers éléments liés à la qualité.

Le test a été créé à l'aide d'une interface HTML utilisée localement. Les phrases sont présentées de manière aléatoire. Pour chaque phrase, les trois étapes de test (intelligibilité, origine géographique, test d'appréciation) sont présentées une à la suite de l'autre.

De plus, au début du test, les participants ont répondu à un questionnaire sociodémographique basé sur les langues connues et sur les endroits où ils ont vécu. Chacune de ces étapes sera présentée plus en détail ci-dessous.

3.4.4.1 Choix des participants

Un total de 15 participants a été retenu pour ce test de perception. Divers critères ont guidé leur sélection. D'abord, les participants devaient impérativement avoir le FQ comme langue maternelle. Ils devaient aussi être majeurs, ne pas avoir de trouble auditif ou langagier. De plus, puisque le test comprenait une tâche de transcription à l'ordinateur, il était requis des participants d'être à l'aise avec le clavier. Les données sociodémographiques des participants sont présentées ci-dessous.

Tableau 3-12 Données sociodémographiques des participants

	ÂGE	Sexe	ORIGINE
P1	27	F	Montréal
P2	23	F	Saint-Jean-Port-Joli
P3	52	F	Saint-Romuald
P4	21	F	Québec
P5	26	F	Montréal
P6	26	F	Québec
P7	58	F	Sainte-Thérèse
P8	57	M	Montréal
P9	37	M	Gaspé
P10	58	M	Sainte-Thérèse
P11	30	M	Montréal
P12	27	M	Saint-Jean-Port-Joli
P13	28	M	Québec
P14	22	M	Cornwall
P15	48	M	Québec

3.4.4.2 Présentation du test

L'ensemble des étapes du test est présenté dans la figure 3-10. Les tâches ont été présentées une à une avec leurs instructions. Il était possible pour le participant de poser des questions à l'expérimentatrice. De plus, une phase de familiarisation de six énoncés a été ajoutée afin de permettre au participant de se familiariser avec les diverses tâches. Le test a été d'une durée approximative d'une heure. Des pauses ont été incluses à tous les vingt stimuli. 81 stimuli sont présentés, en plus des six stimuli de la phase de familiarisation.

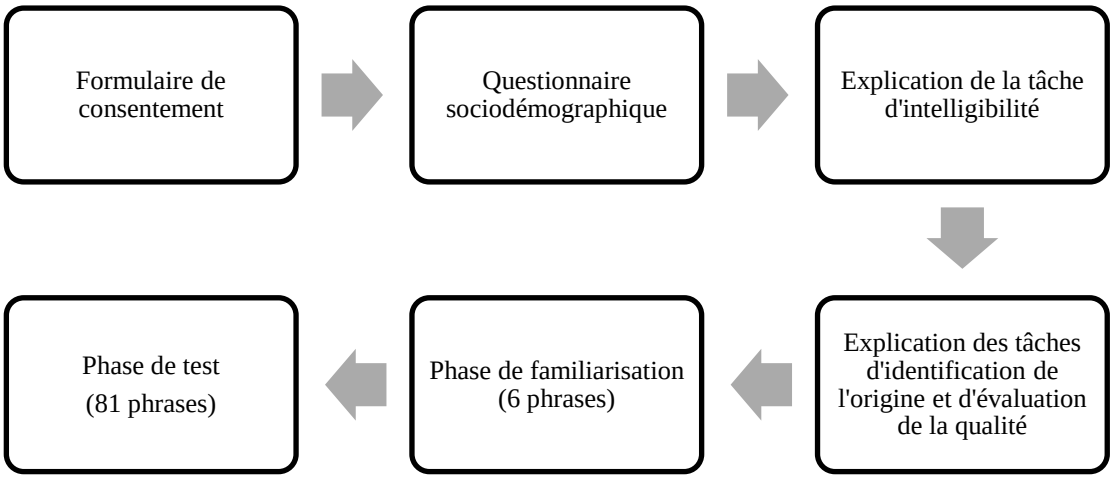


Figure 3-10 Étapes du test de perception

3.4.4.3 Questionnaire

Le questionnaire est présenté dans l'appendice G. Il s'agit d'un court questionnaire sociodémographique permettant de déterminer le sexe et l'âge du participant, de même que l'absence de difficultés auditives et langagières. Des questions ont aussi été posées afin de préciser le bagage langagier des participants. Une série de questions a été créée comprenant des questions sur la ville d'origine, les langues connues et les endroits où le participant a vécu.

3.4.4.4 Tâche d'intelligibilité

L'interface de cette tâche est présentée dans l'appendice G. Le participant a entendu l'énoncé à évaluer une seule fois. Il a dû ensuite, à l'aide du clavier, transcrire cette phrase le plus justement possible. Le pré-test aura permis de constater que l'utilisation de phrases SUS rendait la tâche suffisamment difficile sans qu'un bruit supplémentaire n'ait à être ajouté.

3.4.4.5 Évaluation de la tâche d'intelligibilité

L'évaluation des transcriptions produites par les participants s'est faite sur la base d'une notation par phonème correctement identifié. Ainsi, une phrase comportant 14 phonèmes serait notée sur 14 puis transférée en pourcentage.

Outre le simple rappel phonémique, certains cas ont entraîné eux aussi la perte de points. L'exemple suivant est une transcription offerte par le participant P8 à la phrase « Ce chou combat des râteaux ».

(1) Ce choucon bat des rateau

Cette transcription, bien qu'elle respecte la transcription phonémique, ne marque pas les coupures de mots aux bons endroits, ce qui affecte clairement la compréhension de la phrase. Il a donc été décidé de retirer un point pour ce type d'erreur.

Les erreurs liées aux déterminants ont elles aussi fait l'objet d'une modification de points. En effet, il était très commun que les participants inversent divers déterminants dans la phrase, comme dans l'exemple 2 écrit par le participant P3.

(2) C'est un vœu qui récite **ce** mot, pas **le** mouton.
*C'est un vœu qui récite **le** mot, pas **ce** mouton.*

Il est aussi arrivé que des participants se contentent de mettre le même déterminant partout dans l'énoncé, comme dans l'exemple 3, transcrit par le participant P13.

(3) Ce sont **les** rats qui consignent **les** ponts, pas **les** rats.
*Ce sont **des** rats qui consignent **des** ponts, pas **des** rats.*

Il est apparu évident, lors de la notation, que ces erreurs, bien qu'affectant plusieurs phonèmes, abaissaient beaucoup trop les résultats liés à l'intelligibilité. Par exemple, dans l'exemple 2, le participant avait correctement identifié tous les mots de contenu

et avait simplement inversé les déterminants. Or, une notation par rappel phonémique entraînait une note de 10/14 – il perdait alors quatre points – une note qui ne rend pas justice à l'intelligibilité de la phrase, ni même à la capacité de rappel du participant. Il a donc été décidé de n'enlever qu'un seul point pour une erreur d'inversion, et un point par déterminant pour les erreurs comme celles de l'exemple 3. Ainsi, l'exemple 2 s'est mérité une note de 24/25 et l'exemple 3 une note de 24/25.

Dans les cas où un participant a ajouté des phonèmes et que cela a clairement affecté la compréhension de la phrase, comme dans l'exemple 4 (participant P14), des points ont aussi été enlevés. Ainsi, la transcription 4 s'est vue attribuée une note de 26/28.

(4) C'est son bureau qui recadre les **courbes**, pas son palais.
*C'est son bureau qui recadre les **coups**, pas son palais.*

3.4.4.6 Origine géographique

Lors de cette étape, les participants peuvent réécouter la phrase – un bouton est mis à leur disposition. Ils doivent ensuite, à l'aide d'une échelle, donner leur opinion quant à l'origine géographique du locuteur de l'énoncé écouté. L'échelle avait, d'un côté, une origine plus française, et de l'autre, une origine plus québécoise. L'interface, de même que l'énoncé utilisé, peuvent être vus dans l'appendice G.

3.4.4.7 Tâche d'appréciation

Ce test, quant à lui, présentait trois questions. Ces questions sont tirées de l'ensemble d'évaluation ACR présenté au chapitre 1 dans le tableau 1-2. Sur l'ensemble de sept questions, trois seulement ont été retenues pour ce mémoire. En effet, les modifications apportées aux phrases ne devaient théoriquement pas affecter les résultats aux questions retirées – questions qui portaient principalement sur les qualités segmentales de la voix ou encore sur la qualité du son de source. Seules les questions pouvant être affectées par des éléments prosodiques ont été conservées. Les

trois questions retenues sont celles de la qualité générale, du confort d'écoute et du naturel. Afin de répondre à ces questions, les participants pouvaient réécouter les énoncés au besoin. Ils devaient répondre à chaque question à l'aide d'une échelle. La question liée à la qualité générale, « Comment appréciez-vous globalement ce que vous venez d'entendre » avait une échelle notée de 1 à 10 allant de « très mauvais » à « très bon ». La question liée au confort d'écoute « Comment décririez-vous cette voix », tant qu'à elle, avait une échelle allant de « très désagréable » à « très agréable ». Finalement, la question liée au naturel, « Comment appréciez-vous le naturel de ce que vous venez d'entendre », avait une échelle allant de « très artificiel » à « très naturel ». L'interface utilisée peut être visualisée dans l'appendice G.

3.5 Survol des objectifs

Dans cette section, nous souhaitons survoler chaque objectif afin de clarifier comment la méthodologie présentée dans ce chapitre permettra d'y répondre.

Le premier objectif a été formulé comme suit : « Quelles sont les principales caractéristiques prosodiques du français québécois, dans sa perception et sa production ». Nous croyons que les archétypes dégagés, suite à l'accord interjuges, permettent de répondre à cet objectif. En effet, les caractéristiques prosodiques de ces archétypes sont par la suite finement analysées. Les résultats ainsi obtenus permettront de dresser un portrait clair des éléments prosodiques permettant l'identification du FQ. C'est ce que nous entendons par l'expression « dans sa perception ». Il s'agit de trouver les paramètres prosodiques du FQ qui sont les mieux perçus. Les archétypes nous permettent aussi de répondre à l'objectif « dans sa production ». En effet, les archétypes FQ non-concordants peuvent nous éclairer sur des caractéristiques prosodiques produites mais mal perçues.

Le second objectif « dans quelle mesure la prosodie du FQ affecte-t-elle l'intelligibilité ou la qualité d'une synthèse » sera traité à l'aide du test de perception. En effet, il sera alors possible de vérifier comment les résultats à ce test sont affectés par diverses modifications prosodiques.

Le troisième objectif « quels paramètres prosodiques ou combinaison de paramètres sont les plus pertinents dans l'amélioration d'une synthèse de la parole en FQ » sera lui aussi évalué à l'aide du test de perception. Puisque diverses manipulations prosodiques seront testées, il sera alors possible de vérifier l'impact de chacune d'elles sur les résultats aux tests de perception.

CHAPITRE IV

PRÉSENTATION DES RÉSULTATS

Dans ce chapitre, nous aborderons les résultats des diverses manipulations présentées dans le chapitre précédent. Dans un premier temps, nous verrons les résultats de l'accord interjuges (section 3.3.3). Par la suite, les résultats de l'analyse acoustique des archétypes seront étudiés (section 3.4.2). Finalement, les résultats au test de perception seront présentés (section 3.4.4).

4.1 Accord interjuges

Les résultats du test d'accord interjuges sont présentés ici. Lors de ce test, 482 phrases délexicalisées, 233 FQ et 249 FF, ont été présentées à cinq juges québécois. Ceux-ci avaient pour tâche d'identifier l'origine géographique du locuteur entendu. Les énoncés ayant reçu un accord de quatre ou plus sur cinq ont été retenus pour la suite de l'analyse. Le tableau suivant permet de visualiser les résultats.

On observe d'abord que les juges ont eu de la difficulté à identifier correctement l'origine géographique des locuteurs. En effet, le taux d'identification correcte se situe à 54%. Néanmoins, ils s'entendent à 4/5 sur un nombre assez important d'énoncés (236/482 ou 48%).

Tableau 4-1 Résultats de l'accord interjuges

	Origine FF Identifié FF	Origine FF Identifié FQ	Origine FQ Identifié FQ	Origine FQ Identifié FF	Total
Accord 5/5	10	23	16	8	57
Accord 4/5	45	46	54	34	179
Accord 3/5	69	56	70	51	246
Accord 2/5	56	69	51	70	246
Accord 1/5	46	45	34	54	179
Accord 0/5	23	10	8	16	57
Total	249	249	233	233	

C'est donc à partir des énoncés ayant reçu un accord interjuges de quatre ou plus sur cinq que l'analyse acoustique, présentée dans la section suivante, a été effectuée.

4.2 Analyse acoustique

Suite à l'accord interjuges, une analyse acoustique a été réalisée, telle que décrite dans la section 3.4 du chapitre précédent. Cette analyse acoustique a fait l'objet d'une analyse statistique – la régression logistique, elle aussi présentée dans le chapitre précédent. Les résultats nécessaires à la compréhension de la suite de la méthodologie ont aussi été présentés dans le chapitre 3. Dans cette section, les résultats complémentaires seront présentés.

L'objectif de la régression logistique était d'identifier les éléments prosodiques permettant de séparer les deux groupes : FF et FQ. Cette séparation peut se faire sur la base de deux critères : l'origine réelle des locuteurs ou l'identification qui en a été faite dans l'accord interjuges.

En plus de la régression logistique, un autre test statistique de classification sera présenté ici : l'arbre de classification. Celui-ci, tout comme la régression logistique,

crée un modèle cherchant à classer les énoncés en fonction de la variable indépendante. Cependant, contrairement à la régression logistique, les données sont présentées graphiquement, sous la forme d'un arbre de prise de décision. Avant chaque coupure (*split*), une variable et un seuil sont utilisés comme critères de prise de décision.

Dans ce cas-ci, deux arbres de classification ont été créés afin de jeter un éclairage différent sur les données. Le premier arbre se base sur la variable indépendante d'origine perçue. Il s'agit donc de modéliser le comportement des juges lors de leurs choix quant à l'origine géographique des locuteurs. L'arbre créé est assez complexe, comme on peut le voir dans la figure 4-1. Il compte quinze coupures (moments de décision) et seize nœuds terminaux. Chaque nœud terminal indique le nombre d'énoncés classés à l'aide de ce « parcours ». Chaque coupure indique la variable et le seuil ayant permis de faire la classification.

Dans le cadre de ce mémoire, cette classification est intéressante puisqu'il s'agit d'un second modèle venant sélectionner des variables pertinentes lors de l'identification de l'origine géographique, le premier étant la régression logistique.

Une observation plus détaillée de l'arbre de classification pour l'origine perçue permet, au premier coup d'œil, de confirmer l'importance du paramètre acoustique de la F0. En effet, neuf des quinze coupures sont relatives à la F0 : une pour le registre, deux pour la différence de F0 entre les tons hauts et les tons non marqués, une pour la différence de F0 entre les tons bas et les tons non marqués et cinq pour la F0 moyenne.

De plus, contrairement à la régression logistique, cet arbre de classification semble accorder une importance relative à l'intensité. Finalement, les deux variables liées à la durée relevées par la régression logistique, la durée des voyelles et la différence de durée entre les tons hauts et bas, sont aussi présentes dans cet arbre.

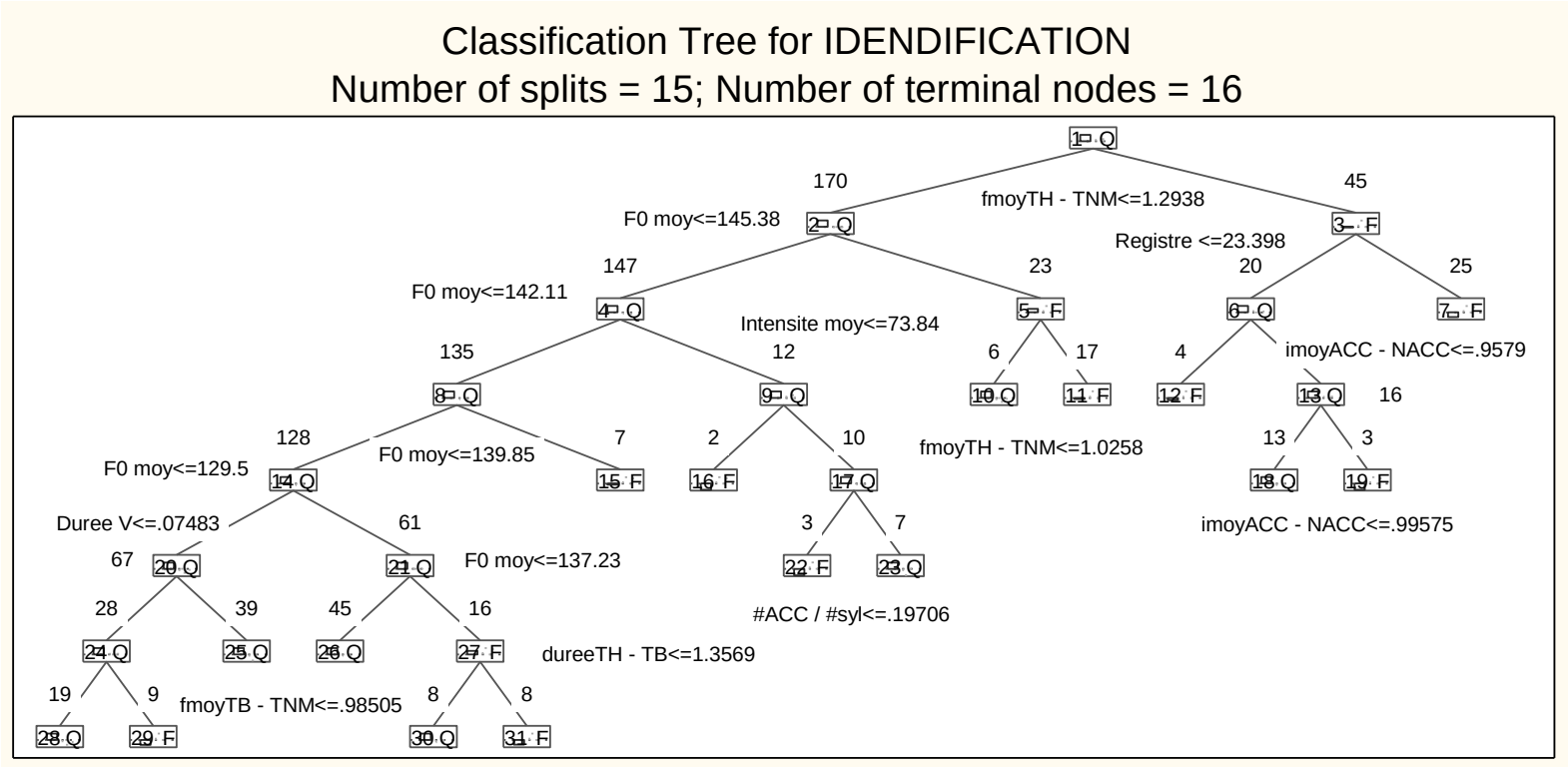


Figure 4-1 Arbre de classification pour l'origine perçue

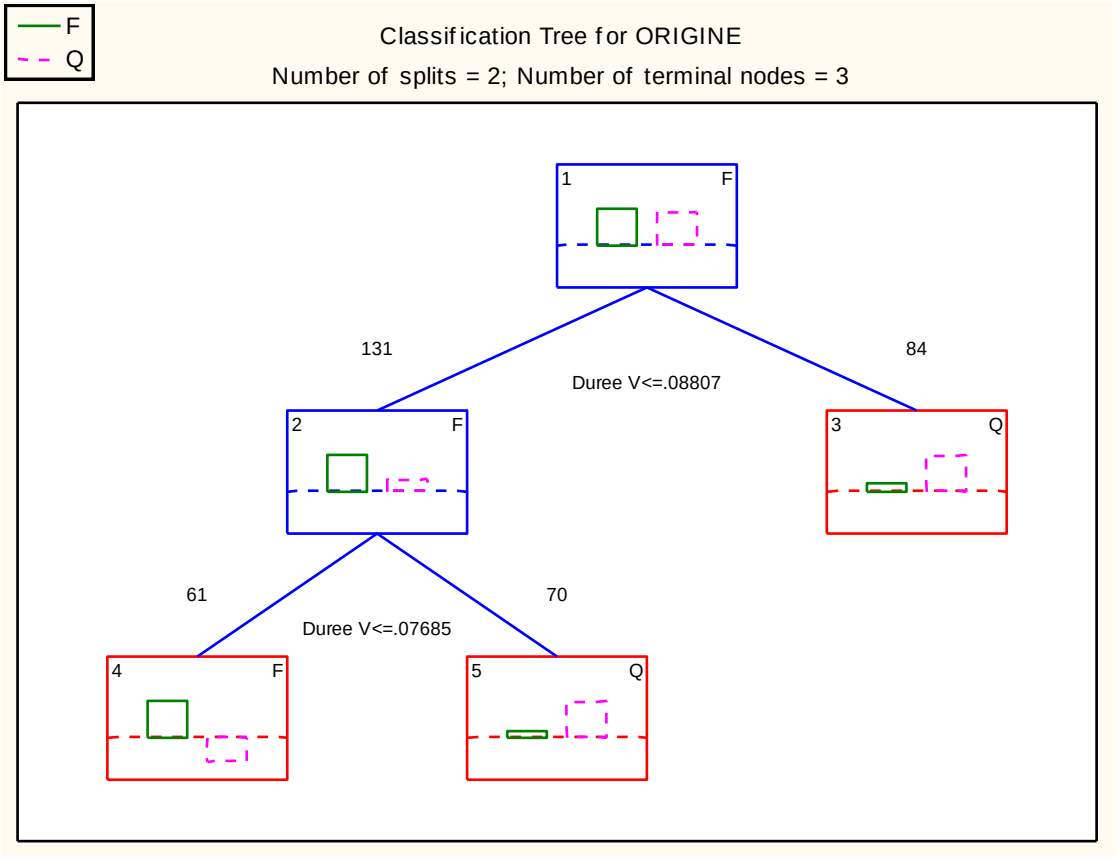


Figure 4-2 Arbre de classification pour l'origine réelle

L'arbre de classification se basant sur l'origine réelle, quant à lui, est beaucoup plus simple. Il est si simple, en fait, qu'il ne se base que sur une seule variable : la durée des voyelles.

Les données issues de la régression logistique sont aussi intéressantes à ce sujet. En effet, quatre variables ont été identifiées (tableau 3-6) : la durée des voyelles, la durée des syllabes, la durée des syllabes accentuées sur la durée des syllabes non accentuées et le nombre de syllabes accentuées sur le nombre de syllabes non accentuées. Il est intéressant de noter le peu de recoupement entre les variables issues des deux classifications.

4.3 Test de perception

Plusieurs manipulations statistiques ont été effectuées afin de bien analyser les résultats au test de perception. Dans un premier temps, les résultats généraux (moyennes et écart types) seront présentés. Ensuite, deux séries d'ANOVAS ont été effectuées afin d'évaluer les neuf séries de phrases. D'abord elles sont évaluées par rapport à l'ensemble des résultats, puis la cohérence interne de chaque série est revue. Enfin, des matrices de corrélations ont été effectuées et seront présentées.

4.3.1 Résultats généraux

Les résultats généraux sont d'abord présentés ici. Ils peuvent être visualisés dans le tableau 4-2. Chaque ligne du tableau correspond à une des neuf séries étudiées. Les colonnes sont divisées selon les cinq critères d'évaluation au test : l'intelligibilité, l'origine géographique et la qualité, comprenant la qualité générale (MOS), le confort d'écoute et le naturel. Pour l'intelligibilité, les scores sont ramenés sur 10, 0 voulant dire une réponse vide (ou aucun rappel) – 10 étant une réponse parfaite. Pour la qualité, les scores vont de 1 à 10, 10 étant le meilleur résultat. Pour l'origine, un score de 1 indique une origine géographique perçue plus française, alors que 10 indique une origine géographique perçue plus québécoise. Pour chacun de ces cinq critères d'évaluation, trois résultats sont présentés : la moyenne, l'écart type, et le niveau de significativité. Celui-ci a été calculé à l'aide d'une analyse de variance univariée (ANOVA) dans le logiciel Statistica . Le seuil de significativité retenu est de $p < 0.05$. Ainsi, les résultats présentés en gras sont significatifs. De plus, tous les graphiques issus des calculs d'ANOVA sont présentés dans l'appendice I. Les ANOVAs ont été faites de manière à comparer chaque série par rapport à l'ensemble des résultats.

Tableau 4-2 Résultats généraux du test de perception

	Origine			MOS			Confort d'écoute			Naturel			Intelligibilité		
	Moy	ÉT	P	Moy	ÉT	P	Moy	ÉT	P	Moy	ÉT	P	Moy	ÉT	P
F0 moyenne	5.02	1.69	0.464	5.16	2.25	0.084	4.47	1.91	0.483	**4.30**	**2.18**	**0.036**	**9.49**	**0.98**	**0.003**
Registre	5.37	1.80	0.078	4.69	2.28	0.417	4.39	1.90	0.835	3.87	1.99	0.632	9.06	1.34	0.270
SylDur TH/TB	5.17	1.77	0.741	4.87	2.32	0.857	4.50	1.90	0.343	4.12	2.01	0.324	9.34	1.06	0.226
F0 TH/TB	5.27	1.61	0.307	4.94	2.38	0.591	4.36	1.95	1	4.01	2.07	0.715	9.01	1.50	0.054
Duree V	**5.42**	**1.60**	**0.033**	5.10	2.33	0.170	4.47	1.81	0.452	4.19	2.15	0.152	**9.43**	**1.15**	**0.021**
Duree	5.15	1.78	0.864	**4.46**	**2.08**	**0.041**	4.27	1.71	0.561	3.73	1.98	0.172	**8.96**	**1.46**	**0.024**
F0	**4.83**	**1.73**	**0.035**	5.07	2.50	0.225	4.33	1.86	0.852	3.96	2.04	1	9.21	1.27	0.973
Duree-F0	5.05	1.70	0.604	4.99	2.24	0.438	4.43	1.90	0.649	3.90	2.09	0.763	9.34	1.11	0.398
Duree-F0 inverse	**4.84**	**1.70**	**0.039**	**4.29**	**2.17**	**0.003**	**4.01**	**1.68**	**0.021**	**3.50**	**1.91**	**0.006**	**8.98**	**1.44**	**0.045**
Total	5.12	1.72		4.84	2.30		4.36	1.85		3.95	2.05		9.20	1.28	

De nombreuses informations peuvent être tirées de ces résultats. Dans un premier temps, les résultats associés à chaque tâche - origine, qualité, intelligibilité - seront présentés. Dans un second temps, les résultats associés aux neuf séries de variables acoustiques seront analysés.

4.3.1.1 Origine

Pour ce qui a trait à l'origine, un résultat de 1 indique une origine FF alors qu'un score de 10 indique une origine FQ. En observant les résultats du tableau 4-2, on remarque d'abord que la moyenne globale se situe à 5.12. Ce résultat est étonnant parce que la voix de synthèse utilisée était une voix de synthèse québécoise, incluant donc des éléments segmentaux typiquement québécois (affrication, relâchement des voyelles hautes, voyelles nasales [e)]). Il est possible que les participants aient simplement utilisé le centre de l'échelle comme point de référence, penchant d'un côté ou de l'autre selon le fait que l'origine québécoise soit plus ou moins marquée.

Les résultats montrent, de plus, que trois variables entraînent des résultats significativement différents lorsqu'il est question de l'origine : la durée des voyelles, la série F0 combinée et la série Durée et F0 inversées.

Les modifications effectuées sur la durée des voyelles, ainsi, entraînent une identification plus québécoise que la moyenne des résultats. Les modifications de F0 et de durée F0 inverse, quant à elles, entraînent une identification plus FF. Dans les deux cas, une analyse interne des séries, effectuée plus loin dans ce chapitre, permettra de jeter un éclairage plus important sur ces données.

En plus de ces résultats, il convient de présenter la distribution des réponses liées à l'origine. Celle-ci peut être vue dans la figure 4-3. On voit que les données suivent une courbe normale, avec la majorité des réponses entre quatre et six. On remarque ensuite que les participants avaient plus tendance à identifier les phrases comme étant

plus FF que FQ (il y a un poids plus important des données à gauche qu'à droite de la courbe).

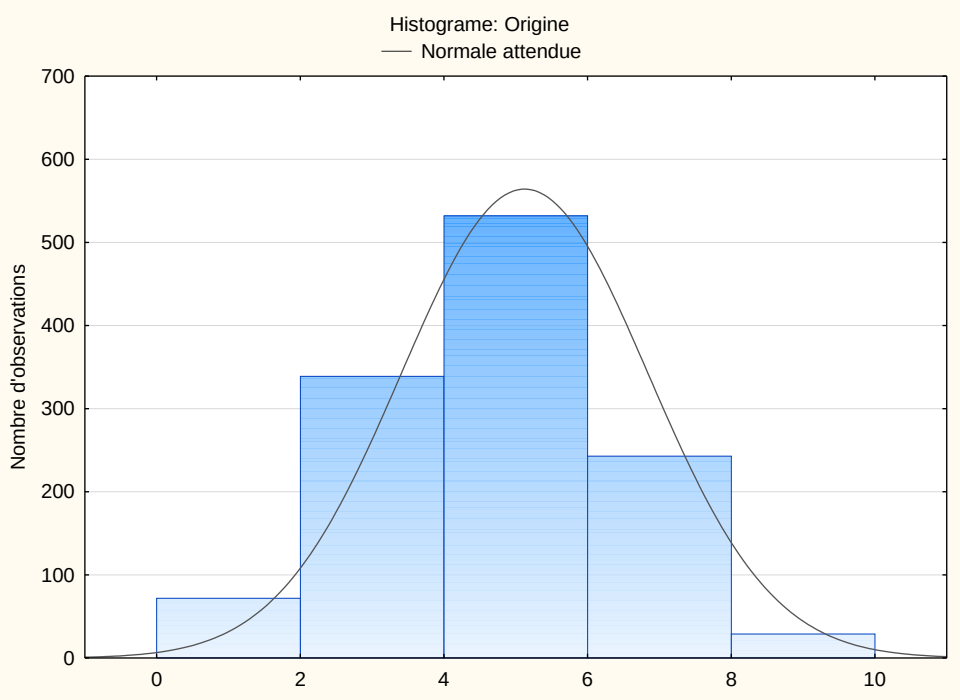


Figure 4-3 Distribution des données origine

Ce fait peut être expliqué en partie. En effet, suite aux tests, plusieurs participants ont mentionné à l'expérimentatrice qu'ils avaient tendance à identifier les phrases qu'ils ne jugeaient ni FF ni FQ du côté FF. Ainsi, l'objectif à atteindre, la « bonne réponse », selon eux, était l'identification FQ. Une voix qu'ils considéraient robotisée ou qui semblait avoir un accent « autre » avait été classée FF.

4.3.1.2 Qualité

La qualité, quant à elle, se composait de trois questions dans le test : le MOS (qualité générale), le naturel et le confort d'écoute (*pleasantness*). Les réponses à ces trois questions sont présentées dans le tableau 4-2. On remarque, d'abord, que les résultats sont généralement assez faibles - la moyenne étant en dessous du centre de l'échelle.

Ensuite, pour le MOS la moyenne se situe à 4.84. De plus, c'est la question ayant l'écart type le plus grand, à 2.30. Ceci indique une assez grande variabilité au niveau de cette réponse. Deux variables ont entraîné des changements significatifs pour cette réponse : la série durée et la série F0-durée inversée. Ces deux variables entraînent une baisse significative dans les résultats par rapport à l'ensemble des données – baisse qui sera vue en détail plus loin, dans la section 4.3.2.9.

La répartition des données pour le MOS peut être visualisée dans la figure 4-4. Les données sont réparties selon une courbe normale. Tout comme pour l'histogramme précédent, le poids de la courbe est un peu plus grand à gauche, avec un nombre plus important de données entre 0 et 4 qu'entre 6 et 10. Autrement dit, les participants avaient plus tendance à juger défavorablement les énoncés.

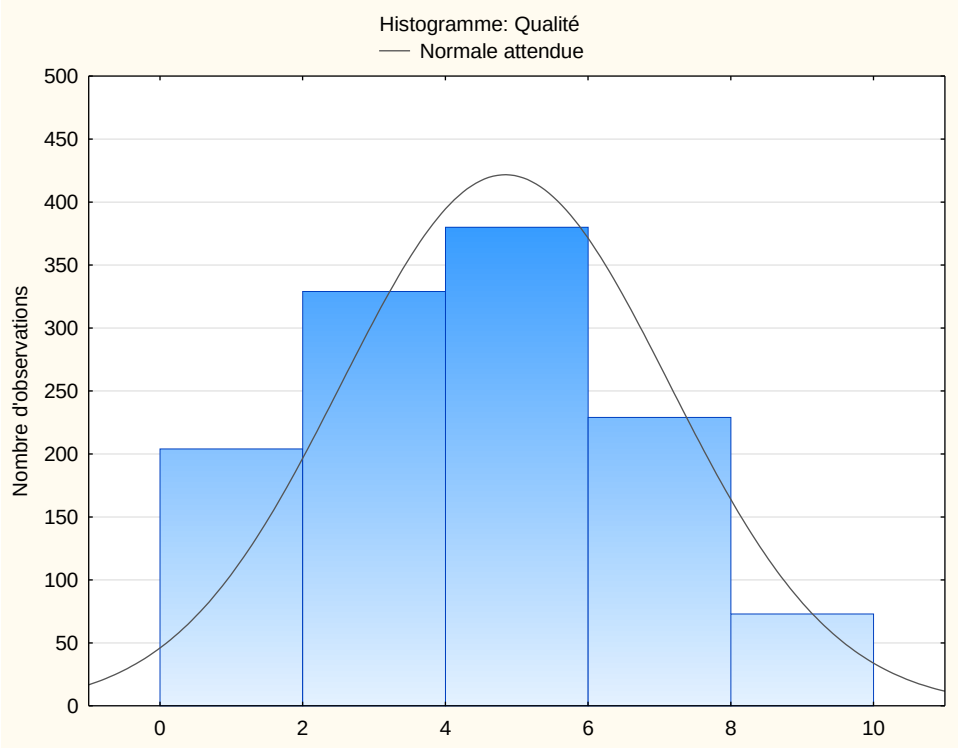


Figure 4-4 Répartition des données MOS

La seconde question liée à la qualité est celle portant sur le confort d'écoute, ou le *pleasantness*. Encore ici, la moyenne se situe en dessous du centre de l'échelle, à 4.36, avec un écart-type de 1.85. Seulement une série est statistiquement différente de l'ensemble des résultats : la série durée-F0 inversée qui est significativement en deçà de la moyenne.

La répartition des résultats pour cette question peut être vue dans la figure 4-5. Ces données suivent elles aussi une courbe normale. Cependant, le sommet de celle-ci se situe entre 2 et 4. Ainsi, le poids de la courbe se situe nettement à gauche, indiquant que les participants ont trouvé la voix très peu plaisante.

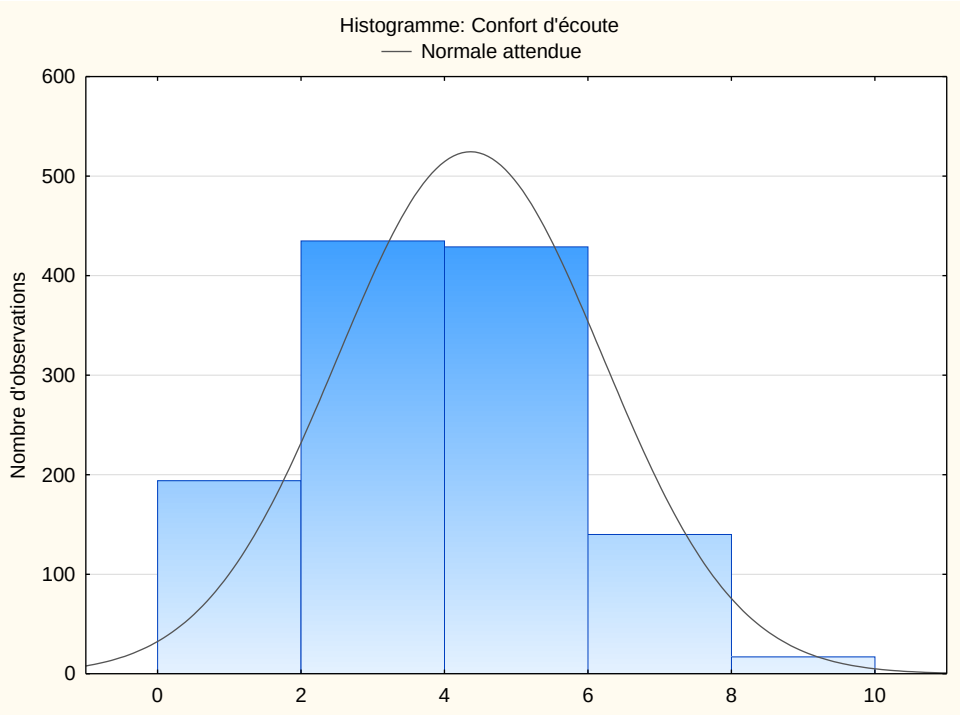


Figure 4-5 Répartition des données pour le confort d'écoute

La dernière question traitant de qualité est celle du naturel. La moyenne pour cette question est la plus faible du test, à 3.95, avec un écart-type de 2.05. De plus, une variable, la F0 moyenne, a eu un impact positif sur cette réponse par rapport au reste des données. Autrement dit, les modifications apportées à la F0 ont augmenté le

naturel de la voix. La série durée-F0 inversée a, encore une fois, entraîné des résultats plus bas que la moyenne pour cette question.

La distribution des données pour le naturel peut être vue dans la figure 4-6. Elles suivent une courbe normale, avec un sommet entre 2 et 4. Le poids de cette courbe se situe nettement à gauche, avec peu de réponses se situant entre 6 et 10. Ceci indique que les participants ont trouvé, d'une manière générale, que la voix manquait énormément de naturel.

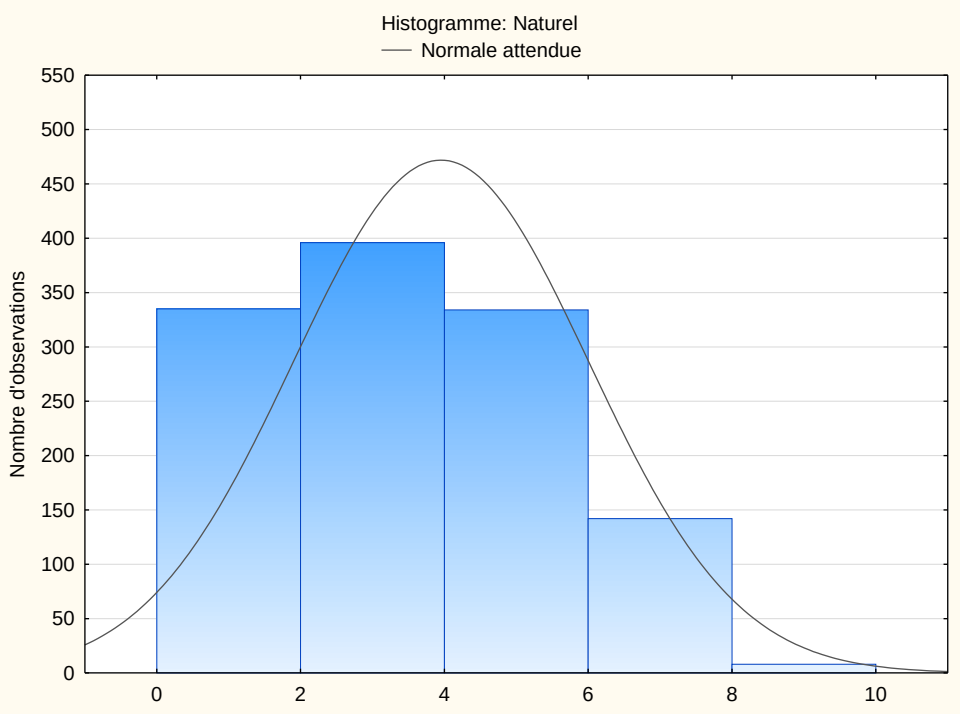


Figure 4-6 Répartition des données pour le naturel

4.3.1.3 Intelligibilité

Le dernier aspect du test était le test d'intelligibilité. Les résultats peuvent être vus dans le tableau 4-2. On remarque d'abord que la moyenne est très haute, à 9.20 sur 10. En effet, bien que des phrases SUS aient été utilisées et que les participants aient

rapporté avoir eu énormément de difficulté à percevoir les énoncés - la majorité de ceux-ci ont été transcrits convenablement.

Malgré tout, il est possible d'identifier des variables qui ont eu un impact significatif sur l'intelligibilité. La F0 moyenne et la durée des voyelles ont eu un impact positif, alors que la série durée et la série durée-F0 inversée ont eu un impact négatif sur les résultats.

La distribution des données d'intelligibilité peut être vue dans la figure 4-7. Cette courbe suit une distribution non normale (liliefors p <.01). En effet, la majorité des réponses se trouve à l'extrême droite de l'histogramme. Cette distribution nous amène à employer des tests non paramétriques pour cette variable lors des diverses analyses présentées dans ce chapitre.

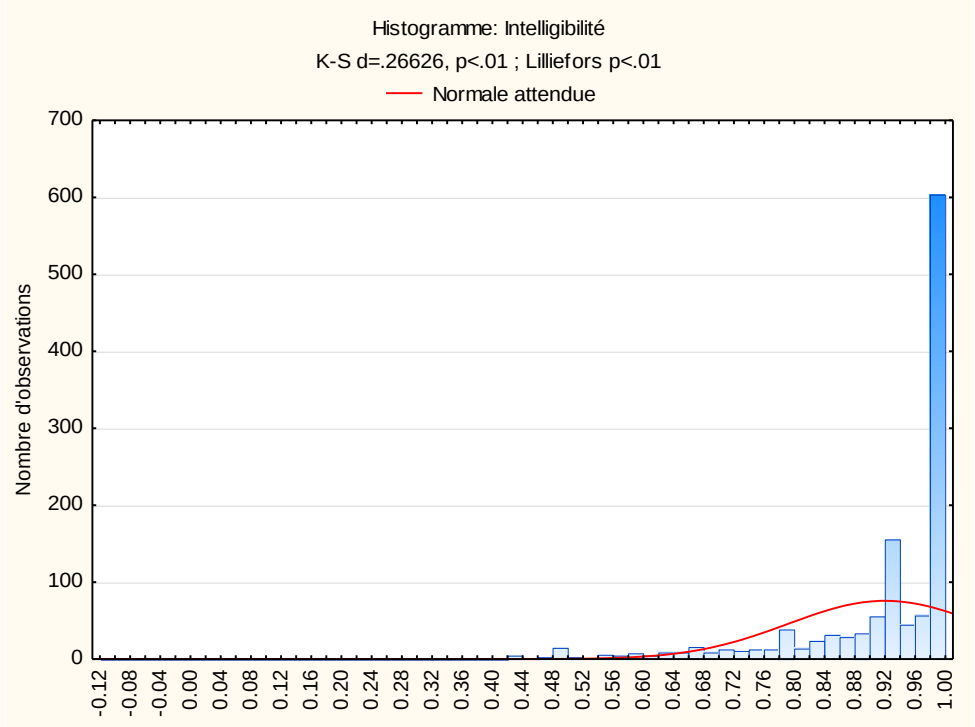


Figure 4-7 Répartition des données pour l'intelligibilité

4.3.2 Cohérence interne des séries

Après l'analyse des résultats généraux, la cohérence interne de chaque série a été évaluée. Le lecteur se souviendra que neuf séries d'énoncés ont été synthétisés : cinq pour étudier les paramètres acoustiques de manière isolée, quatre pour étudier les paramètres acoustiques de manière combinée. De plus, pour chaque paramètre acoustique, trois seuils ont été créés – FF, N et FQ. Les tableaux suivants présentent les différences existant, pour chaque série et chaque question, entre ces différents seuils. Chaque tableau présente une série différente. Les différentes questions présentées lors du test sont sur le plan horizontal, alors que les seuils sont sur le plan vertical, de même que les résultats totaux et le seuil de significativité (p). Les résultats totaux sont en fait un rappel du tableau 4-2. Les résultats statistiquement significatifs à $p < 0.05$ sont ombragés.

4.3.2.1 Série 1 : F0 moyenne

Dans le cas de notre première série, la F0 moyenne, on n'observe pas de différence significative entre les trois seuils, et ce, pour les cinq questions. Les seuils étaient les suivants : FF - 150Hz, N 135 Hz et FQ 120 Hz, donc du plus aigu au plus grave.

Tableau 4-3 Cohérence interne des séries : F0 moyenne

Série 1 : F0 moyenne										
	Origine		**MOS**		**Confort**		**Naturel**		**Intel**	
	Moy	ET	Moy	ET	Moy	ET	Moy	ET	Moy	ET
p	0.424		0.398		0.926		0.126		0.119	
FF - 150 Hz	4.93	1.79	5.20	2.23	4.53	2.07	3.78	2.01	9.25	1.19
N - 135 Hz	4.84	1.65	4.82	2.20	4.49	1.94	4.47	2.33	9.48	1.09
FQ - 120 Hz	5.29	1.65	5.47	2.33	4.38	1.75	4.67	2.12	9.74	0.46
Total	5.02	1.69	5.16	2.25	4.47	1.91	**4.30**	**2.18**	**9.49**	**0.98**

Il est surprenant que cette variable ne soit pas significative, puisque la F0 a été identifiée comme un des marqueurs principaux de l'attribution de l'origine géographique, comme on l'a vu plus haut dans les arbres de décision. En regardant de plus près, on observe une tendance en ce sens au niveau de l'identification de l'origine (les énoncés plus aigus sont plus identifiés FF, les énoncés plus graves plus FQ) mais cette tendance est loin d'atteindre un niveau significatif. Une échelle plus grande, allant par exemple de 175 à 100 Hz, aurait peut-être généré des résultats différents.

Au niveau de la qualité (MOS, confort d'écoute, naturel), on remarque aussi que les écarts entre les trois seuils n'entraînent pas de différence significative au niveau des résultats. Seuls les résultats totaux pour le naturel sont significatifs. Ceci veut dire que les énoncés de cette série sont significativement plus naturels que l'ensemble des énoncés. En y regardant de plus près, on remarque que les énoncés des seuils N et FQ ont tous deux reçu un bon score au niveau du naturel. Ceci indique que les participants jugeaient que les voix plus graves – entre 135 et 120 Hz – étaient plus naturelles que les autres.

La valeur d'intelligibilité pour cette série, elle aussi, est significativement meilleure que pour l'ensemble des énoncés, avec un score moyen de 9.49 comparé à 9.20 pour l'ensemble des données. En y regardant de plus près, on observe la même tendance que plus le naturel : les voix plus graves obtiennent un meilleur score. En effet, on constate une progression de FF à FQ, bien qu'elle n'atteigne pas le seuil de la significativité ($p = 0.062$).

4.3.2.2 Série 2 : Registre

La seconde série, quant à elle, traitait du registre – la différence, en hertz, entre la valeur la plus haute (aigüe) et basse (grave) de F0. Les trois seuils étaient identifiés comme suit : FF à 120 Hz, N à 60 Hz et FQ à 10 Hz. Ces trois seuils indiquent qu'une voix plus FF a plus d'espace de modulations, entre l'aigu et le grave, alors qu'une voix plus FQ a très peu d'espace de variation d'intonation.

Ces seuils ont été obtenus suite à l'analyse acoustique des résultats de l'accord interjuges. Ainsi, il avait été identifié qu'un registre très petit – jusqu'à 10 Hz (appendice I) – était identifié comme québécois. Ce résultat était très surprenant, mais a néanmoins été utilisé dans le test de perception, précisément pour l'évaluer.

On peut observer les résultats dans le tableau 4-4. Trois questions ont des résultats significativement différents entre les trois seuils : origine, MOS et intelligibilité. Aucune question n'avait de résultat significatif par rapport à l'ensemble des données (ligne *Total*).

Tableau 4-4 Cohérence interne des séries : Registre

Série 2 : F0 registre										
	Origine		**MOS**		**Confort**		**Naturel**		**Intel**	
	Moy	ET	Moy	ET	Moy	ET	Moy	ET	Moy	ET
p	**0.041**		**0.034**		0.059		0.083		**0.001**	
FF – 120 Hz	5.20	1.60	5.13	2.12	4.62	1.91	4.13	2.01	9.23	1.01
N – 60 Hz	5.91	1.93	4.96	2.41	4.71	2.12	4.16	2.11	9.37	1.26
FQ – 10 Hz	5.00	1.77	3.98	2.18	3.84	1.57	3.33	1.78	8.57	1.57
Total	5.37	1.80	4.69	2.28	4.39	1.90	3.87	1.99	9.06	1.34

Sur le plan de l'origine, selon nos prévisions, on devrait s'attendre à une courbe linéaire, FQ ayant le score le plus élevé, FF le plus faible. Or, ce n'est pas ce que nous observons dans ce tableau. En effet, le seuil ayant été le mieux identifié comme québécois est celui qui avait été préalablement identifié comme neutre (N). Le seuil ayant été le plus identifié comme français est plutôt celui qui avait été prévu pour FQ. On peut voir une représentation graphique de cette courbe dans la figure 4-8.

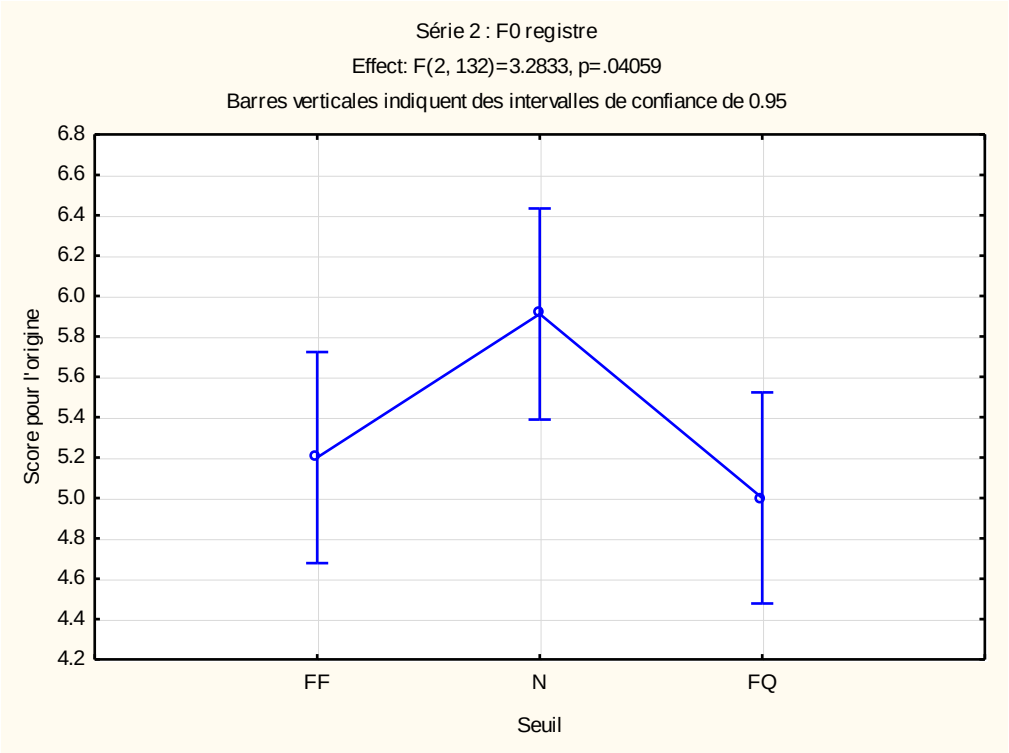


Figure 4-8 ANOVA cohérence interne de la série registre pour la question origine

Ces résultats peuvent apparaître surprenants mais s'expliquent probablement par le choix des seuils. En effet, un seuil de 10 Hz est non seulement très faible, il est certainement linguistiquement inapproprié dans le cas des phrases avec accent d'emphase ou pour les interrogations. Or, un registre de 60 Hz est un registre qui, bien que relativement réduit, reste dans des limites permettant la réalisation de ces phénomènes. Encore ici, un choix différent de seuils aurait peut-être contribué à une meilleure identification de l'origine.

Le fait qu'un registre de 10 Hz ait été préalablement identifié comme québécois lors de l'accord interjuges s'explique probablement par les énoncés utilisés lors de ce test. On observe notamment, dans le corpus initial, qu'il y avait de très courtes phrases, telles que « Il la mange ». Il est normal d'observer un registre moins étendu pour ce genre d'énoncés produits de manière neutre. Ainsi, dans le cas des courtes phrases,

des énoncés délexicalisés ayant un registre de 10 Hz ont été identifiés comme FQ lors de l'accord interjuges. Tenter d'appliquer ce registre à des phrases interrogatives plus longues pose certainement problème. En effet, il faut un espace spectral assez grand pour créer une courbe interrogative – 10 Hz sont nettement insuffisants.

De plus, si on considère que les résultats associés à un registre de 10 Hz sont, dans plusieurs cas, des erreurs linguistiques, on observe tout de même une tendance entre les deux seuils restants : FF et N. Dans ce cas, FF est correctement identifié comme tel et N est identifié comme FQ.

Les résultats de cohérence interne pour le MOS sont eux aussi significatifs. On observe ici un argument de plus en faveur de l'idée selon laquelle un registre de 10 Hz est linguistiquement inapproprié. En effet, on observe une baisse importante de l'appréciation entre les seuils N et FQ.

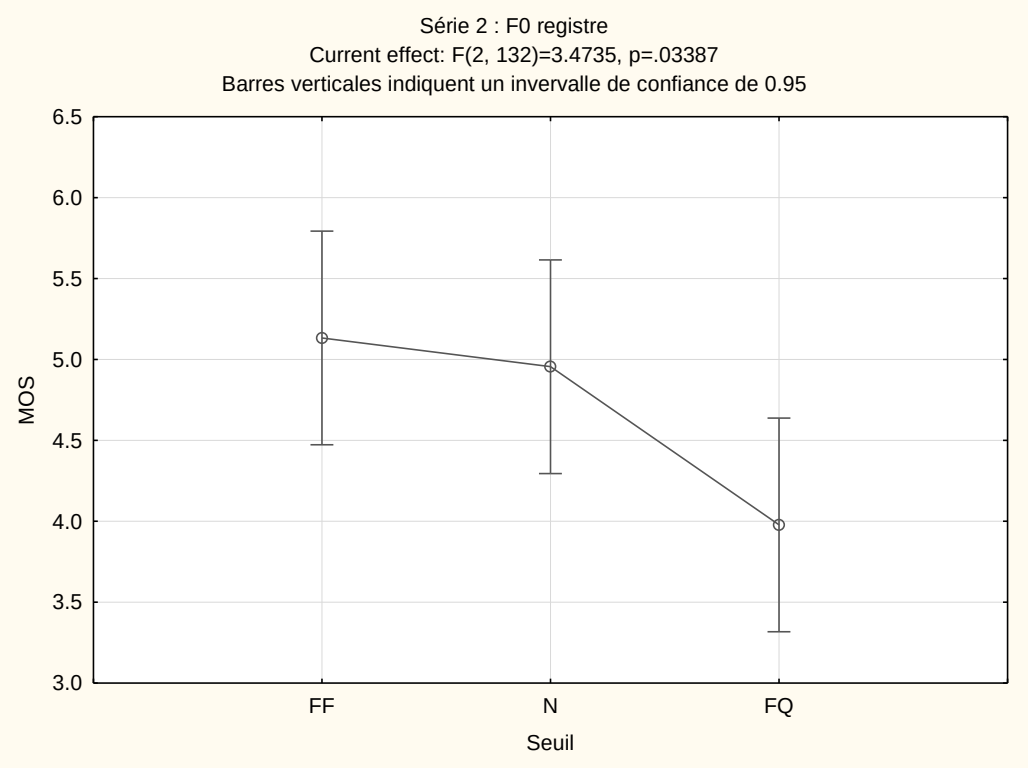


Figure 4-9 ANOVA cohérence interne de la série registre pour la question MOS

Néanmoins, contrairement à la question précédente, on observe une légère baisse de la qualité entre les seuils FF et N. Ceci semble indiquer qu'alors qu'un registre relativement étroit est associé à une identification FQ, c'est plutôt le registre plus large qui est le plus apprécié des participants. Cette différence est assez faible, mais mériterait d'être explorée plus en profondeur.

Finalement, pour la série registre, les résultats liés à l'intelligibilité sont également significatifs. On observe alors que les phrases ayant le seuil FQ – ou le seuil inapproprié - sont clairement moins intelligibles que les autres. Il est également pertinent de souligner le fait que les énoncés ayant un seuil N sont légèrement plus intelligibles que celles qui ont un seuil FF. On observe donc la même forme de courbe que pour l'origine, c'est-à-dire avec un sommet au centre, comme on peut le voir dans la figure 4-9.

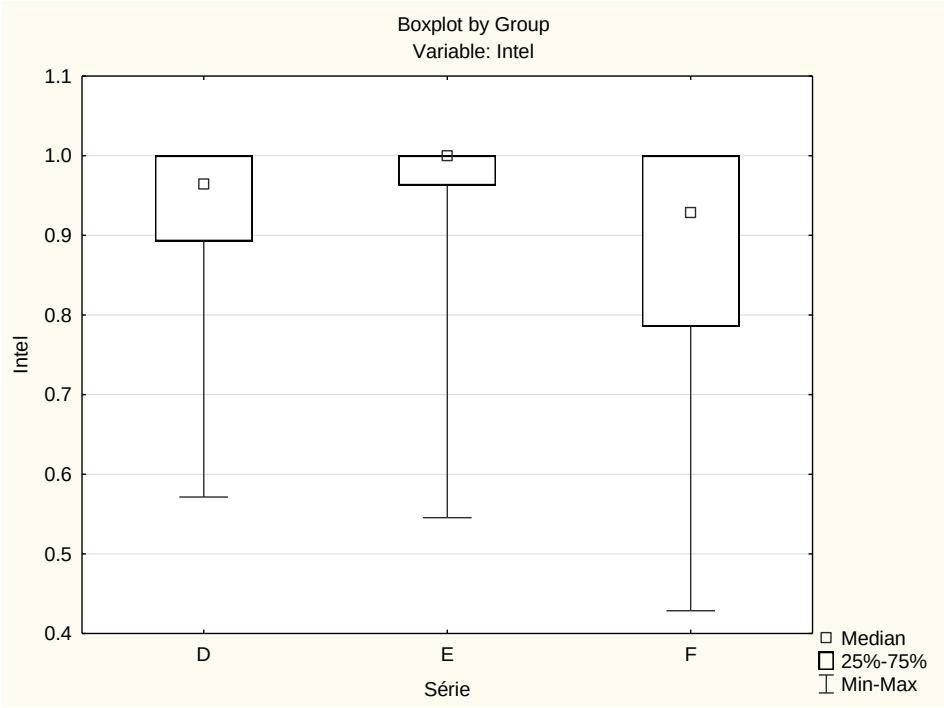


Figure 4-10 Kruskal-Wallis cohérence interne de la série registre pour la question intelligibilité

Ces résultats sont assez significatifs. Ils indiquent que les énoncés identifiés par les participants comme FQ sont plus intelligibles que les autres. Ces résultats, mis en relation avec le MOS, sont particulièrement intéressants. D'un côté, les phrases identifiées FQ sont plus intelligibles. Néanmoins, les participants préfèrent légèrement les énoncés identifiés FF. Il sera pertinent de vérifier si cette tendance se confirme parmi les autres séries.

4.3.2.3 Série 3 : Durée Syllabes TH/TB

La troisième série, quant à elle, traite de la différence de durée entre les syllabes étiquetées par un TH et celles étiquetées par un TB. Trois seuils avaient été choisis : un ratio de 1.5 pour FF, 1.3 pour le seuil neutre et 1 pour FQ. Ceci indique que, selon nos analyses statistiques liées à l'accord interjuges, un ratio plus grand entre les durées des TH et TB indiquerait une origine plus française.

Tableau 4-5 Cohérence interne de la série sylDur TH/TB

Série 3 : sylDur TH/TB										
	Origine		**MOS**		**Confort**		**Naturel**		**Intel**	
	Moy	ET	Moy	ET	Moy	ET	Moy	ET	Moy	ET
p	0.845		0.155		0.928		0.363		0.439	
FF r = 1.5	5.04	1.54	4.98	2.24	4.42	1.83	3.84	1.98	9.51	0.60
N r = 1.3	5.24	1.96	5.29	2.35	4.58	1.90	4.44	2.12	9.51	0.91
FQ r = 1	5.22	1.86	4.36	2.36	4.51	2.02	4.07	1.95	9.02	1.44
Total	5.17	1.77	4.87	2.32	4.50	1.90	4.12	2.01	9.34	1.06

On remarque qu'aucune variable n'atteint le seuil de la significativité. Au plan de l'origine, bien que les différences ne soient pas significatives, on remarque que le seuil FF est effectivement celui qui a été le plus reconnu comme tel par les

participants. Les deux autres seuils, N et FQ, ont été reconnus à des niveaux similaires, relativement FQ.

Pour l'intelligibilité, on observe une différence entre les deux premiers seuils, très intelligibles, et le troisième qui se situe légèrement en dessous de la moyenne pour l'ensemble des énoncés (9.20 - tableau 4.2). Un ratio de 1 indique qu'il n'y a pas de différence de durée entre les TH et TB. Ceci affecte visiblement l'intelligibilité, comme on peut le voir dans la figure 4-11.

Considéré en parallèle avec les résultats liés à l'origine géographique, ce résultat est assez intéressant. En effet, une différence peu importante entre les durées des syllabes TH et TB est reconnu comme FQ. Néanmoins, un ratio de 1, ou l'absence de différence de durée entre les TH et les TB, entraîne une perte de l'intelligibilité.

En observant le reste des résultats – ceux liés à la qualité, bien qu'ils n'atteignent pas le seuil de significativité, on réalise que le seuil favorisé par les participants est le seuil N. Ceci indique qu'il doit y avoir une différence dans la durée des syllabes entre les TH et les TB, puisqu'une absence de variation entraîne une perte d'intelligibilité et une légère diminution de la qualité (MOS, confort et naturel). Néanmoins, cette différence ne doit pas être trop grande. En effet, un ratio de 1.5 entraîne une légère diminution des résultats par rapport à un ratio de 1.3. Autrement dit, un ratio plus élevé – ou plus FF – entraîne une légère baisse de qualité et d'intelligibilité.

4.3.2.4 Série 4 : F0 TH/TB

La quatrième série porte sur la différence de F0 entre les tons hauts et bas. Comme pour la série précédente, chaque seuil est un ratio. Le ratio identifié FQ est de 1 (ou pas de différence), le ratio N de 1.5 et le ratio FF est de 2 (TH deux fois plus aigu que TB). Les résultats sont présentés dans le tableau 4-6.

Tableau 4-6 Cohérence interne des séries : F0 TH/TB

Série 4 : F0 TH/TB										
	Origine		**MOS**		**Confort**		**Naturel**		**Intel**	
	Moy	ET	Moy	ET	Moy	ET	Moy	ET	Moy	ET
p	**0.04**		0.203		0.184		0.108		0.462	
FF r = 2	4.80	1.63	4.42	2.31	3.98	1.92	3.49	1.87	8.76	1.83
N r = 1.5	5.36	1.68	5.18	2.42	4.38	1.99	4.22	2.17	9.31	0.78
FQ r = 1	5.64	1.43	5.22	2.38	4.73	1.89	4.33	2.09	8.96	1.66
Total	5.27	1.61	4.94	2.38	4.36	1.95	4.01	2.07	9.01	1.50

La seule donnée significative est celle de l'origine : la différence entre les trois seuils est significative. On observe que, tel que prévu, un ratio de 2 reçoit effectivement un jugement d'origine plus FF alors qu'un ratio de 1 reçoit un jugement plus FQ, tel qu'on peut le voir dans la figure 4-12.

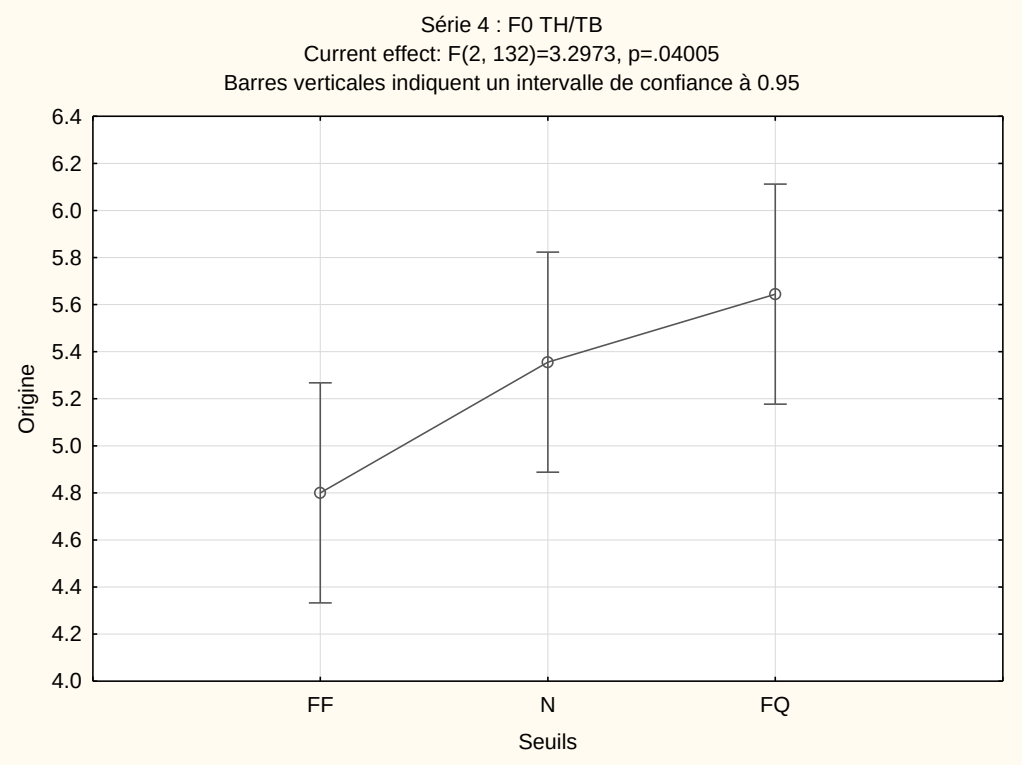


Figure 4-11 ANOVA cohérence interne de la série F0 TH/TB pour la question origine

Le choix de mettre un des seuils à un ratio de 1 était une décision audacieuse. En effet, ce qui caractérise la différence entre les tons hauts et bas, de manière empirique, ce sont les différences de F0. Or, on observe ici que cette absence de différence a plutôt guidé les participants dans leur identification de l'origine. Une absence de différence de F0 entre les TH et TB était associée clairement à une identification québécoise. De plus, ce ratio de 1 entraîne une légère baisse de l'intelligibilité par rapport aux énoncés N. Autrement dit, l'absence de saillance au niveau de la F0 entre les TH et TB entraîne une diminution – non significative – de l'intelligibilité. Du point de vue de la qualité, par ailleurs, ce ratio de 1 entraîne plutôt une légère hausse de la qualité pour les trois indicateurs.

Or, si on compare ces résultats avec ceux de la série précédente, on réalise que, dans ce cas, le seuil de 1 avait un effet significatif sur l'intelligibilité, un léger effet négatif

sur la qualité et n'avait pas d'effet si important dans l'attribution de l'origine géographique. Or, ici, en plus d'avoir un impact clair sur l'attribution de l'origine géographique, le seuil de 1 entraîne une légère perte d'intelligibilité (non significative) et une tendance à la hausse non significative au niveau de la qualité. Autrement dit, l'absence de saillance entre les TH et TB entraîne une diminution de l'intelligibilité, tant pour la F0 que pour la durée des syllabes. Cependant, peut-être parce qu'ils sont identifiés plus FQ, ce seuil entraîne une légère hausse de qualité.

Ces résultats mériteraient à être explorés plus en profondeur. En effet, ils semblent révéler des différences dans la réalisation des tons hauts et bas en FQ, où les différences de durée pour les syllabes – peut-être plus que la F0 – auraient un impact important à jouer.

Au-delà de ce ratio de 1, les résultats entre FF et N suivent la tendance générale observée dans les résultats jusqu'ici : les résultats FF ont une qualité et une intelligibilité un peu moins grande que les résultats N.

4.3.2.5 Série 5 : Durée V

La cinquième série visait à moduler le dernier paramètre acoustique identifié dans la régression logistique : la durée des voyelles. Les données statistiques avaient aussi révélé que des voyelles plus courtes étaient identifiées plus FF alors que des voyelles plus longues étaient identifiées plus FQ. Ainsi, pour la durée moyenne des voyelles, le seuil FF était de 60 ms, le seuil N de 80 ms et le seuil FQ de 95 ms. Les résultats aux diverses questions pour cette série peuvent être vus dans le tableau 4-7.

Tableau 4-7 Cohérence interne des séries : durée des voyelles

Série 5 : durée V										
	Origine		**MOS**		**Confort**		**Naturel**		**Intel**	
	Moy	ET	Moy	ET	Moy	ET	Moy	ET	Moy	ET
p	0.449		0.557		0.627		0.562		0.384	
FF - 60 ms	5.18	1.54	5.11	2.33	4.62	1.83	4.18	2.03	9.35	1.24
N - 80 ms	5.51	1.66	5.36	2.43	4.53	1.87	4.44	2.31	9.51	1.17
FQ - 95 ms	5.58	1.60	4.82	2.24	4.27	1.75	3.96	2.13	9.42	1.06
Total	**5.42**	**1.60**	5.10	2.33	4.47	1.81	4.19	2.15	**9.43**	**1.15**

En observant le tableau, on remarque d'abord qu'aucune variable n'est significative dans la cohérence interne de la série, entre les trois seuils. Néanmoins, les données pour l'origine sont globalement supérieures à la moyenne pour l'ensemble des résultats (5.12, tableau 4-2). Ce résultat signifie que les énoncés ont été globalement plus identifiés comme étant FQ. Autrement dit, les seuils choisis font en sorte que même le seuil FF est considéré plus FQ que la moyenne des énoncés présents dans le test. En observant les données internes liées à l'origine, bien qu'elles ne soient pas significatives, on constate aussi cette tendance en faveur de FQ. Ainsi, les résultats au test de perception confirment que les voyelles plus longues sont jugées plus FQ.

Ces résultats indiquent aussi que les trois seuils choisis sont globalement FQ. Un seuil ayant une durée de voyelles plus petite aurait probablement entraîné une identification plus FF. Les résultats au test d'intelligibilité sont, eux aussi, globalement meilleurs que la moyenne (9.20 tableau 4-2).

4.3.2.6 Série 6 : Durée

Les quatre dernières séries ne reposent pas sur un paramètre acoustique particulier, mais plutôt sur une combinaison de plusieurs des cinq paramètres vus plus haut. La

première série de combinaisons visait à combiner les deux paramètres acoustiques liés à la durée : la durée des voyelles et la différence de durée des syllabes entre les TH et les TB. Les seuils sont les mêmes que pour le paramètre acoustique pris de manière isolée. Ainsi, le seuil FF aura une durée moyenne de ses voyelles de 60 ms et un ratio entre ses TH et TB de 2. Le seuil N aura des voyelles d'une durée moyenne de 80 ms et un ratio de 1.5, et le seuil FQ aura des voyelles de 95 ms et un ratio de 1. Les réponses au test pour cette série peuvent être vues dans le tableau 4-8.

Tableau 4-8 Cohérence interne des séries : durée

Série 6 : Durée										
	Origine		**MOS**		**Confort**		**Naturel**		**Intel**	
	Moy	ET	Moy	ET	Moy	ET	Moy	ET	Moy	ET
p	0.552		0.663		0.521		0.139		**0.008**	
FF	5.27	1.85	4.67	1.93	4.51	1.74	4.11	1.94	8.76	1.70
N	4.91	1.76	4.27	2.17	4.18	1.72	3.78	2.11	8.77	1.44
FQ	5.27	1.74	4.44	2.17	4.13	1.67	3.29	1.82	9.34	1.14
Total	5.15	1.78	**4.46**	**2.08**	4.27	1.71	3.73	1.98	**8.96**	**1.46**

Observons d'abord les différences entre la série et l'ensemble des données. Le MOS, dans un premier temps, présente un résultat significativement inférieur à la moyenne (4.84 tableau 4-2). En effet, malgré le fait que les deux résultats de durée pris indépendamment soient sensiblement dans la moyenne (4.87 pour duréeS TH/TB, 5.10 pour duréeV), leur combinaison entraîne une baisse significative dans l'appréciation des énoncés.

L'inadéquation entre les variables prises séparément et leur combinaison se reproduit au niveau de l'intelligibilité. En effet, les résultats pour la série durée sont significativement en dessous de la moyenne (9.20, tableau 4.2). Néanmoins, les

résultats d'intelligibilité pour les deux variables prises de manière isolée étaient significativement supérieures à la moyenne, 9.34 pour duréeS TH/TB et 9.43 pour duréeV.

Il semble donc y avoir une interaction entre ces deux variables qui entraîne une baisse de qualité et d'intelligibilité. Ce résultat met en lumière l'équilibre délicat existant entre ces divers paramètres prosodiques en parole naturelle.

Au plan de la cohérence interne, une variable se distingue, l'intelligibilité. Puisque les données liées à cette variable suivent une courbe non-paramétrique, un test statistique différent a été employé, l'ANOVA de Kruskal-Wallis. On peut voir une représentation graphique des résultats dans la figure 4-12. On y remarque ainsi que les valeurs de durée associées à FQ sont nettement plus intelligibles que les autres.

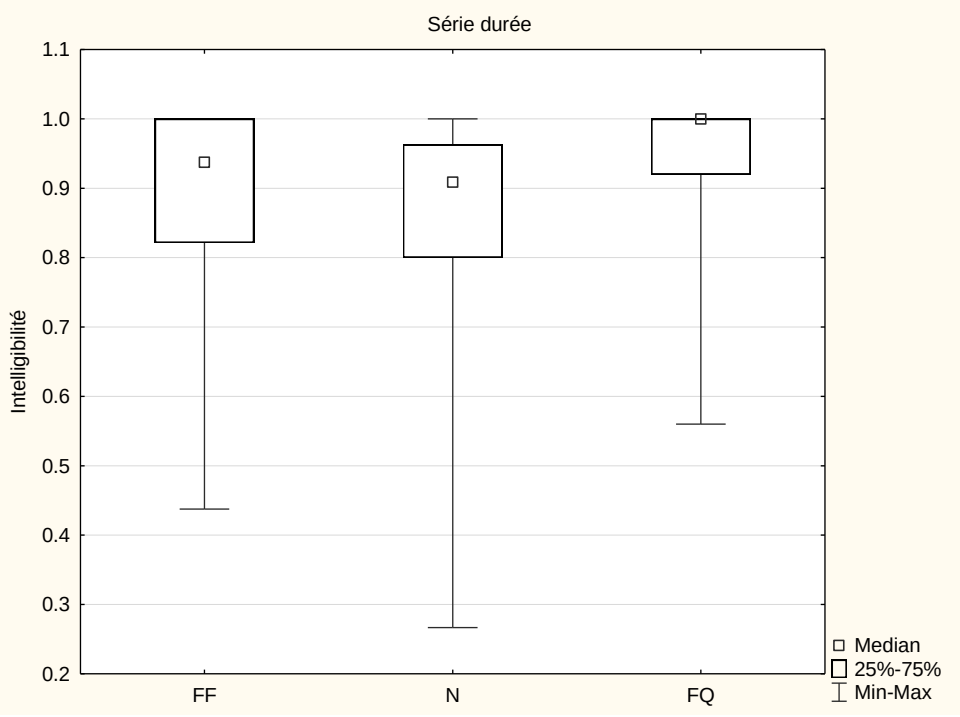


Figure 4-12 Anova Kruskal-Wallis pour la variable intelligibilité

4.3.2.7 Série 7 : F0

L'objectif de cette septième série était de combiner les trois paramètres acoustiques liés à la F0 : la F0 moyenne, le registre et la différence de F0 entre les TH et les TB. Les seuils ont été définis de la manière suivante :

- **FF** : F0 moy. = 150 Hz; registre = 120 Hz, F0 TH/TB r = 2,
- **N** : F0 moy. = 135 Hz, registre = 60 Hz, F0 TB/TB r = 1.5,
- **FQ** : F0 moy. = 120 Hz, registre = 10 Hz, F0 TH/TB r = 1.

Les résultats de cette série au test de perception peuvent être vus dans le tableau 4-9. On observe d'abord la très grande cohérence interne entre les seuils de cette série.

Tableau 4-9 Cohérence interne des séries : F0

Série 7 : FO										
	Origine		**MOS**		**Confort**		**Naturel**		**Intel**	
	Moy	ET	Moy	ET	Moy	ET	Moy	ET	Moy	ET
p	**0.009**		**0.004**		0.122		**0.026**		**0.000**	
FF	5.47	1.67	5.80	2.40	4.71	2.02	4.62	2.21	9.60	0.71
N	4.47	1.69	4.11	2.36	3.91	1.81	3.60	1.84	8.57	1.71
FQ	4.56	1.67	5.29	2.50	4.38	1.70	3.64	1.92	9.46	0.92
Total	**4.83**	**1.73**	5.07	2.50	4.33	1.86	3.96	2.04	9.21	1.27

Pour l'origine, d'abord, on remarque que, globalement, les résultats sont identifiés plus FF que la moyenne des énoncés (5.12 tableau 4-2). Or, en y regardant de plus près, on constate que cette identification FF ne correspond pas aux seuils identifiés. En effet, le seuil FF est identifié plus FQ que le seuil FQ - comme on peut le voir dans la figure 4-13.

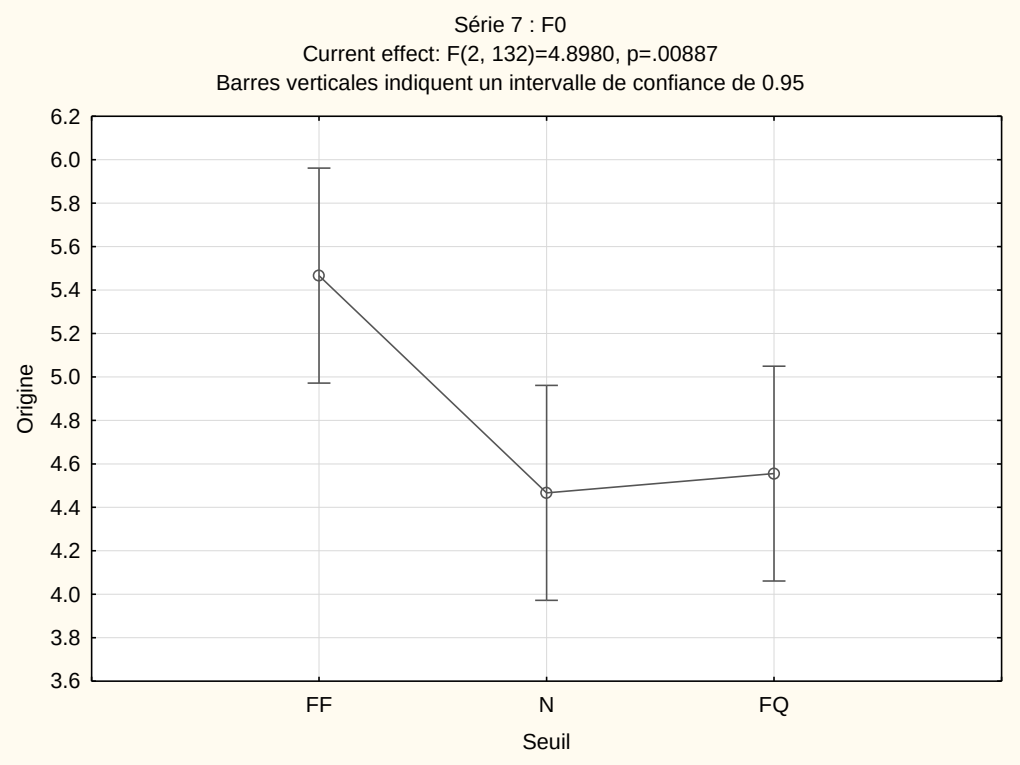


Figure 4-13 ANOVA cohérence interne F0 pour la question origine

Pris indépendamment, les résultats étaient pourtant dans la moyenne (F0 5.02, registre 5.37, F0 TH/TB 5.27) avec une cohérence interne en faveur du seuil FQ pour la F0 TH/TB et le registre. Les résultats présentés ici montrent donc qu'il existe une interaction entre ces trois variables venant affecter les jugements liés à l'origine.

Du point de vue de la qualité, deux variables sont significatives au niveau de leur cohérence interne : le MOS et le naturel. Les graphiques illustrant ces variables peuvent être vus dans les figures 4-14, pour le MOS, et 4-15 pour le naturel.

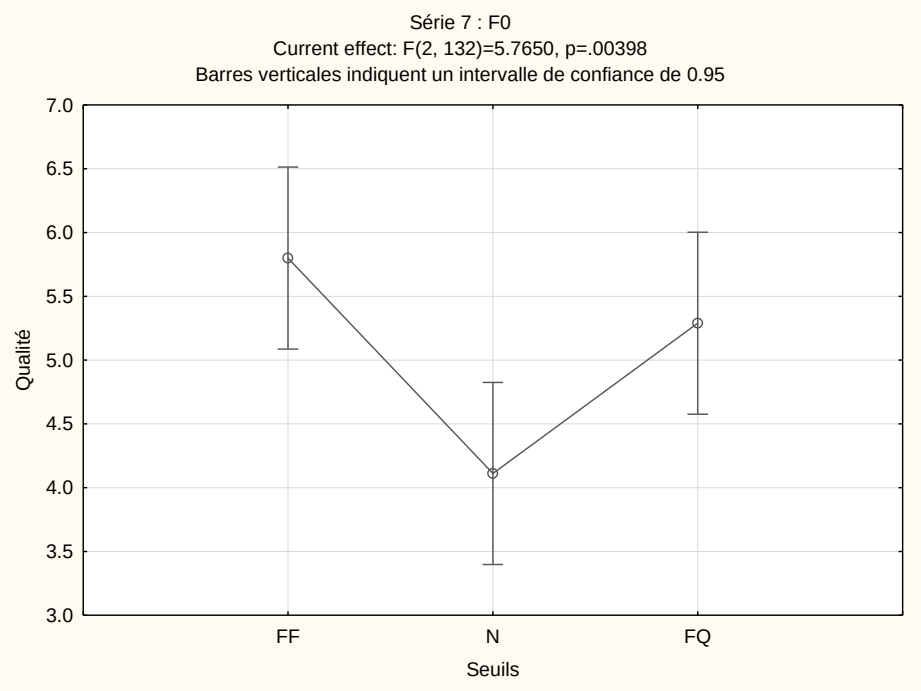


Figure 4-14 ANOVA cohérence interne série F0 : MOS

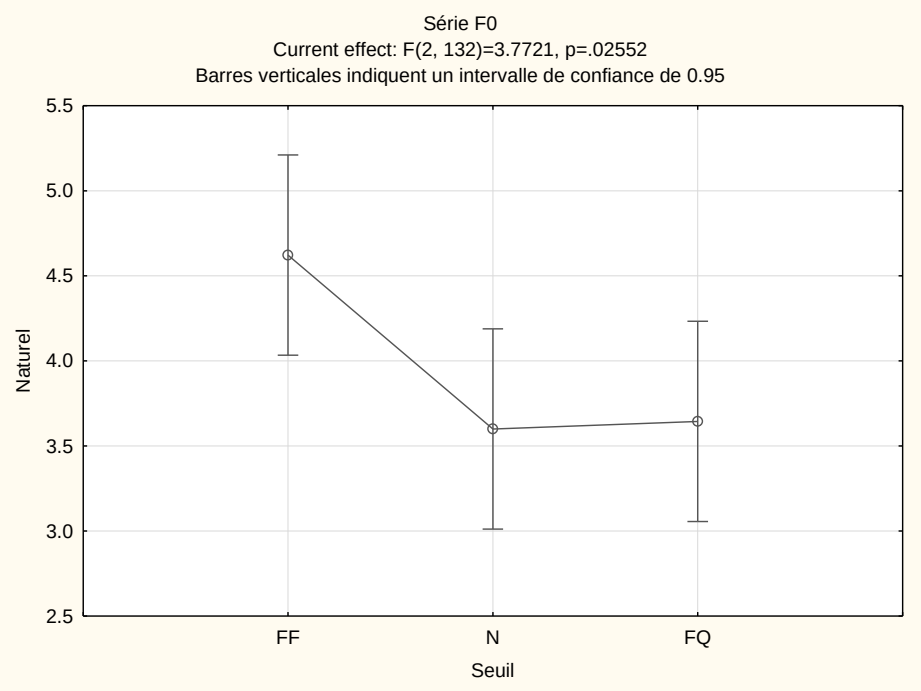


Figure 4-15 ANOVA cohérence interne série F0 : naturel

Dans un premier temps, on remarque que, tant pour le MOS que pour le naturel, c'est le seuil FF qui obtient les meilleurs résultats. Ensuite, dans le cas du MOS, le seuil N a les pires résultats alors que le seuil FQ a des résultats moyens. Pour le naturel, les résultats pour les seuils N et FQ sont semblables. Ainsi, non seulement le seuil FF était jugé FQ, il était aussi de meilleure qualité que les deux autres.

Du point de vue de l'intelligibilité, finalement, on observe que, encore, le seuil FF obtient les meilleurs résultats (figure 4-16). Le seuil N, quant à lui a, avec 8.57, un taux d'intelligibilité nettement en dessous de la moyenne (9.20 tableau 4.2).

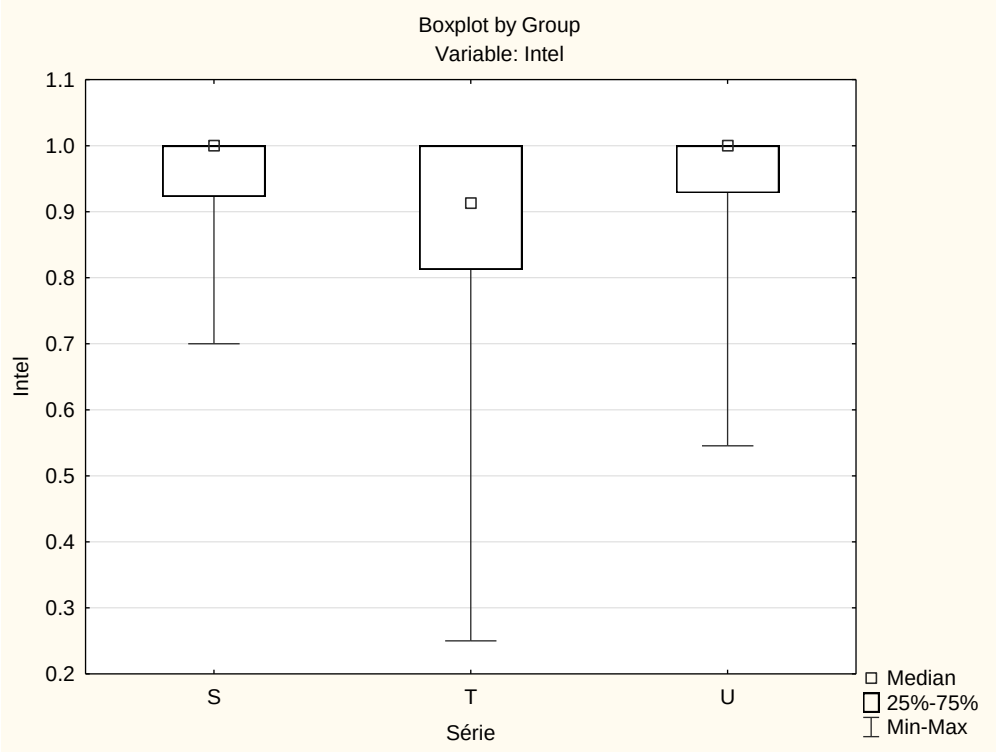


Figure 4-16 ANOVA Kruskal-Wallis pour la variable intelligibilité, série 7

Ainsi, les modifications combinées de F0 génèrent des résultats intéressants. Dans un premier temps, les seuils identifiés isolément comme FF deviennent, mis ensemble, un unique seuil identifié FQ. De plus, ce seuil obtient des résultats particulièrement élevés, tant au niveau de la qualité que de l'intelligibilité. De plus, la combinaison des

trois variables de F0 peut avoir un effet négatif tant sur la qualité que l'intelligibilité, comme on l'observe pour le seuil N.

4.3.2.8 Série 8 : F0 et durée

La huitième série visait à combiner les cinq paramètres acoustiques. Ainsi, les seuils sont définis tel que démontré dans le tableau ci-dessous.

Tableau 4-10 Seuils retenus pour la série 8

	FF	N	FQ
F0 moyenne	150 Hz	135 Hz	120 Hz
Registre	120 Hz	60 Hz	10 Hz
Durée V	60 ms	80 ms	95 ms
F0 TH/TB	2	1.5	1
Durée S TH/TB	1.5	1.3	1

Les données issues de cette série sont présentées dans le tableau 4-10.

Tableau 4-11 Cohérence interne des séries : F0 et durée

Série 8 : F0 et Durée										
	Origine		**MOS**		**Confort**		**Naturel**		**Intel**	
	Moy	ET	Moy	ET	Moy	ET	Moy	ET	Moy	ET
p	0.826		0.346		0.798		0.276		0.072	
FF	4.93	1.74	4.64	2.02	4.31	1.86	4.00	2.16	9.47	0.81
N	5.16	1.87	5.33	2.27	4.58	1.92	4.20	2.10	9.16	1.00
FQ	5.07	1.51	4.98	2.40	4.40	1.96	3.51	2.00	9.39	1.43
Total	5.05	1.70	4.99	2.24	4.43	1.90	3.90	2.09	9.34	1.11

On observe que cette série n'a engendré aucun résultat significatif. Autrement dit, la combinaison des cinq paramètres acoustiques à l'étude n'a d'effet ni positif ni négatif sur les résultats.

4.3.2.9 Série 9 : F0 et durée inversées

L'objectif de cette série était de présenter à nouveau les cinq paramètres acoustiques. Cependant, alors que l'ensemble des variables liées à la F0 étaient d'une origine, par exemple FF, l'ensemble des variables liées à la durée étaient de l'autre, soit FQ, et vice versa. Les seuils pour cette série peuvent être vus dans le tableau 4-12.

Tableau 4-12 Seuils pour la série F0 et durée inversées

	FF	N	FQ
F0 moyenne	150 Hz	135 Hz	120 Hz
Registre	120 Hz	60 Hz	10 Hz
Durée V	95 ms	80 ms	60 ms
F0 TH/TB	1	1.5	2
Durée S TH/TB	1.5	1.3	1

Les résultats issus de cette série sont présentés dans le tableau 4-13.

Tableau 4-13 Cohérence interne des séries : F0 et durée inversées

Série 9 : F0 et durée inversées										
	Origine		**MOS**		**Confort**		**Naturel**		**Intel**	
	Moy	ET	Moy	ET	Moy	ET	Moy	ET	Moy	ET
p	0.648		0.538		0.427		0.864		0.087	
FF	4.64	1.61	4.09	2.24	3.76	1.88	3.44	1.91	8.59	1.84
N	4.91	1.66	4.20	2.03	4.20	1.63	3.62	1.71	8.97	1.22
FQ	4.96	1.85	4.58	2.25	4.09	1.49	3.42	2.13	9.37	1.04
Total	**4.84**	**1.70**	**4.29**	**2.17**	**4.01**	**1.68**	**3.50**	**1.91**	**8.98**	**1.44**

Dans un premier temps, on observe que l'ensemble des résultats aux cinq questions est plus bas que la moyenne. Ce résultat était prévisible pour les questions liées à la qualité et l'intelligibilité. En effet, inverser ainsi les données liées à la durée et la F0 crée des énoncés linguistiquement inappropriés, ce qui a un impact négatif sur la qualité et l'intelligibilité.

Cependant, il était peu prévisible que ces manipulations aient comme effet d'entraîner un jugement d'origine plus FF. Il semble que les participants aient identifié les énoncés qu'ils jugeaient incorrects comme étant français.

Du point de vue de la cohérence interne, cette série a eu un effet mineur sur l'intelligibilité. Bien que cette série soit globalement moins intelligible que les autres, le seuil FQ a reçu un score supérieur aux deux autres seuils. Autrement dit, la combinaison particulière des paramètres acoustiques de ce seuil a produit des énoncés significativement plus intelligibles que pour les énoncés N et FF.

4.3.3 Matrice de corrélation

Il est toujours pertinent de vérifier les variables qui sont corrélées. Ainsi, une matrice de corrélation a été produite afin de déterminer si des variables étaient liées entre elles. Ces corrélations peuvent être observées dans le tableau 4-14. Les données en gras indiquent une corrélation significative à $p < 0.05$.

Tableau 4-14 Matrice de corrélation pour le test de perception

Corrélations significant at p < .05 N=1215							
	Moyennes	**Écart type**	**Origine**	**MOS**	**Confort**	**Naturel**	**Intel**
Origine	5.12	1.71	1.00	**0.11**	**0.30**	**0.20**	**0.15**
Qualité	4.84	2.29	**0.11**	1.00	**0.61**	**0.56**	**0.31**
Confort	4.36	1.84	**0.30**	**0.61**	1.00	**0.67**	**0.23**
Naturel	3.95	2.05	**0.20**	**0.56**	**0.67**	1.00	**0.27**
Intel	0.92	0.12	**0.15**	**0.31**	**0.23**	**0.27**	1.00

On remarque d'abord que toutes les variables sont corrélées entre elles. De plus, il s'agit toujours d'une corrélation positive. La plupart de ces corrélations étaient attendues. Par exemple, le fait que les trois variables liées à la qualité (MOS, confort d'écoute et naturel) soient corrélées se comprend aisément : ce sont trois variables qui se rapportent à un même phénomène global : la qualité. De plus, il est normal que l'intelligibilité et la qualité soient corrélées. En effet, une phrase peu intelligible sera forcément de moins bonne qualité. Inversement, une phrase très intelligible sera jugée plus favorablement.

Cependant, les corrélations liées à l'origine géographique méritent qu'on y porte une plus grande attention. En effet, ce que ces corrélations indiquent, c'est que plus une phrase est notée FQ, plus elle est intelligible et meilleure est sa qualité. Des graphiques de corrélation ont été créés en lien avec l'origine.

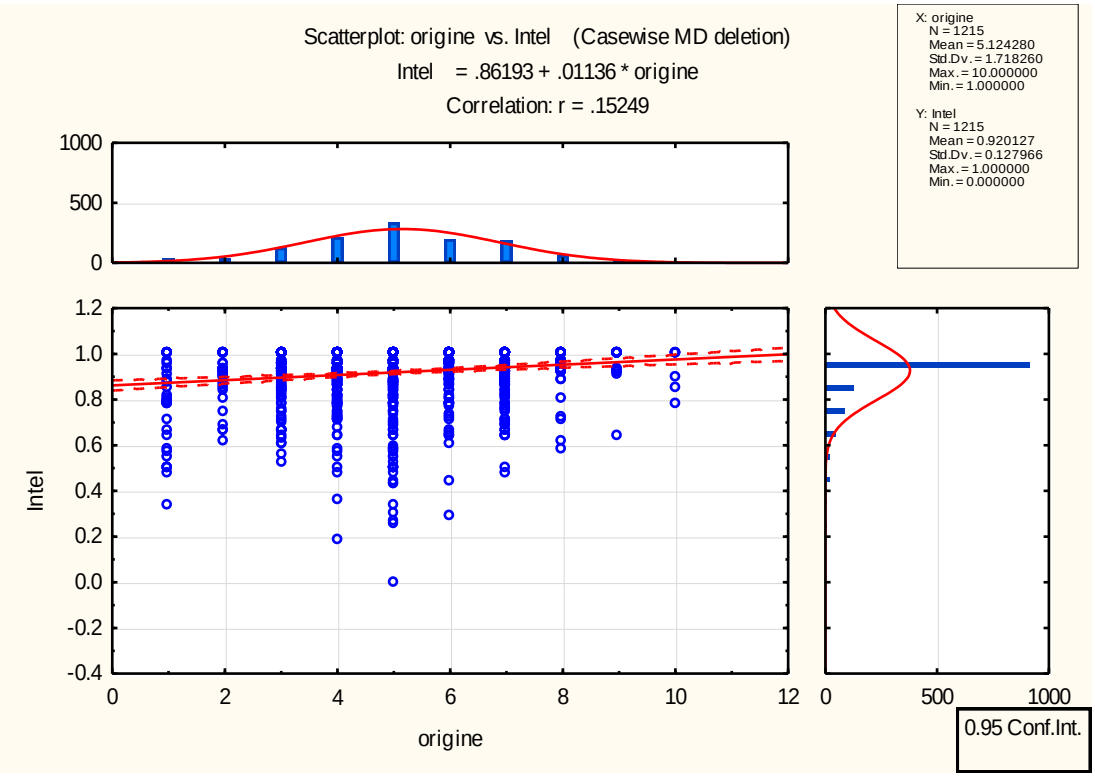


Figure 4-17 Corrélation origine vs. intel

Cette corrélation est intéressante puisqu'elle indique que les participants avaient un meilleur score au test d'intelligibilité pour les phrases qu'ils jugeaient plus FQ, ce qu'on peut voir dans la figure 4-17. Même si cette corrélation est assez faible ($r = .15$) elle est néanmoins significative à $p < 0.05$.

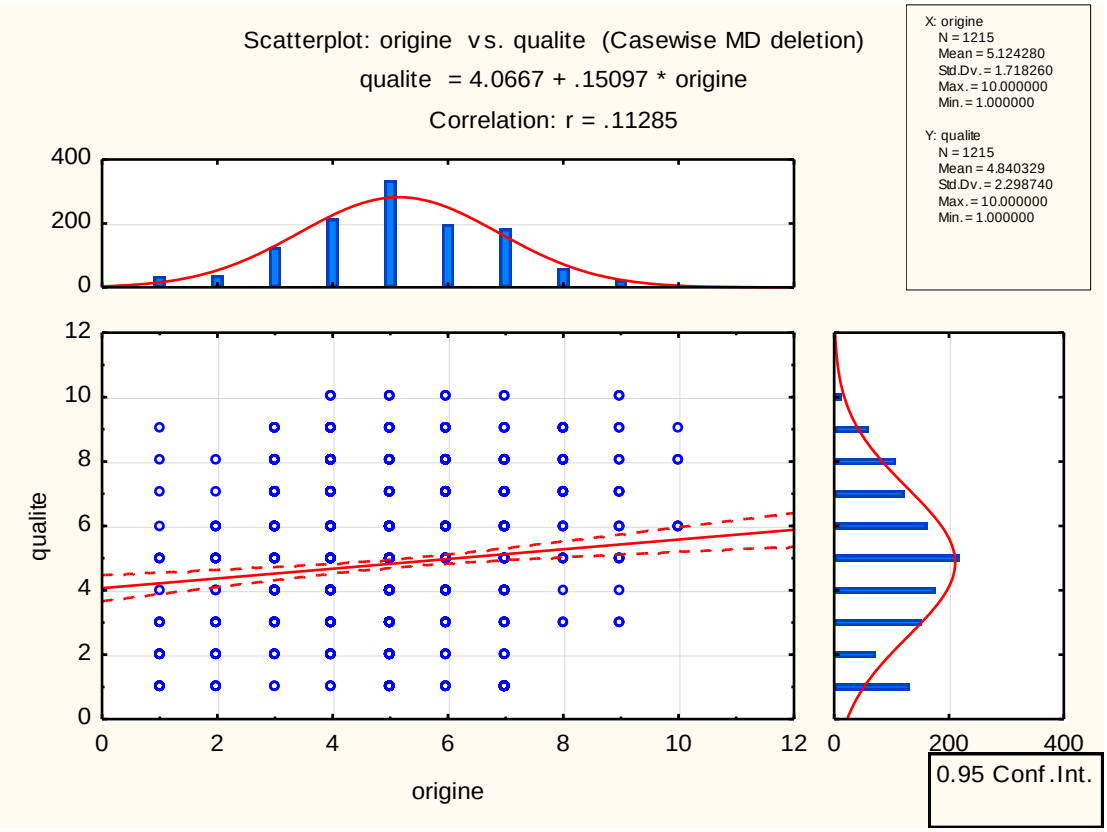


Figure 4-18 Corrélation MOS vs origine

La figure 4-18 présente la corrélation entre le MOS et l'origine. Il s'agit d'une corrélation plutôt faible (r = .11) mais statistiquement significative. Elle indique que plus une phrase était jugée FQ, plus les participants considéraient que sa qualité générale était élevée.

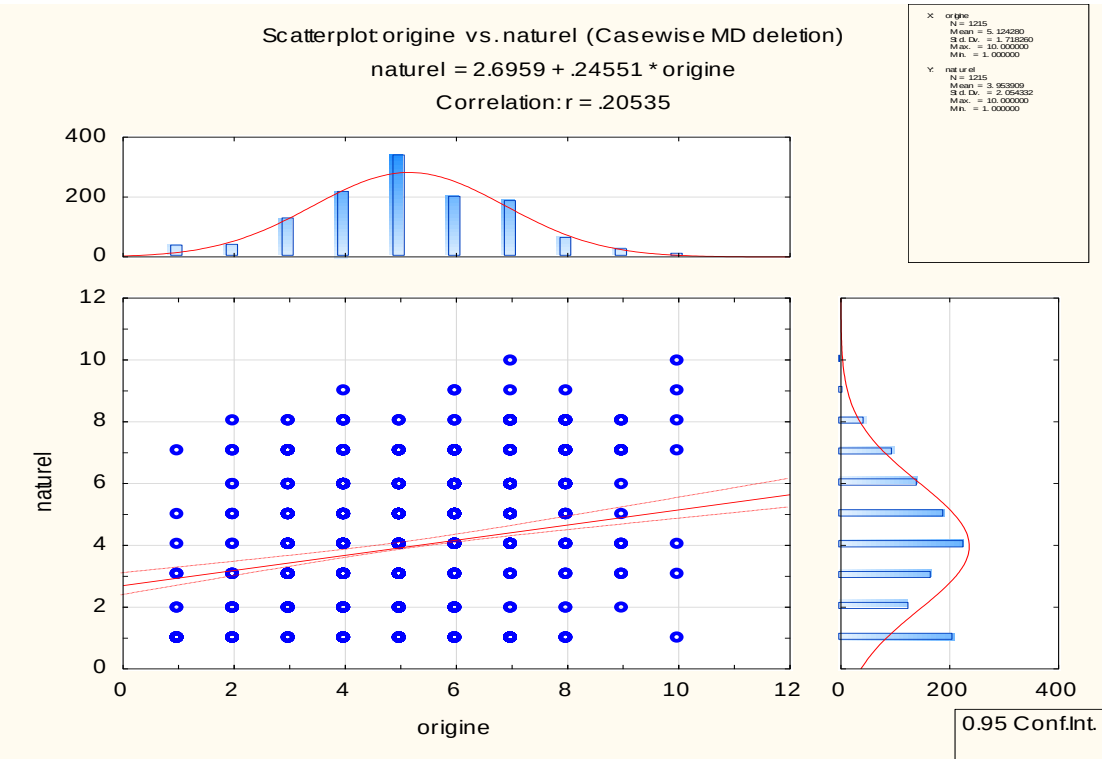


Figure 4-19 Corrélation naturel vs origine

La figure 4-19, finalement, présente la corrélation entre le naturel et l'origine. Elle est plus élevée que les deux dernières, à r = .20. Cette corrélation était assez prévisible. En effet, il est normal que des participants québécois jugent qu'un énoncé FQ soit plus naturel. C'est ce qu'on constate ici.

Ce chapitre visait à présenter les résultats des diverses analyses menées dans cette étude. Dans un premier temps, l'accord interjuges a été revu, de même que les analyses statistiques complémentaires. Ensuite, les résultats au test de perception ont été présentés en détail. Deux groupes d'ANOVAs ont été effectués, de même qu'une matrice de corrélation, afin de mieux comprendre comment les participants ont jugé les neuf séries d'énoncés.

CHAPITRE V

DISCUSSION

Ce chapitre vient apporter une interprétation aux divers résultats présentés dans le chapitre précédent, de façon à les mettre en lien avec les études antérieures. Il vient aussi présenter comment ces résultats permettent d'atteindre les objectifs de recherche que nous nous étions fixés et vérifier les hypothèses posées.

5.1 Premier objectif : Identifier certaines caractéristiques prosodiques principales du français québécois, dans sa perception et sa production

Le premier objectif a été choisi afin de mieux comprendre les différences prosodiques existant entre FF et FQ. Pour atteindre cet objectif, plusieurs démarches ont été entreprises. D'abord, une revue de la littérature a été effectuée (chapitre 2). Ensuite, une série de manipulations ont été faites. Six locuteurs ont été enregistrés, trois français et trois québécois, leurs énoncés ont été délexicalisés par resynthèse puis évalués par cinq juges québécois. Les énoncés les plus prototypiques, à la fois de FF et de FQ, ont été analysés acoustiquement et cette analyse acoustique a fait l'objet de deux analyses statistiques : la régression logistique et l'arbre de classification. Suite à ces démarches, il est possible de vérifier les trois hypothèses liées à ce premier objectif.

5.1.1 Hypothèse 1 : Des différences prosodiques seront observées, à la perception comme à la production entre le FF et le FQ

Cette première hypothèse a été posée à la lumière des diverses études présentées au chapitre 2. En effet, la quasi-totalité des études notait des différences prosodiques entre FF et FQ. Cependant, ces études pouvaient être classées en deux catégories : les études de production et les études perceptives.

Avec cette hypothèse, nous tentions de vérifier les conclusions de ces études. Autrement dit, nous espérions observer des différences significatives autant au niveau de la production qu'au niveau de la perception. C'est effectivement une des conclusions qui peuvent être tirées de cette étude.

En effet, il a été possible de produire des modèles de classification significatifs à la fois sur la base de l'origine réelle des locuteurs (tableau 3-5, figure 4-1) que sur leur origine perçue (tableau 3-6, figure 4-2). Ces résultats indiquent clairement qu'il existe des différences prosodiques, tant au niveau de la perception que de la production, entre FF et FQ.

5.1.2 Hypothèse 2 : Les différences prosodiques observées à la **production** se trouveront principalement au niveau de la **durée**

Cette hypothèse a été posée suite à l'analyse des études antérieures, présentée dans le chapitre 2. En effet, bien que les études soulèvent une grande variété de différences prosodiques entre FF et FQ, les études qui se concentrent sur les différences au niveau de la production tendent à identifier principalement des écarts liés au paramètre acoustique de la durée.

Cette tendance s'est confirmée dans les données que nous avons recueillies. En effet, les modèles de classification effectués pour l'origine réelle des locuteurs ont relevé principalement des variables liées au paramètre acoustique de la durée, tel qu'on peut le voir dans le tableau résumé ci-dessous.

Tableau 5-1 Classification en fonction de l'origine réelle des locuteurs

Méthode statistique	Variables
Régression logistique	Durée V Durée S # ACC / # S Durée Acc/NAcc
Arbre de classification	Durée V

Ces données tendent à démontrer que les différences prosodiques observées à la production se manifestent principalement à travers le paramètre acoustique de la durée.

5.1.3 Hypothèse 3 : Les différences prosodiques observées à la **perception** se trouveront principalement au niveau de la **F0**

Cette hypothèse, quant à elle, a été formulée suite à l'analyse des études centrées sur la perception (chapitre 2). En effet, ces études tendent à relever des paramètres acoustiques différents des études centrées sur la production. Ménard (1998), Bissonnette (2000) et Kaminskaïa (2004), par exemple, isolent des paramètres acoustiques liés à la F0 – notamment le registre. Il s'agirait du paramètre acoustique dont se servent le plus les auditeurs afin d'identifier l'origine géographique de leur interlocuteur sur la base de la prosodie.

Les résultats issus de cette étude viennent confirmer notre hypothèse. En effet, comme on peut le voir dans le tableau ci-dessous, la plupart des variables identifiées dans les tests statistiques de classification sont associées à la F0.

Tableau 5-2 Classification en fonction de l'origine perçue des locuteurs

Méthode statistique	**Variables**
Régression logistique	dureeV, dureeS, TH/TB Registre, F0 moyenne, F0 TH/TB
Arbre de classification	F0 TH/TNM, F0 moy, Duree V, F0 TB/TNM, dureeS TH/TB, #ACC/ N Syl, intensite moy, registre, imoy Acc/NAcc

Cependant, il convient d'apporter ici quelques nuances. D'abord, contrairement à l'étude de Ménard (1998), le registre ne semble pas s'imposer comme étant le paramètre acoustique principal guidant la perception de l'origine géographique.

De plus, les résultats obtenus pour la tâche d'identification de l'origine au test de perception viennent compléter ce portrait. En effet, parmi les variables qui avaient été identifiées lors des tests statistiques, certaines ont influencé plus grandement l'identification de l'origine géographique.

En effet, il est apparu que, contrairement aux études antérieures, la durée des voyelles s'est avérée être un paramètre favorisant une identification FQ. Les voyelles plus longues sont plus identifiées à FQ, alors que les voyelles plus courtes sont plus identifiées à FF.

Le registre, quant à lui, était effectivement un paramètre influençant l'origine géographique, avec un registre moins étendu étant caractéristique de FQ alors qu'un registre plus grand étant caractéristique de FF.

Au niveau de la F0, les différences de F0 entre les TH et les TB ont, elles aussi, influencé l'origine géographique, une différence plus grande étant identifiée plus FF alors qu'une différence plus petite – voire nulle – a été identifiée plus FQ.

Finalement, en ce qui a trait aux résultats pour l'origine au test de perception, il reste à présenter le résultat intriguant pour la série combinant les trois paramètres de F0. En

effet, cette série combinait trois paramètres acoustiques de F0 : la F0 moyenne, le registre et la différence de F0 entre TH et TB. Or, lorsque ces paramètres ont été testés de manière isolée, le seuil acoustique que nous avions attribué à FF avait effectivement été identifié comme tel par les participants. Cependant, lorsque ces trois paramètres sont appliqués sur un même énoncé, inexplicablement, les trois seuils préalablement identifiés comme FF sont alors identifiés comme FQ.

À la lumière de ces résultats, il est clair que les manipulations de F0 viennent influencer les jugements liés à l'origine géographique, comme on peut le voir dans les tableaux 4-2 et 4-9. Néanmoins, la direction de ce jugement, ou les seuils nécessaires pour faire un jugement FF ou FQ restent à être explorés plus longuement.

5.2 Deuxième objectif : Dans quelle mesure la prosodie du FQ affecte-t-elle l'intelligibilité ou la qualité d'une synthèse

Cet objectif a été posé suite à l'analyse des études de Brousseau (1992) et Côté-Giroux et coll. (2011). Ces deux études démontraient que l'utilisation d'une synthèse segmentale FQ augmentait l'intelligibilité des énoncés auprès d'auditeurs québécois. Il nous a donc semblé pertinent de vérifier si les paramètres prosodiques associés à FQ amélioraient eux aussi l'intelligibilité d'une synthèse et si, incidemment, ils pourraient en améliorer la qualité. Au plan de la qualité, l'étude de Côté-Giroux avait relevé que les auditeurs québécois semblaient préférer les voix de synthèse FF, même si les voix FQ étaient plus intelligibles.

Pour atteindre cet objectif, plusieurs choses ont été entreprises. D'abord, il fallait, pour chaque paramètre acoustique relevé précédemment, identifier des seuils pour lesquels les juges avaient considéré les phrases plus FF ou plus FQ. Ensuite, ces divers seuils ont été implémentés en synthèse à l'aide du synthétiseur Mbrola. Deux tâches ont été créées pour le test de perception : une tâche d'intelligibilité – composée d'une transcription d'une phrase sémantiquement non prévisible – et une tâche de

qualité, composée de trois questions portant sur l'appréciation générale, le confort d'écoute et le naturel des énoncés. À partir des données recueillies, il est possible d'aborder l'hypothèse associée à cet objectif.

5.2.1 Hypothèse 4 : Un modèle FF de la prosodie (F0 et durée) sur une production segmentale FQ entraîne une baisse d'intelligibilité et de qualité

Afin de répondre à cet objectif, les résultats apportés par les participants aux énoncés correspondant aux seuils FF seront analysés. Ainsi, les résultats pour les énoncés FF dans chacune des neuf séries seront revus.

Le tableau 5-2 présente les résultats FF pour chacune des quatre questions à l'étude : le MOS, le confort, le naturel et l'intelligibilité. Pour chaque série, deux résultats sont présentés. Le seuil testé représente le seuil qui avait été préalablement choisi comme FF. L'origine perçue quant à elle représente la série d'énoncés qui a été le plus perçue FF par les participants au test de perception (le score d'origine le plus bas). Par exemple, si on retourne voir le tableau 4-4, celui de la cohérence interne de la série 2, le registre, on remarque que les énoncés les plus identifiés FF par les participants étaient ceux correspondant au seuil FQ. C'est donc ce résultat qui est présenté dans la catégorie origine perçue du tableau suivant. Les résultats présentés correspondent à l'écart par rapport à la moyenne pour chaque série d'énoncés. Ainsi, si un résultat est positif, cela signifie que le résultat pour cet ensemble d'énoncés était supérieur à la moyenne de la série.

Tableau 5-3 Résultats pour le seuil FF

		MOS	Confort	Naturel	Intel
Série 1	Seuil testé	+ 0.04	+ 0.06	- 0.52	- 0.24
	Origine perçue	- 0.34	+ 0.02	+ 0.17	- 0.01
Série 2	Seuil testé	+ 0.44	+ 0.23	+ 0.26	- 0.17
	Origine perçue	- 0.71	- 0.55	- 0.54	- 0.49
Série 3	Seuil testé et origine perçue	+ 0.11	- 0.08	- 0.28	+ 0.16
Série 4	Seuil testé et origine perçue	- 0.46	- 0.38	- 0.52	- 0.25
Série 5	Seuil testé et origine perçue	+ 0.01	+ 0.15	- 0.01	- 0.08
Série 6	Seuil testé	- 0. 21	+ 0.24	+ 0.38	- 0.20
	Origine perçue	- 0.19	- 0.09	+ 0.05	- 0.19
Série 7	Seuil testé	+ 0.73	+ 0.38	+ 0.66	+ 0.39
	Origine perçue	- 0.96	- 0.42	- 0.36	- 0.64
Série 8	Seuil testé et origine perçue	- 0.35	- 0.12	+ 0.10	+ 0.13
Série 9	Seuil testé et origine perçue	- 0.20	- 0.25	- 0.06	- 0.39

D'une manière générale, nous considérons que les résultats liés à l'origine perçue sont plus fiables que ceux du seuil testé. En effet, les seuils établis provenant d'un autre ensemble d'énoncés, il est donc probable, tel qu'il a été démontré lors du test de perception, que les participants assignent une origine géographique différente à chaque ensemble d'énoncés.

Dans un premier temps, lorsqu'il est question d'intelligibilité, on remarque que l'ensemble des résultats pour le seuil FF est généralement négatif. Ceci indique que des éléments prosodiques FF apposés sur une composition segmentale FQ entraînent

d'une manière générale une baisse de l'intelligibilité. Seules deux séries dérogent à cette observation : la série 3, correspondant à la durée des syllabes entre les tons hauts et les tons bas, et la série 8, présentant l'ensemble des paramètres prosodiques à l'étude.

Dans un second temps, lorsqu'il est question de qualité, on remarque aussi que la grande majorité des écarts pour l'origine perçue sont négatifs. Il est ainsi possible d'affirmer que des composantes prosodiques FF sur une production segmentale FQ entraînent une baisse de la qualité. Ainsi, il est donc possible d'affirmer que l'hypothèse 4 est confirmée par notre étude.

Ceci vient donc renforcer les résultats de l'étude de Miller, Schlauch et Watson (2010) qui indiquait que les manipulations non naturelles de F0 entraînaient des pertes d'intelligibilité. Nous allons plus loin, puisque dans notre étude des manipulations non naturelles de la durée entraînent elles aussi des pertes d'intelligibilité. De plus, nous avons prouvé que ces manipulations entraînent également des pertes aussi au niveau de la qualité.

5.3 Troisième objectif : Quels paramètres prosodiques ou combinaison de paramètres sont les plus pertinents dans l'amélioration d'une synthèse de la parole en FQ ?

Cet objectif a été formulé afin de potentiellement dégager des aspects prosodiques qui auraient plus d'impact en synthèse, de manière à guider l'implémentation de la prosodie du FQ. En effet, si certains paramètres ont plus d'importance que d'autres, il convient d'y porter une attention particulière dans le développement des systèmes.

5.3.1 Hypothèse 5 : D'une manière générale, c'est la F0 qui aura le plus d'impact, à la fois sur l'intelligibilité et sur la qualité

Cette hypothèse avait été formulée suite aux travaux de Ménard (1998), Bissonnette (2000) et Kaminskaïa (2004). En effet, ces auteures ont toutes traité principalement

de F0. Ménard, notamment, avait affirmé l'importance de la F0, sous la forme du registre, quant à l'attribution de l'origine géographique. Ces données nous avaient amenée à postuler une plus grande importance de la F0 au niveau de la perception. De plus, la durée n'a pas véritablement été abordée dans les travaux en prosodie du FQ depuis 1995, tel que le lecteur peut voir dans les tableaux synthèse du chapitre 2.

Les résultats que nous avons trouvés ne nous permettent pas de confirmer cette hypothèse. En effet, tel que le lecteur peut voir dans le tableau 4-2, contrairement à notre hypothèse, des variables liées à la durée ont eu un impact significatif à la fois sur la qualité et l'intelligibilité de nos énoncés de synthèse. De plus, nous n'observons pas de différence quant aux résultats des séries liées à la F0 ou à la durée.

CONCLUSION

L'objectif premier de ce mémoire était d'offrir une contribution linguistique aux systèmes de synthèse de la parole en français québécois. Au cours de notre recherche sur la synthèse de la parole, l'importance primordiale d'un aspect phonétique, la prosodie, a été démontré. Ce mémoire cherchait donc à vérifier l'impact que pouvait avoir la prise en compte de la prosodie du FQ sur la qualité et l'intelligibilité d'un système de synthèse.

Afin d'atteindre cet objectif principal, une méthodologie en trois temps a été développée. D'abord, six hommes, trois originaires de France, trois du Québec, ont été enregistrés. Ils ont produits une série de courts énoncés en contexte de lecture à voix haute. Ces énoncés présentaient une certaine variété syntaxique : des phrases neutres, des interrogations et des phrases sous accent d'emphase.

Afin de vérifier l'impact précis de la prosodie, ces énoncés ont été délexicalisés. La méthode de la resynthèse a été retenue puisqu'elle permettait de conserver l'aspect prosodique de la durée, éliminé dans la plupart des autres méthodes de délexicalisation. Ces énoncés délexicalisés ont alors été jugés par cinq juges québécois afin de déterminer l'origine géographique des locuteurs qui les ont produits. À partir de ces jugements, les énoncés les plus fortement identifiés comme québécois ou français ont été retenus.

Ces énoncés ont fait l'objet d'une analyse acoustique fine. Un ensemble varié de paramètres acoustiques ont été relevés, à la fois pour la durée et pour la F0, et sur un certain nombre de niveaux d'analyse – de la syllabe au syntagme intonatif. Ces divers paramètres acoustiques ont fait l'objet d'une analyse statistique de catégorisation, la régression logistique, afin d'identifier les éléments qui ont eu le plus d'impact au moment de l'attribution de l'origine géographique.

L'analyse statistique a permis de relever l'importance de cinq paramètres acoustiques : la F0 moyenne, le registre, la différence de F0 entre les tons hauts et les tons bas, la durée des voyelles et la différence entre la durée des syllabes pour les tons hauts et les tons bas. Pour chacun de ces cinq paramètres acoustiques, une échelle comportant trois seuils a été créée, du plus FQ au plus FF. Ces seuils ont ensuite été implémentés en synthèse à l'aide du synthétiseur Mbrola.

Ces énoncés synthétisés ont été soumis à un ensemble de 15 participants, 8 hommes et 7 femmes lors d'un test de perception. Ce test comportait trois tâches principales, une tâche d'identification de l'origine géographique, une tâche d'intelligibilité – sous la forme d'une transcription – et une tâche de qualité, sous la forme de trois questions.

Les résultats, suite à ce test de perception, permettent d'affirmer, conformément aux études antérieures, qu'il existe bel et bien des différences prosodiques entre le FF et le FQ. De plus, ce test permet de conclure que ces différences se trouvent tant à la perception qu'à la production. L'analyse des résultats acoustiques permet aussi d'affirmer qu'il existe une différence entre les paramètres prosodiques les plus pertinents à la production et à la perception. En effet, les paramètres acoustiques liés à la durée étaient les plus proéminents à la production, alors que les paramètres acoustiques liés à la F0 étaient plus importants à la perception.

Une autre grande conclusion à tirer de cette étude est que des informations prosodiques linguistiquement erronées, notamment sous la forme d'une prosodie FF

sur un contenu segmental FQ, entraîne une diminution de l'intelligibilité et une perte de la qualité sur des énoncés générés en synthèse.

Finalement, alors que nous espérions isoler certaines composantes prosodiques ayant plus d'impact que d'autres sur l'intelligibilité ou la qualité des systèmes de synthèse, il semble plutôt que ce soient les erreurs linguistiques – et non un élément prosodique précis – qui aient le plus grand impact sur la qualité ou l'intelligibilité des énoncés.

Cette étude répond à quelques questions, mais en pose surtout plusieurs autres. En effet, plusieurs avenues restent à être explorées afin de mieux comprendre ces résultats.

Dans un premier temps, il pourrait être pertinent de reprendre cette étude en comparant les résultats en fonction du bagage linguistique des participants. En effet, il serait pertinent de comparer les perceptions en fonction du sexe, de l'âge, de l'origine géographique (la ville ou les régions, l'est ou l'ouest du Québec), de l'expérience musicale ou encore du degré de contact avec des variétés européennes de français.

Il pourrait aussi être intéressant de poursuivre l'étude des composantes prosodiques associées au FQ. En effet, il est apparu évident, lors de la revue de la littérature, puis dans les diverses décisions liées à l'attribution de seuils FF ou FQ, que les différences prosodiques entre ces deux variétés sont peu documentées. Une étude visant à explorer ces seuils perceptifs pour divers aspects prosodiques serait extrêmement pertinente.

Ce mémoire est donc un pas de plus vers une meilleure compréhension de la prosodie du FQ. En effet, à travers divers tests perceptifs, plusieurs composantes prosodiques permettant de mettre en lumière les distinctions entre FF et FQ ont été révélées. De plus, leur impact sur l'intelligibilité et la qualité d'un système de synthèse de la parole a été confirmé.

APPENDICES

APPENDICE A
EXEMPLE DE FORMULAIRE DE CONSENTEMENT

Test de perception

Français du Québec et de France

Laboratoire de phonétique

320, rue Sainte-Catherine Est.
Pavillon J-A De Sève
Université du Québec à Montréal

J'accepte de participer à ce test de perception d'une durée d'environ une heure. Ma participation implique que j'écoute près de 100 énoncés sur un ordinateur, à l'aide d'un casque d'écoute. Je devrai répondre à plusieurs questions pour chaque énoncé. Je n'ai pas de problème d'audition, de langage ou de lecture connu. Je sais que je peux me retirer de cette étude en tout temps. Je sais que cette étude ne me cause aucun danger. Le temps que prend cette étude peut être un inconvénient, mais des pauses ont été prévues tout au long de la procédure afin de limiter ces inconvénients. Les résultats de cette étude sont confidentiels et aucune information permettant de vous identifier ne sera conservée.

Nom :

Prénom :

Signature :

Date :

APPENDICE B
STIMULI DU TEST DE PERCEPTION

Série	Transcriptions	Cible
Série F0 moyenne	Des fous redressent le nez.	1A
	[δεφυ{↔δ{Εσλ↔νε]	
	Le nain lessive un lapin.	1B
	[λ↔νε)λΕσιϖ↵)λαπε)]	
	Le bandit renverse une fée.	1C
	[λ↔βα)δ$_{ζ}$ι{α)ϖΕ{σψνφε]	
	C’est son filou qui cale son bouchon, pas le bébé.	1A
	[σεσ□)φιλυκικαλσ□)βυΣ□)πΑλ↔βεβε]	
	Ce sont ses mulets qui débutent la rue, pas ces mains.	1B
	[σ↔σ□)σεμψλΕκιδεβΨτλα{ψπΑσεμε)]	
	Ce sont ces chats qui dérobent le nez, pas ce robot.	1C
	[σ↔σ□)σεΣΑκιδε{□βλ↔νεπΑσ↔{οβο]	
	Que ce bison répare-t-il ?	1A
	[κ↔σ↔βιζ□){επΑ{τΙλ]	
	Qui le loup rejette-t-il ?	1B
	[κιλ↔λυρ↔ΖΕττΙλ]	
	Que ses bébés cassent-ils ?	1C
	[κ↔σεβεβεκαστΙλ]	
Série registre	Le pou mélange des gants.	2A
	[λ↔πυμελα)Ζδεγα)]	
	La fée recouche ces sapins.	2B
	[λαφε{↔κΥΣσεσαπε)]	
	Ce chou combat des râteaux.	2C
	[σ↔Συκ□)βΑδε{Ατο]	
	C’est son bureau qui recadre les coups, pas son palais.	2A
	[σεσ□)βψ{οκι{↔καδ{λεκυπΑσ□)παλΕ]	
	C’est un mouton qui raconte ce gain, pas le boucher.	2B
	[σετ↵)μυτ□)κι{ακ□)σ↔γε)πΑλ↔βυΣε]	
	C’est le bison qui rogne les bancs, pas les buts.	2C
	[σελ↔βιζ□)κιρ□Νλεβα)πΑλεβψ]	
	Que ces gars retirent-ils ?	2A
	[κ↔σεγΑ{↔τ$_{σ}$ι{τ$_{σ}$ιλ]	
	Où ces bœufs campent-ils ?	2B
	[υσεβΟκα)πτ$_{σ}$Ιλ]	

	Quand son gars bétonne-t-il ?	2C
	[κα)σ□)γAβετ□ντσIλ]	
Série durée des syllabes Th/TB	Un matou menace le tango.	3A
	[↵)ματυμ↔νασλ↔τα)γο]	
	Cette fée manque la fin.	3B
	[σEτφεμα)κλαφε)]	
	Un rat finance le thé.	3C
	[↵){Aφινα)σλ↔τε]	
	C'est le thon qui recrute le banc, pas les marais.	3A
	[σελ↔τ□)κι{↔κ{Ψτλ↔βα)πAλεμα{E]	
	Ce sont ces bébés qui mènent les jeux, pas un loup.	3B
	[σ↔σ□)σεβεβεκιμEνλεZOπA↵)λυ]	
	C'est son minet qui remonte le poulain, pas ses noms.	3C
	[σεσ□)μινEκι{↔μ□)τλ↔πυλε)πAσεν□)]	
	Qui les bouffons congèlent-ils ?	3A
	[κιλεβυφ□)κ□)ZEλτσIλ]	
	Où ses rats virent-ils ?	3B
	[υσε{Aϖι{τσIλ]	
	Quand le rat gigote-t-il ?	3C
	[κα)λ↔{AZιγ□ττσIλ]	
Série F0 moyenne Th/TB	Un sot compte ses bouleaux.	4A
	[↵)σοκ□)τσεβυλο]	
	Le mouton rate ces pains.	4B
	[λ↔μυτ□){ατσεπε)]	
	Son bison mérite des bonbons.	4C
	[σ□)βιζ□)με{Iτδεβ□)β□)]	
	Ce sont des rats qui consignent des ponts, pas des gars.	4A
	[σ↔σ□)δε{Aκικ□)σINδεπ□)πAδεγA]	
	C'est un vœu qui récite le mot, pas ce mouton.	4B
	[σετ↵)ϖOκι{εσIτλ↔μοπAσ↔μυτ□)]	
	C'est ce daim qui bétonne le tas, pas ces nœuds.	4C
	[σεσ↔δε)κιβετ□νλ↔τAπAσενO]	
	Que ces pies tapent-elles ?	4A
	[κ↔σεπιταπτEλ]	
	Que son nain rallume-t-il ?	4B
	[κ↔σ□)νε)ραλΨμτσIλ]	
	Que sa main démêle-t-elle ?	4C
	[κ↔σαμε)δεμEλτEλ]	

Série Durée Voyelle	Des gueux méritent un boni.	5A
	[δεγOμε{Iτ↵)β□νι]	
	Le fou lave son tipi.	5B
	[λ↔φυλαϖσ□)τσιπι]	
	Un matou rase le donjon.	5C
	[↵)ματυ{Aζλ↔δ□)Z□)]	
	Ce sont des veaux qui redressent ce nid, pas un bébé.	5A
	[σ↔σ□)δεϖοκι{↔δ{Eσσ↔νιπA↵)βεβε]	
	C'est le bandit qui raisonne ce machin, pas le volant.	5B
	[σελ↔βα)δζικι{Eζ□νσ↔μαΣε)πAλ↔ϖ□λα)]	
	C'est un bébé qui picosse ce nez, pas le nid.	5C
	[σετ↵)βεβεκιπικ□σσ↔νεπAλ↔νι]	
	Qui ce matou laisse-t-il ?	5A
	[κισ↔ματυλEστσIλ]	
	Que ce mouton concasse-t-il ?	5B
	[κ↔σ↔μυτ□)κκAστσIλ]	
	Que le rat signe-t-il ?	5C
	[κ↔λ↔{AσINτσIλ]	
Série Combo durée	Les nains volent le chaton	6A
	[λενε)ϖ□λλ↔Σατ□)]	
	La pie gèle des loups.	6B
	[λαπιZEλδελυ]	
	Les mulets découpent le thym.	6C
	[λεμψλEδεκYπλ↔τε)]	
	Ce sont les badauds qui retapent ces pins, pas un bonnet.	6A
	[σ↔σ□)λεβαδοκι{↔ταπσεπε)πA↵)β□νE]	
	C'est un bébé qui venge ce jambon, pas ces forêts.	6B
	[σετ↵βεβεκιϖα)Zσ↔Zα)β□)πAσεφ□{E]	
	C'est un cou qui commente ses reins, pas un balai.	6C
	[σετ↵κυκικ□μα)τσερε)πA↵)βαλE]	
	Qui le poulain tamponne-t-il ?	6A
	[κιλ↔πυλε)τα)π□)ντσIλ]	
	Où son boucher dévale-t-il ?	6B
	[υσ□)βυΣεδεϖαλτσIλ]	
	Qui le veau bouge-t-il ?	6C
	[κιλ↔ϖοβYZτσIλ]	
Sér ie	Ce chat donne un rat.	7A
	[σ↔ΣAδ□ν↵){A]	

	Un pou chicane ces pies.	7B
	[↵)πυΣικανσεπι]	
	Un mouton reboise des bâtons.	7C
	[↵)μυτ□){↔βωΑζδεβΑτ□)]	
	C’est un matou qui recourbe la queue, pas les mets.	7A
	[σετ↵)ματυκι{↔κυ{βλακΟπΑλεμΕ]	
	C’est son rat qui cogne sa bougie, pas sa baie.	7B
	[σεσ□){Ακικ□Νσαβυ ZιπΑσαβΕ]	
	C’est la fée qui rajoute un pot, pas des loups.	7C
	[σελαφεκι{αZYτ↵)ποπΑδελυ]	
	Que ses mains colorent-elles ?	7A
	[κ↔σεμε)κ□λ□{τΕλ]	
	Que ces badauds remisent-ils ?	7B
	[κ↔σεβαδο{↔μιζτσΙλ]	
	Qui ces fous visitent-ils ?	7C
	[κισεφυϖιζΙττσΙλ]	
Combo F0 et durée	Un bison donne un cadeau.	8A
	[↵)βιζ□)δ□ν↵)καδο]	
	Son veau bouffe le gâteau.	8B
	[σ□ϖοβΥϕλ↔γΑτο]	
	Le rat ménage le pain.	8C
	[λ↔{ΑμεναΖλ↔πε)]	
	C’est le boucher qui tisse ce mont, pas son minet.	8A
	[σελ↔βυΣ3εκιτΙσσ↔μ□)πΑσ□)μινΕ]	
	C’est ce bureau qui possède le balai, pas le boulot.	8B
	[σεσ↔βψ{οκιπ□σΕδλ↔βαλΕπΑλ↔βυλο]	
	C’est ce mouton qui rabaisse le goût, pas le donjon.	8C
	[σεσ↔μυτ□)κι{αβΕσλ↔γυπΑλ↔δ□)Ζ□)]	
	Que ce matou combine-t-il ?	8A
	[κ↔σ↔ματυκ□)βΙντσΙλ]	
	Que ce forain couve-t-il ?	8B
	[κ↔σ↔ϕ□{ε)κυϖτσΙλ]	
	Qui le sceau cimente-t-il ?	8C
	[κιλ↔σοσιμα)ττσΙλ]	
Combo F0 et	Les veaux condamnent un champ.	9A
	[λεϖοκ□)δΑν↵)Σα)]	
	Le poulain lève cette roue.	9B
	[λ↔πυλε)λΕϖσΕτ{υ]	

	Les gars cessent un jeu.	9C
	[λεγAσEσ↵)ZO]	
	C’est une main qui recompte ses bidons, pas ces boulots.	9A
	[σεψνμε)κι{↔κ□)τσεβιδ□)πAσεβυλο]	
	C’est un gueux qui ravage ces marais, pas ce loup.	9B
	[σε↵)γOκι{αϖαZσεμα{EπAσ↔λυ]	
	C’est son baudet qui rejette sa laitue, pas les dents.	9C
	[σεσ□)βοδEκι{↔ZEτσαλEτσψπAλεδα)]	
	Qui le mouton douche-t-il ?	9A
	[κιλ↔μυτ□)δYΣτσIλ]	
	Que ce minet gère-t-il ?	9B
	[κ↔σ↔μινEZE{τσIλ]	
	Que cette main diffuse-t-elle ?	
	[κ↔σEτμε)δζιϕΨζτEλ]	

Série		Cible
Série F0 moy	1A	150 Hz
	1B	135 Hz
	1C	120 Hz
Série registre	2A	120 Hz
	2B	60 Hz
	2C	10 Hz
Série SylDur TH/TB	3A	1.5
	3B	1.2
	3C	1
Série F0 moy TH/TB	4A	2
	4B	1.5
	4C	1
Série durée V	5A	75 ms
	5B	85 ms
	5C	110 ms
Série durée	6A	3A, 5A
	6B	3B, 5B
	6C	3C, 5C
Série F0	7A	1A, 2A, 4A
	7B	1B, 2B, 4B
	7C	1C, 2C, 4C
Série F0-durée	8A	1A, 2A, 3A, 4A, 5A
	8B	1B, 2B, 3B, 4B, 5B
	8C	1C, 2C, 3C, 4C, 5C
Série F0-durée inversés	9A	1A, 2A, 4A, 3C, 5C
	9B	1B, 2B, 3B, 4B, 5B
	9C	1C, 2C, 4C, 3A, 5A

APPENDICE C
FICHIER .MLC TIRÉ DE EULER

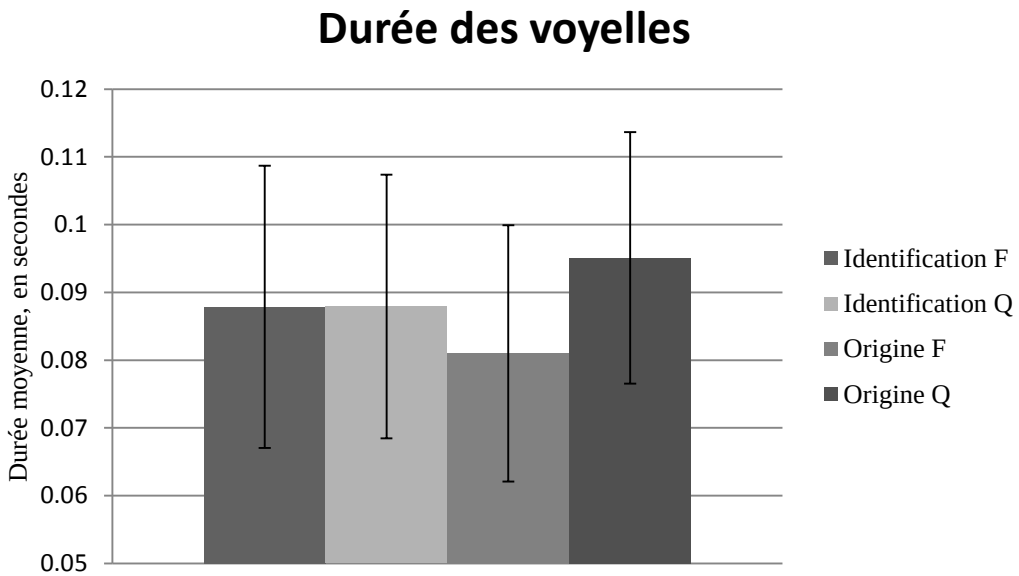
```
Mlc1
'Word'
        '.','ENDPUNCT','UNSTRESSED'; 0; 'GrammarUnit' 0 ; 'Syllable' 0 ;
'Phoneme' 0
        'des','DET','UNSTRESSED'; 1; 'GrammarUnit' 1 ; 'Syllable' 1 ; 'Phoneme' 1
        'fous','NOUN','PRIMARYSTRESS'; 2; 'GrammarUnit' 2 ; 'Syllable' 2 ;
'Phoneme' 3
        'redressent','VERB','PRIMARYSTRESS'; 3; 'GrammarUnit' 3 ; 'Syllable' 3 4 ;
'Phoneme' 5
        'le','DET','UNSTRESSED'; 4; 'GrammarUnit' 4 ; 'Syllable' 5 ; 'Phoneme' 11
        'nez','NOUN','PRIMARYSTRESS'; 5; 'GrammarUnit' 5 ; 'Syllable' 6 ;
'Phoneme' 13
        '.','ENDPUNCT',''; 6; 'GrammarUnit' 6 ; 'Syllable' 7 ; 'Phoneme' 15

'Token'
        '.','PUNCTUATION'; 0; 'GrammarUnit' 0
        'des','WORD'; 1; 'GrammarUnit' 1
        ' ','WHITE'; 2
        'fous','WORD'; 3; 'GrammarUnit' 2
        ' ','WHITE'; 4
        'redressent','WORD'; 5; 'GrammarUnit' 3
        ' ','WHITE'; 6
        'le','WORD'; 7; 'GrammarUnit' 4
        ' ','WHITE'; 8
        'nez','WORD'; 9; 'GrammarUnit' 5
        '.','PUNCTUATION'; 10; 'GrammarUnit' 6

'GrammarUnit'
        'ENDPUNCT','ENDPUNCT'; 0; 'Token' 0 ; 'Word' 0
        'DET','DET DETPREP'; 1; 'Token' 1 ; 'Word' 1
        'NOUN','ADJ NOUN'; 2; 'Token' 3 ; 'Word' 2
        'VERB','VERB'; 3; 'Token' 5 ; 'Word' 3
        'DET','DET PRONCL'; 4; 'Token' 7 ; 'Word' 4
        'NOUN','NOUN'; 5; 'Token' 9 ; 'Word' 5
        'ENDPUNCT','ENDPUNCT'; 6; 'Token' 10 ; 'Word' 6

'Phoneme'
```

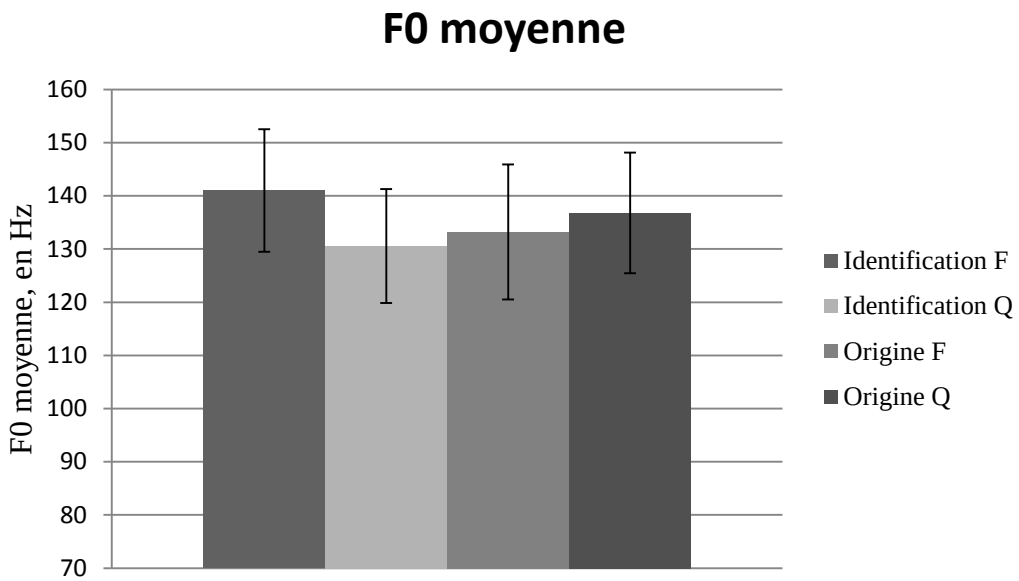
```
        '_',464,'SILENCE','UNVOICED','','NUCLEUS',''; 0; 'Word' 0 ; 'Syllable' 0
        'd',55,'VOICEDPLOSIVE','UNVOICED','','ONSET',''; 1; 'Word' 1 ; 'Syllable'
1
        'e',58,'ORALVOWEL','VOICED','','NUCLEUS',''; 2; 'Syllable' 1
        'f',99,'FRICATIVE','UNVOICED','','ONSET',''; 3; 'Word' 2 ; 'Syllable' 2
        'u',71,'ORALVOWEL','VOICED','','NUCLEUS',''; 4; 'Syllable' 2
        'R',69,'LIQUID','UNVOICED','','ONSET',''; 5; 'Word' 3 ; 'Syllable' 3
        '@',45,'ORALVOWEL','VOICED','','NUCLEUS',''; 6; 'Syllable' 3
        'd',54,'VOICEDPLOSIVE','UNVOICED','','ONSET',''; 7; 'Syllable' 4
        'R',61,'LIQUID','UNVOICED','','ONSET',''; 8; 'Syllable' 4
        'E',66,'ORALVOWEL','VOICED','','NUCLEUS',''; 9; 'Syllable' 4
        's',84,'FRICATIVE','UNVOICED','','ONSET',''; 10; 'Syllable' 5
        'l',42,'LIQUID','UNVOICED','','ONSET',''; 11; 'Word' 4 ; 'Syllable' 5
        '@',48,'ORALVOWEL','VOICED','','NUCLEUS',''; 12; 'Syllable' 5
        'n',62,'NASAL','VOICED','','ONSET',''; 13; 'Word' 5 ; 'Syllable' 6
        'e',83,'ORALVOWEL','VOICED','','NUCLEUS',''; 14; 'Syllable' 6
        '_',416,'SILENCE','UNVOICED','','NUCLEUS',''; 15; 'Word' 6 ; 'Syllable' 7

'Syllable'
        '_','SIL','UNDEFINED','P1',464,1; 0; 'Word' 0 ; 'Phoneme' 0
        'de','CV','NA','ll',113,2; 1; 'F0Target' 0 ; 'Word' 1 ; 'Phoneme' 1 2
        'fu','CV','AF','HH',170,2; 2; 'F0Target' 5 10 ; 'Word' 2 ; 'Phoneme' 3 4
        'R@','CV','NA','ll',114,2; 3; 'F0Target' 15 ; 'Word' 3 ; 'Phoneme' 5 6
        'dRE','CV','AF','HH',181,3; 4; 'F0Target' 20 25 ; 'Word' 3 ; 'Phoneme' 7 8 9
        'sl@','CV','NA','ll',174,3; 5; 'F0Target' 30 ; 'Word' 4 ; 'Phoneme' 10 11 12
        'ne','CV','AF','L-L-',145,2; 6; 'F0Target' 35 40 ; 'Word' 5 ; 'Phoneme' 13 14
        '_','SIL','UNDEFINED','P1',416,1; 7; 'Word' 6 ; 'Phoneme' 15

'F0Target'
        519,107; 0; 'Syllable' 1
        519,107; 1
        519,107; 2
        519,104; 3
        519,103; 4
        676,101; 5; 'Syllable' 2
        676,101; 6
        676,101; 7
        676,100; 8
        676,100; 9
        676,100; 10; 'Syllable' 2
        676,100; 11
```

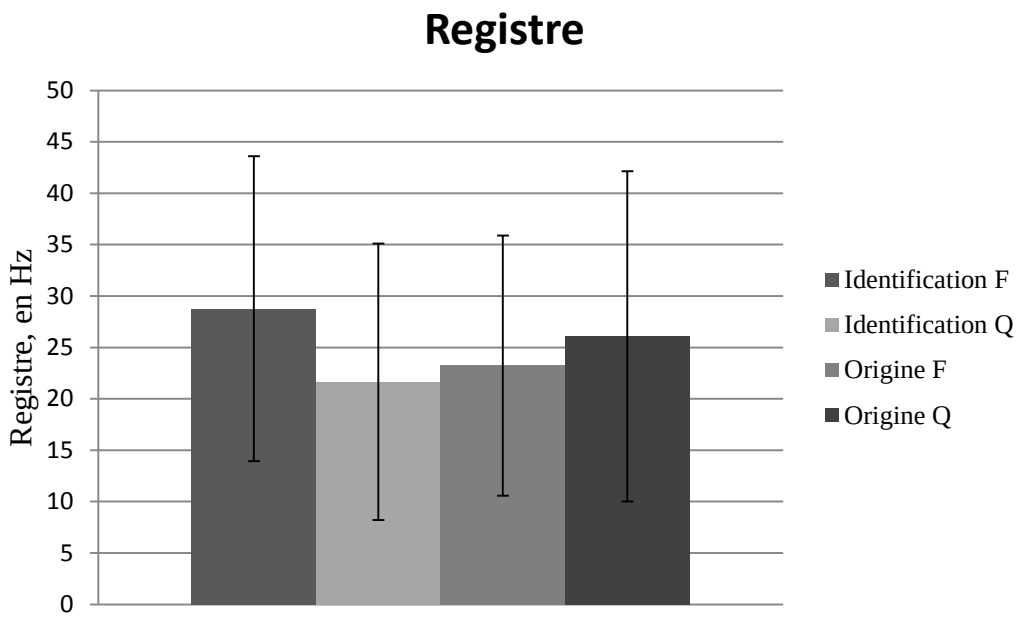
676,100; 12
676,100; 13
676,100; 14
816,115; 15; 'Syllable' 3
816,120; 16
816,125; 17
816,130; 18
816,135; 19
976,150; 20; 'Syllable' 4
976,155; 21
976,160; 22
976,165; 23
976,170; 24
976,176; 25; 'Syllable' 4
976,176; 26
976,176; 27
976,176; 28
976,171; 29
1168,156; 30; 'Syllable' 5
1168,152; 31
1168,152; 32
1168,152; 33
1168,152; 34
1216,154; 35; 'Syllable' 6
1228,133; 36
1240,112; 37
1253,91; 38
1265,70; 39
1216,139; 40; 'Syllable' 6
1228,134; 41
1240,129; 42
1253,124; 43
1265,119; 44

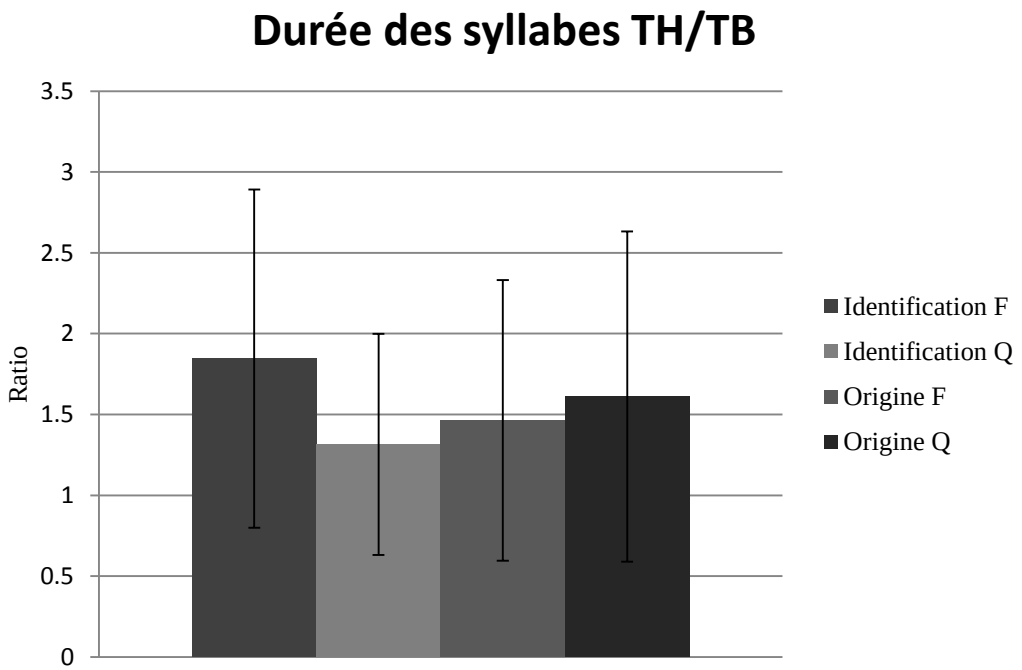
Mlc2

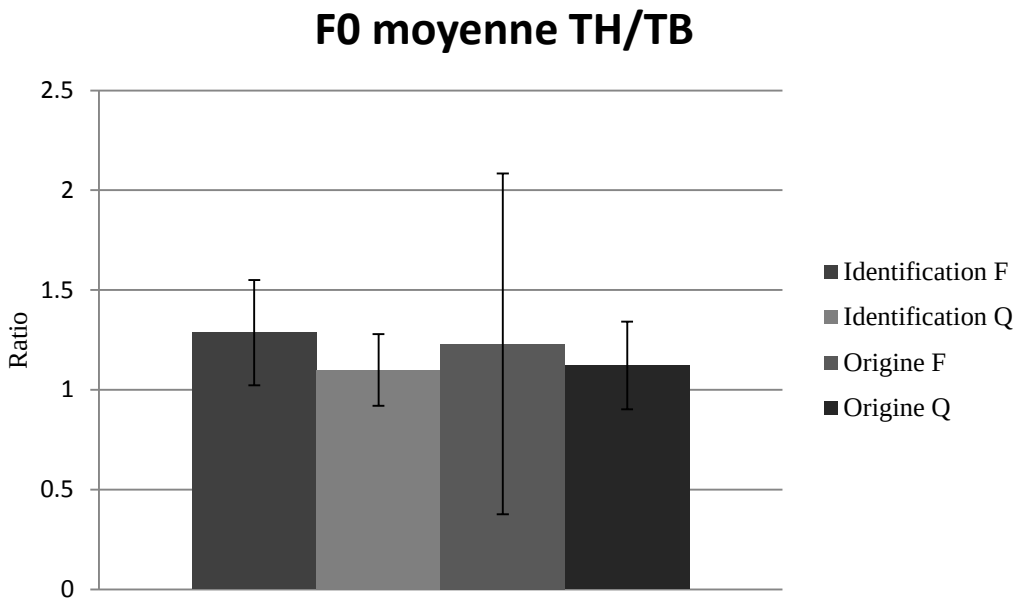
APPENDICE D
GRAPHIQUES






Registre
Registre, en Hz
50
45
40
35
30
25
20
15
10
5
0
Identification F
Identification Q
Origine F
Origine Q


Durée des syllabes TH/TB
Ratio
3.5
3
2.5
2
1.5
1
0.5
0
Identification F
Identification Q
Origine F
Origine Q


F0 moyenne TH/TB
2.5
2
1.5
1
0.5
0
Ratio
Identification F
Identification Q
Origine F
Origine Q

APPENDICE E
MODIFICATION DES PARAMÈTRES PROSODIQUES

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
import sys
import os.path

voyelles = ["i", "e", "E", "a", "O", "o", "u", "y", "2", "9", "@", "e~", "a~", "o~", "9~", "Y", "I", "3", "U", "J", "A"]

f0_moyen_list = [1,3,5,9,11,13,20,22,24]
registre_list = [25,27,29,33,35,37,44,46,48]
syl_dur_list = [49,51,53,59,61,63,66,68,70]
syl_f0_list = [73,75,79,83,85,87,92,94,96]
duree_v_list = [97,99,103,105,107,109,116,118,120]
duree_comb_list = [121,123,127,131,133,135,138,140,144]
f0_comb_list = [147,149,151,155,157,159,162,164,166]
comb_list = [169,171,173,177,181,183,186,188,192]
comb_inv_list = [193,195,199,203,205,207,210,214,216]

list_total = f0_moyen_list + registre_list + syl_dur_list + syl_f0_list + duree_v_list + duree_comb_list + f0_comb_list +
comb_list + comb_inv_list

def registre(f0, cible_id = None):
  cible = {1:120, 2:60, 3:10}    #{1:45, 2:30, 3:20, 4:10}
  f0_max = 0
  f0_min = 90000 #vraiment haute valeure
  f0_med = 0
  for k in f0:
    for f_zero in f0[k]:
      if float(list(f_zero.values())[0]) > f0_max:
        f0_max = float(list(f_zero.values())[0])
      elif float(list(f_zero.values())[0]) < f0_min:
        f0_min = float(list(f_zero.values())[0])
  f0_med = (f0_max - f0_min) / 2 + f0_min
  if cible_id is None:
    offset = (cible[registre_list.index(i)%3 +1] - (f0_max - f0_min)) / 2
  else:
    offset = (cible[cible_id] - (f0_max - f0_min)) / 2
  '''
  if i%8:
    offset = (cible[round(i%8/2+0.1)] - (f0_max - f0_min)) / 2
  else:
    offset = (cible[4] - (f0_max - f0_min)) / 2
  '''
  distance_max = f0_med - f0_min

  for k in f0:
    for f_zero in f0[k]:
      for a in f_zero:
        #print(f_zero[a] + " ", end=" ")
        f_zero[a] = float(f_zero[a]) + (((float(f_zero[a]) - f0_med) / distance_max) * offset)
        #print(f_zero[a])

def moyen(f0, cible_id = None):
  cible = {1:150, 2:135, 3:120}   #{1:150, 2:140, 3:130, 4:120}
  f0_moy = 0
```

```
  total = 0
  for k in f0:
    for f_zero in f0[k]:
      f0_moy += float(list(f_zero.values())[0])
    total += len(f0[k])
  f0_moy /= total
  if cible_id is None:
    ratio = cible[f0_moyen_list.index(i)%3 +1] / f0_moy
  else:
    ratio = cible[cible_id] / f0_moy
  '''
  if i%8:
    ratio = cible[round(i%8/2+0.1)] / f0_moy
  else:
    ratio = cible[4] / f0_moy
  '''
  for k in f0:
    for f_zero in f0[k]:
      for a in f_zero:
        #print(str(f_zero[a]) + " ", end=" ")
        f_zero[a] = float(f_zero[a]) * ratio
        #print(round(f_zero[a]))

def syl_dur(duree, syllable, cible_id = None):
  cible = {1:1.5, 2:1.2, 3:1}   #{1:1.5, 2:1.3, 3:1, 4:0.7}
  H_moy = 0
  tot_H = 0
  L_moy = 0
  tot_L = 0
  ratio = 0
  for k in syllable:
    if k.ton == "H":
      H_moy += int(k.duree[0])
      tot_H += 1
    else:
      L_moy += int(k.duree[0])
      tot_L += 1
  ratio = (H_moy / tot_H) / (L_moy / tot_L)
  if i==70:
    print("ratio = " + str(ratio))
  #print("total TH : " + str(H_moy) + " moy_H : " +  str(H_moy / tot_H))
  #print("total TL : " + str(L_moy) + " moy_L : " +  str(L_moy / tot_L))
  #print("ratio : " + str((H_moy / tot_H) / (L_moy / tot_L)))
  if cible_id is None:
    ratio_offset = cible[syl_dur_list.index(i)%3 +1] / ratio
  else:
    ratio_offset = cible[cible_id] / ratio
  if i == 70:
    print("ratio_offset = "+str(ratio_offset))
  '''
  if i%8:
    ratio_offset = (cible[round(i%8/2+0.1)] / ratio)
  else:
    ratio_offset = (cible[4] / ratio)
  '''

  for k in duree:
    for dur in duree[k]:
      #print(dur + " " + str(duree[k][dur]) + " ", end=" ")
      not_in_s = True
      for s in syllable:
        if duree[k] in s.duree:
```

```
        not_in_s = False
        if s.ton == "H":
          if i == 70:
            print(str(dur) + " " + str(duree[k][dur]) + " ", end=" ")
          duree[k][dur] = float(duree[k][dur]) * ratio_offset
          if i==70:
            print(duree[k][dur])
        else:
          if i == 70:
            print(str(dur) + " " + str(duree[k][dur]) + " ", end=" ")
          duree[k][dur] = float(duree[k][dur])
          if i == 70:
            print(duree[k][dur])
    if not_in_s:
      pass
      #print()

 H_moy = 0
 tot_H = 0
 L_moy = 0
 tot_L = 0
 ratio = 0
 H = ""
 prev_s = ""
 for k in duree:
   for dur in duree[k]:
     for s in syllable:
       if duree[k] in s.duree:
         if s.ton == "H":
           H_moy += duree[k][dur]
           H = True
         else:
           L_moy += duree[k][dur]
           H = False
         if prev_s != s:
           if H:
             tot_H += 1
           else:
             tot_L += 1
         prev_s = s
 #print("total TH : " + str(H_moy) + " moy_H : " +  str(H_moy / tot_H))
 #print("total TL : " + str(L_moy) + " moy_L : " +  str(L_moy / tot_L))
 #print("ratio : " + str((H_moy / tot_H) / (L_moy / tot_L)))

def syl_f0(f0, duree, syllable, cible_id = None):
  cible = {1:2, 2:1.5, 3:1}   #{1:2, 2:1.5, 3:1, 4:0.5}
  H_moy = 0
  tot_H = 0
  L_moy = 0
  tot_L = 0
  ratio = 0
  for k in syllable:
    if k.ton == "H":
      for f in k.f0:
        H_moy += float(list(f.values())[0])
        tot_H += 1
    else:
      for f in k.f0:
        L_moy += float(list(f.values())[0])
        tot_L += 1
  ratio = (H_moy / tot_H) / (L_moy / tot_L)
```

```
#print("total TH : " + str(H_moy) + " moy_H : " +  str(H_moy / tot_H))
#print("total TL : " + str(L_moy) + " moy_L : " +  str(L_moy / tot_L))
#print("ratio : " + str((H_moy / tot_H) / (L_moy / tot_L)))
if cible_id is None:
  ratio_offset = cible[syl_f0_list.index(i)%3 +1] / ratio
else:
  ratio_offset = cible[cible_id] / ratio
'''
if i%8:
  ratio_offset = (cible[round(i%8/2+0.1)] / ratio)
else:
  ratio_offset = (cible[4] / ratio)
'''
#print("ratio_offset : " + str(ratio_offset))

for k in duree:
  for dur in duree[k]:
    #print(dur + " " + duree[k][dur] + " ")#, end=" ")
    not_in_s = True
    for s in syllable:
      if duree[k] in s.duree:
        not_in_s = False
        if s.ton == "H":
          if k in f0:
            for f in f0[k]:
              #print(list(f.keys())[0], end=" ")
              #print(list(f.values())[0], end=" ")
              f[list(f.keys())[0]] = float(f[list(f.keys())[0]]) * ratio_offset
              #print(list(f.values())[0], end=" ")
            #print()
        else:
          if k in f0:
            for f in f0[k]:
              pass
              #print(list(f.keys())[0], end=" ")
              #print(list(f.values())[0], end=" ")
            #print()
    if not_in_s:
      pass
      #print()

H_moy = 0
tot_H = 0
L_moy = 0
tot_L = 0
ratio = 0
for k in syllable:
  if k.ton == "H":
    for f in k.f0:
      H_moy += float(list(f.values())[0])
      tot_H += 1
  else:
    for f in k.f0:
      L_moy += float(list(f.values())[0])
      tot_L += 1
ratio = (H_moy / tot_H) / (L_moy / tot_L)

#print("total TH : " + str(H_moy) + " moy_H : " +  str(H_moy / tot_H))
#print("total TL : " + str(L_moy) + " moy_L : " +  str(L_moy / tot_L))
#print("ratio : " + str((H_moy / tot_H) / (L_moy / tot_L)))
```

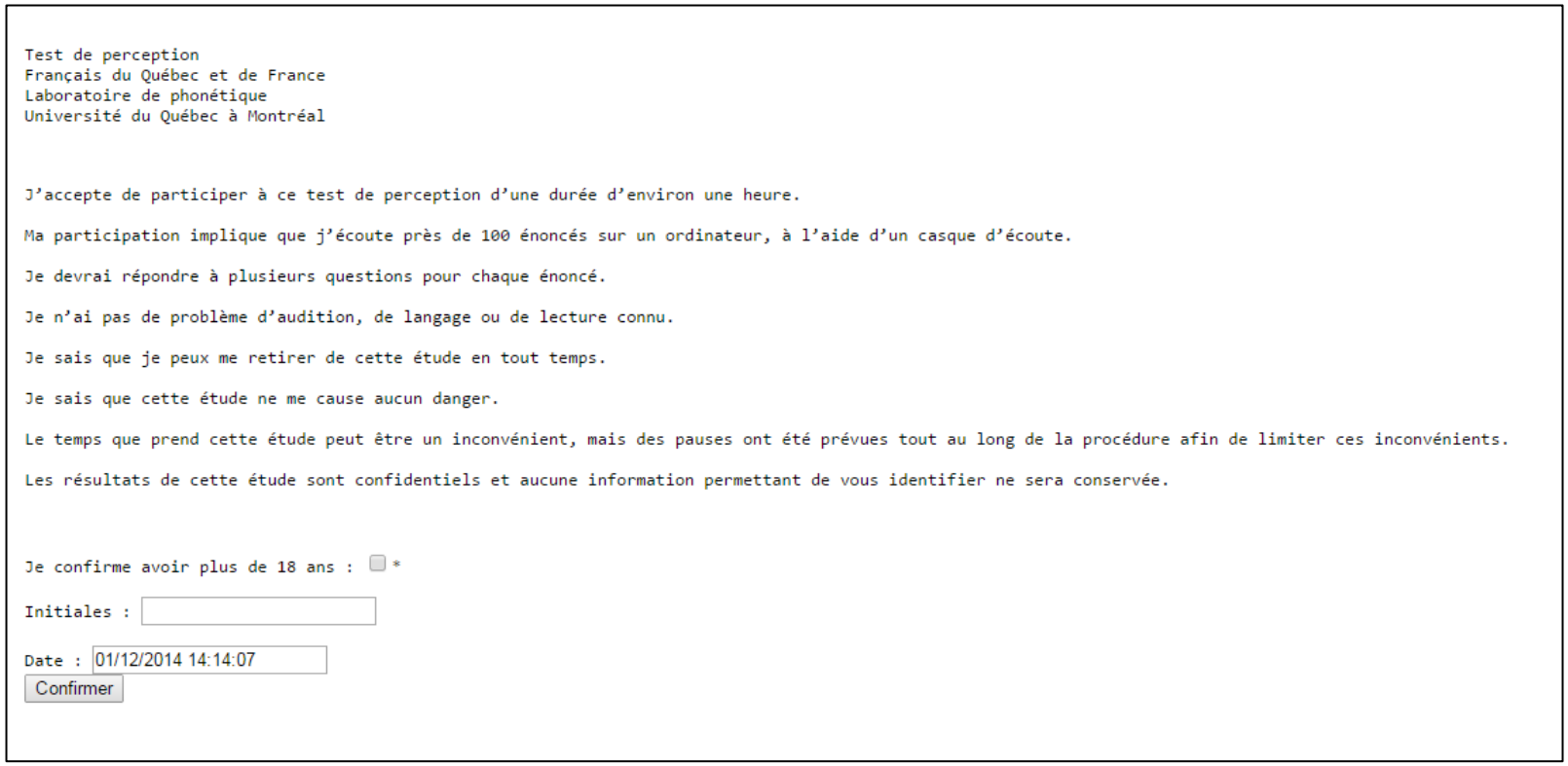
```
def duree_v(duree, syllable, cible_id = None):
  cible = {1:75, 2:85, 3:110}   #{1:60, 2:80, 3:95, 4:110}
  v_moy = 0
  total = 0
  for s in syllable:
    if s.index_voy in duree:
      v_moy += float(list(duree[s.index_voy].values())[0])
    total += 1
  v_moy /= total
  if cible_id is None:
    ratio = cible[duree_v_list.index(i)%3 +1] / v_moy
  else:
    ratio = cible[cible_id] / v_moy
  '''
  if i%8:
    ratio = cible[round(i%8/2+0.1)] / v_moy
  else:
    ratio = cible[4] / v_moy
  '''

  #print("v_moy : " + str(v_moy))
  #print("ratio : " + str(ratio))

  for k in duree:
    for dur in duree[k]:

      for s in syllable:
        if k == s.index_voy:
          if i == 70:
            print(str(dur) + " " + str(duree[k][dur]) + " ", end=" ")
          duree[k][dur] = float(duree[k][dur]) * ratio
          if i == 70:
            print(duree[k][dur], end=" ")
      if i == 70:
        print()

class syl:
  def __init__(self, ton):
    self.ton = ton
    self.duree = []
    self.f0 = ""
    self.index_voy = ""

  def add_duree(self, dur):
    self.duree.append(dur)
  def add_f0(self, fzero):
    self.f0 = fzero
  def set_voyelle(self, voyelle):
    self.index_voy = voyelle

  def __str__(self):
    str_duree = ""
    for d in range(len(self.duree)):
      str_duree += str(self.duree[d]) + " "

    return "ton : " + self.ton + " duree : " + str_duree + " f0 : " + str(self.f0) + " index voyelle : " + str(self.index_voy)

for i in list_total:

  duree = {}
  f0 = {}
  if os.path.exists(r""+str(i)+"_debit.pho"):
```

```
  lines = tuple(open(r""+str(i)+"_debit.pho", "r"))
else:
  print("pho file read error\n")
  continue

fout = open(r""+str(i)+"_debit_mod.pho", 'w')
j = 0
for line in lines:
  if ";" in line:
    fout.write(line)
    continue
  if line.strip() != "" and line.strip() != '\x00':
    val = line.strip().split(" ")
    duree[j] = {val[0] : val[1].strip()}
    if len(val) > 2:
      f0[j] = []
      for k in range(2, len(val), 2):
        f0[j].append({val[k] : val[k+1].strip()})
    j += 1

print(i)
#print(duree)
if os.path.exists(r""+str(i)+".txt.mlc"):
  mlc_lines = tuple(open(r""+str(i)+".txt.mlc", "r"))
else:
  print("mlc file read error\n")
  continue

good = False
syllable = []
phoneme_num = 0
phoneme_actuel = 0
for line in mlc_lines:
  if line.find("Syllable") == 1:
    good = True
    continue
  if good:
    if line.strip() == "":
      continue
    if line.find("'") == 0:
      break
    a = line.split(",")
    for j in range(len(a)):
      a[j] = a[j].lstrip().strip()
    #print(a)
    phoneme_num += int(a[5][:1])
    #print(phoneme_actuel)
    #print(phoneme_num)
    if "L" in a[3] or "l" in a[3]:
      syllable.append(syl("L"))
      syllable[-1].add_duree(a[4])
      while phoneme_actuel < phoneme_num:
        if list((duree[phoneme_actuel].keys()))[0] in voyelles:
          syllable[-1].set_voyelle(phoneme_actuel)
        if phoneme_actuel in f0:
          syllable[-1].add_f0(f0[phoneme_actuel])
        syllable[-1].add_duree(duree[phoneme_actuel])
        phoneme_actuel += 1
    elif "H" in a[3] or "h" in a[3]:
      syllable.append(syl("H"))
      syllable[-1].add_duree(a[4])
      while phoneme_actuel < phoneme_num:
```

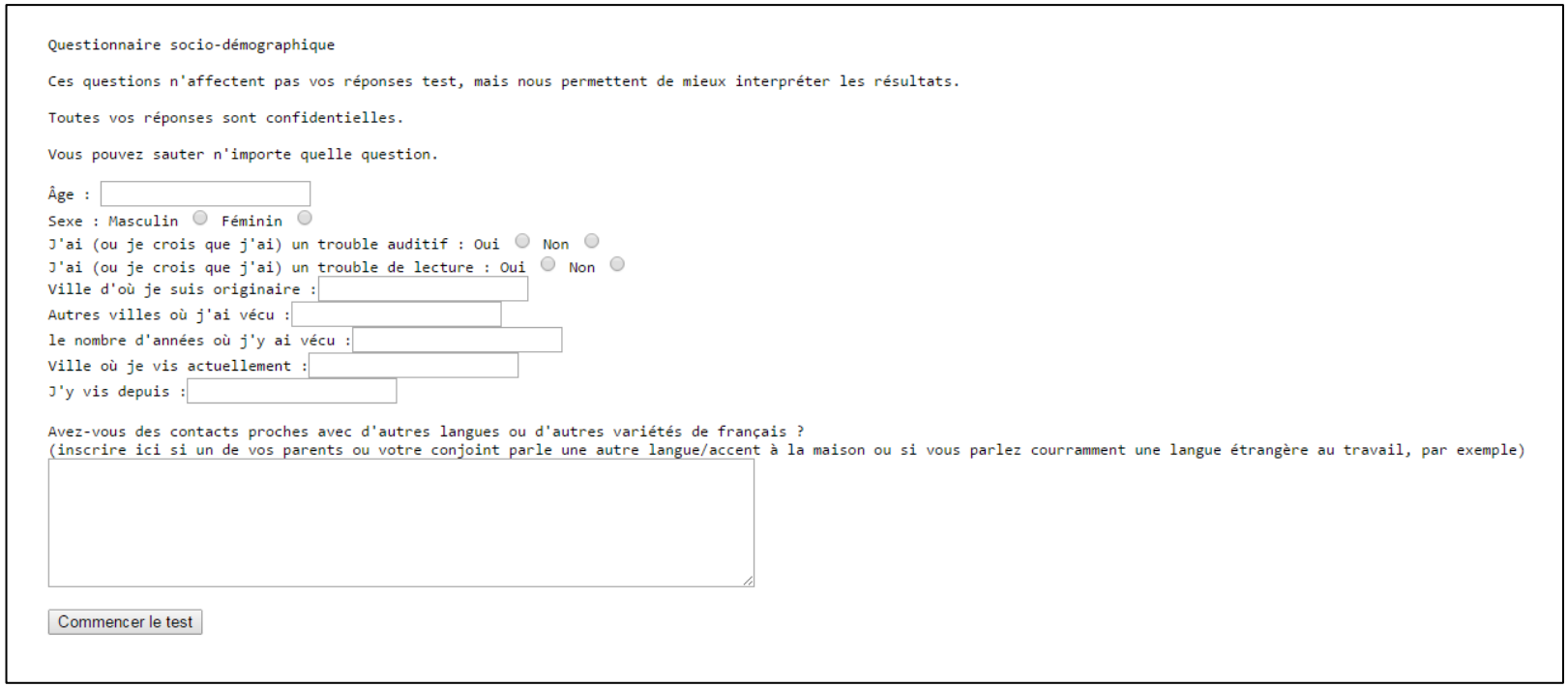
```
        if list((duree[phoneme_actuel].keys()))[0] in voyelles:
          syllable[-1].set_voyelle(phoneme_actuel)
        if phoneme_actuel in f0:
          syllable[-1].add_f0(f0[phoneme_actuel])
        syllable[-1].add_duree(duree[phoneme_actuel])
        phoneme_actuel += 1
    else:
      while phoneme_actuel < phoneme_num:
        phoneme_actuel += 1

for s in syllable:
  pass
  #print(str(s))

if i in f0_moyen_list:          #range(1,25): #f0_moyen (1,3,5; 9,11,13; 20,22,24)
  moyen(f0)

elif i in registre_list:        #range(25,49): #f0_max - f0_min (25,27,29; 33,35,37; 44,46,48)
  registre(f0)

elif i in syl_dur_list:         #range(49, 73): #(49,51,53; 59,61,63; 66,68,70)
  syl_dur(duree, syllable)

elif i in syl_f0_list:          #range(73, 97): #(73,75,79; 83,85,87; 92,94,96)
  syl_f0(f0, duree, syllable)

elif i in duree_v_list:         #range(97, 121): #(97,99,103; 105,107,109; 116,118,120)
  duree_v(duree, syllable)

elif i in duree_comb_list:      #range(121, 145): #(121,123,127; 131,133,135; 138,140,144)
  cible_id = duree_comb_list.index(i)%3 +1
  duree_v(duree, syllable, cible_id)
  syl_dur(duree, syllable, cible_id)
  duree_v(duree, syllable, cible_id)

elif i in f0_comb_list:         #range(145, 169): #(147,149,151; 155,157,159; 162,164,166)
  cible_id = f0_comb_list.index(i)%3 +1
  registre(f0, cible_id)
  syl_f0(f0, duree, syllable, cible_id)
  moyen(f0, cible_id)

elif i in comb_list:            #range(169, 193):#(169,171,173; 177,181,183; 186,188,192)
  cible_id = comb_list.index(i)%3 +1
  registre(f0, cible_id)
  syl_f0(f0, duree, syllable, cible_id)
  moyen(f0, cible_id)
  duree_v(duree, syllable, cible_id)
  syl_dur(duree, syllable, cible_id)
  duree_v(duree, syllable, cible_id)

elif i in comb_inv_list:        #range(193, 217): #(193,195,199; 203,205,207; 210,214,216)
  cible_id = comb_inv_list.index(i)%3 +1
  '''
  if i%8:
    if round(i%8/2+0.1) == 1:
      i = 1
      j = 4
    elif round(i%8/2+0.1) == 2:
      i = 2
      j = 3
    elif round(i%8/2+0.1) == 3:
      i = 3
```

```
      j = 2
    elif round(i%8/2+0.1) == 4:
      i = 4
      j = 1
  else:
    i = 4
    j = 1
  '''
  registre(f0, cible_id)
  syl_f0(f0, duree, syllable, cible_id)
  moyen(f0, cible_id)
  if cible_id == 1:
    cible_id = 3
  elif cible_id == 3:
    cible_id = 1
  #i = j
  duree_v(duree, syllable, cible_id)
  syl_dur(duree, syllable, cible_id)
  duree_v(duree, syllable, cible_id)


for k in range(len(duree)):
  #print(list(duree[k].keys())[0], end=" ")
  #print(round(float(list(duree[k].values())[0]), 2), end=" ")
  fout.write(str(list(duree[k].keys())[0]) + " ")
  fout.write(str(round(float(list(duree[k].values())[0]), 2)) + " ")
  if k in f0:
    #print(f0[k])
    for i in f0[k]:
      #print(list(i.keys())[0], end=" ")
      fout.write(str(list(i.keys())[0]) + " ")
      if float(list(i.values())[0]) % 1:
        #print(round(float(list(i.values())[0]), 2), end=" ")
        fout.write(str(round(float(list(i.values())[0]), 2)) + " ")
      else:
        #print(list(i.values())[0], end=" ")
        fout.write(str(list(i.values())[0]) + " ")

  #print()
  fout.write("\n")

fout.close()
```

APPENDICE F
INTERFACE DU TEST DE PERCEPTION



Formulaire de consentement



Questionnaire socio-démographique

Ces questions n'affectent pas vos réponses test, mais nous permettent de mieux interpréter les résultats.

Toutes vos réponses sont confidentielles.

Vous pouvez sauter n'importe quelle question.

Âge :

Sexe : Masculin Féminin

J'ai (ou je crois que j'ai) un trouble auditif : Oui Non

J'ai (ou je crois que j'ai) un trouble de lecture : Oui Non

Ville d'où je suis originaire :

Autres villes où j'ai vécu :

le nombre d'années où j'y ai vécu :

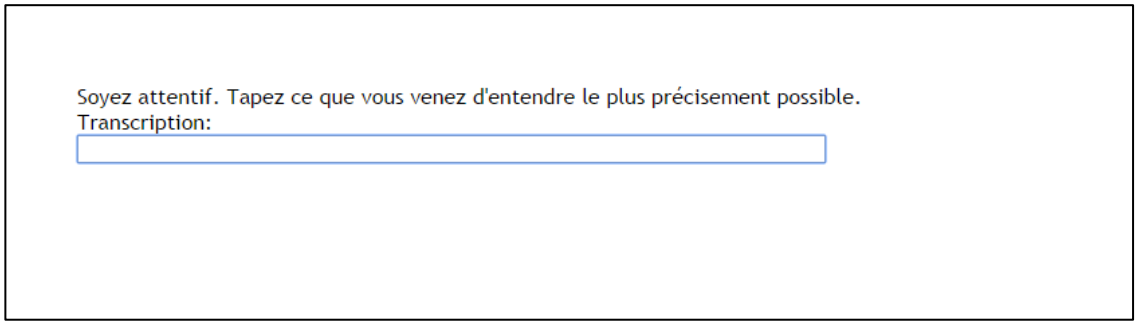
Ville où je vis actuellement :

J'y vis depuis :

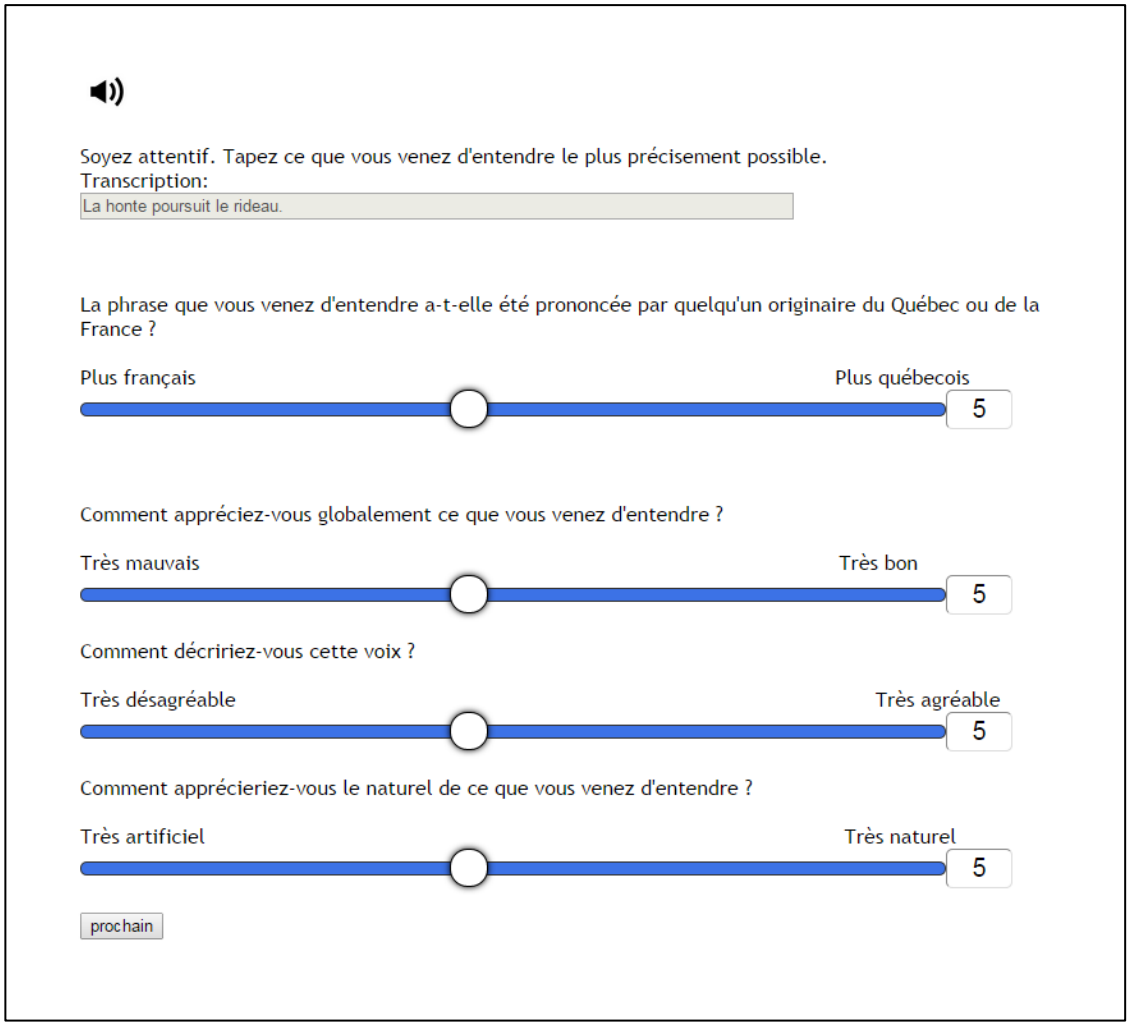
Avez-vous des contacts proches avec d'autres langues ou d'autres variétés de français ?
(inscrire ici si un de vos parents ou votre conjoint parle une autre langue/accent à la maison ou si vous parlez courrament une langue étrangère au travail, par exemple)

Commencer le test

Questionnaire sociodémographique



Interface pour la tâche d'intelligibilité



Interface pour l'évaluation de l'origine géographique et pour l'évaluation de la qualité

APPENDICE G
SCRIPT PRAAT DE DÉLEXICALISATION

```
##
##Ce script permet de transférer la F0, durée et intensité d'une phrase initiale
##à une phrase porteuse composée de la répétition d'une seule syllabe.
##
##Il segmente d'abord le fichier original en syllabes. Les données F0, durée et intensité sont mesurées
##sur les syllabes. Elles sont ensuite transférées à la syllabe porteuse.
##
##

form Transfert prosodique
  comment Ce script vous permet de transférer l'information prosodique (F0, durée, intensité) d'une phrase à une
  comment syllabe porteuse répétée.
  comment Dans votre directory, les fichiers doivent être accompagnés d'un textgrid où les syllabes ont été
  comment segmentées et annotées.
  sentence directory C:\Users\premont_a\Dropbox\Mémoire\enregistrements\F1\corrections\
  sentence output_directory C:\Users\premont_a\Dropbox\Mémoire\enregistrements\F1\corrections\phrasesproso
  comment Quel fichier désirez-vous avoir comme syllabe porteuse ?
  sentence fichierPorteuse C:\Users\premont_a\Dropbox\Mémoire\enregistrements\la_ma\la_isole3.wav
  comment Inscrivez ici vos préférences pour la manipulation.
  comment (Vous pouvez garder les paramètres par défaut)
  positive timeStep 0.01
  positive minPitch 75
  positive maxPitch 600
  comment Inscrivez ici vos préférences pour la modification de l'intensité.
  comment (Vous pouvez garder les paramètres par défaut)
  positive minPitchInt 100
  real timeStepInt 0
endform

Create Strings as file list... list 'directory$'/*.Sound
numberOfFiles = Get number of strings

for ifile from 1 to numberOfFiles
		writeInfoLine ("debut")
  select Strings list
  fichier_temp$ = Get string... ifile
  fichier$ = replace$ (fichier_temp$, ".Sound", "", 1)

  Read from file... 'directory$'/'fichier$'.Sound
  Read from file... 'directory$'/'fichier$'.TextGrid

  @trouver_duree_silence_debut()

  @extraire_syllabe()

  #Sélectionner la syllabe porteuse à partir d'un fichier
  Read from file... 'fichierPorteuse$'
  Rename... syllabeporteuse

  #Trouver la durée de la syllabe porteuse
  select Sound syllabeporteuse
  dureePorteuse = Get duration

  nbSyllabes = 0
  nbPauses = 0
```

```
 @trouver_nb_syllabe()
  @duree()
  @concatene_syllabe()
  @changer_pitch()
  @intensite()
  @resynth_finale()
  pause
  @remove()
endfor
select all
Remove

 procedure remove ()
              select all
              minus Strings list
              Remove
 endproc

  procedure trouver_duree_silence_debut()

    #Extraire les syllabes

    select Sound 'fichier$'
    plus TextGrid 'fichier$'
    Extract all intervals... 1 no

              select Sound untitled
       Remove
              select Sound untitled
              dureeDebut = Get duration
              select all
              minus Strings list
              minus Sound 'fichier$'
              minus TextGrid 'fichier$'
              Remove

  endproc

procedure resynth (i)
#writeInfoLine(i)
#Resynthétiser avec la durée changée
    select Manipulation 'nomManip$'
    Get resynthesis (overlap-add)
    Rename... lasynth 'i'
endproc

procedure resynth_finale ()
#Resynthétiser avec la durée et la F0 changées
  select Sound phraseDuree_int
  Save as WAV file... 'output_directory$'/'fichier$'_synth.wav
endproc

procedure extraire_syllabe ()

  #Extraire les syllabes

  select Sound 'fichier$'
  plus TextGrid 'fichier$'
  Extract non-empty intervals... 1 no

endproc
```

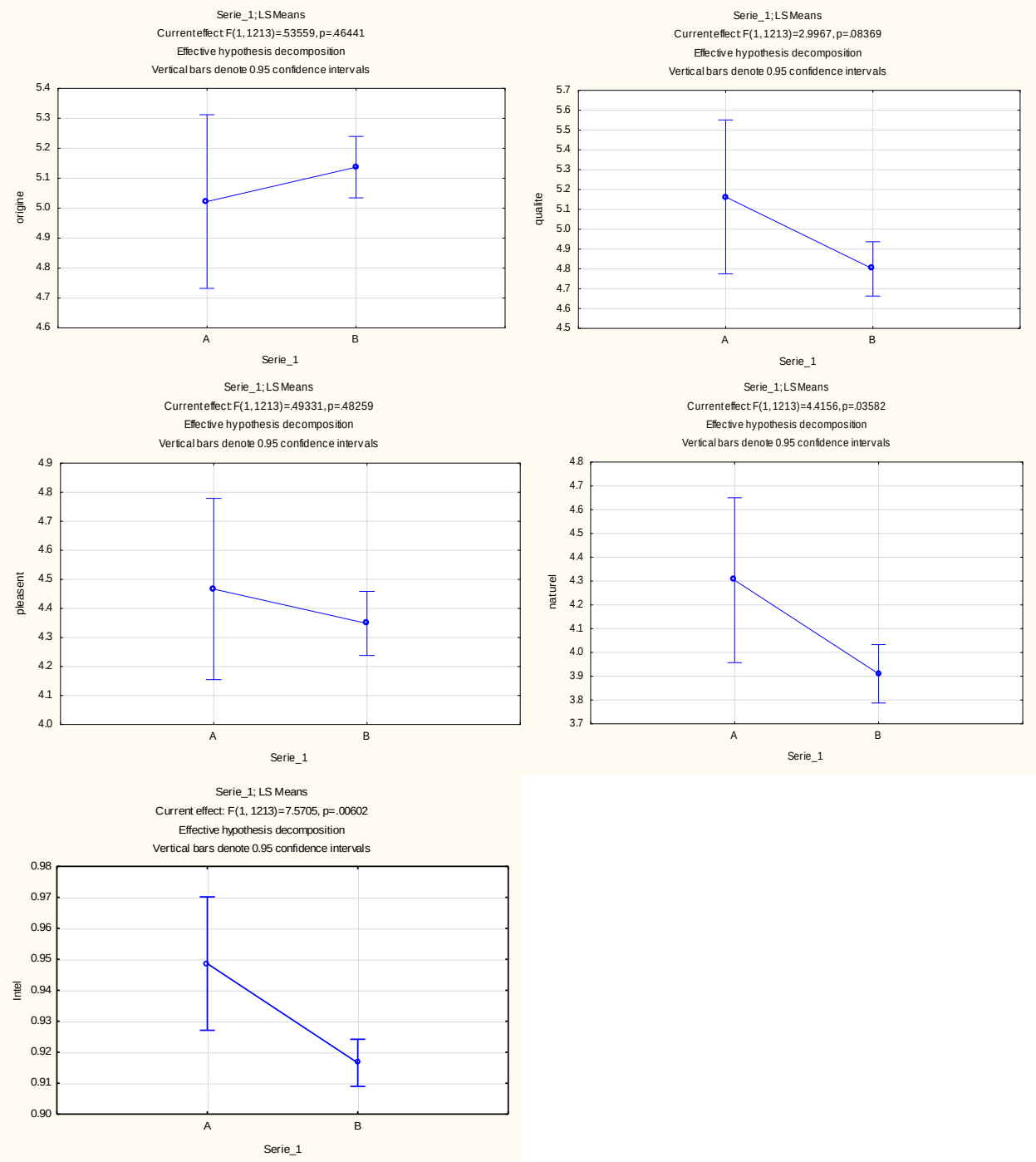
```
procedure intensite ()
    #Intensité
    select Sound 'fichier$'
    To Intensity... minPitchInt timeStepInt
    select Intensity 'fichier$'
    Down to IntensityTier
    Shift times by... -'dureeDebut'
    select Sound phraseDuree
    plus IntensityTier 'fichier$'
    Multiply
endproc

procedure trouver_nb_syllabe ()
  #Trouver le nombre de syllabes & pauses
  #Créer des copies de la syllabe porteuse
  #Créer des pauses vides
  select TextGrid 'fichier$'
  nInterv = Get number of intervals... 1

  for j from 1 to 'nInterv'
    select TextGrid 'fichier$'
    lab$ = Get label of interval... 1 'j'

                if lab$ != "" and lab$ != "pause"
      nbSyllabes = nbSyllabes + 1
      select Sound syllabeporteuse
      nomCopie$ = "la" + "'nbSyllabes'"
      Copy... 'nomCopie$'
      #nom$ = nomCopie$ + "_"
      #select Sound 'nom$'
      #To Manipulation... timeStep minPitch maxPitch
    endif

                if lab$ = "pause"
      nbPauses = nbPauses + 1
                  nom_pause$ = "pause" + "'nbPauses'"
                  appendInfo ( nom_pause$, newline$)
      Create Sound from formula... 'nom_pause$' mono 0 1 44100 0
    endif

  endfor
endproc

procedure concatene_syllabe ()

  #Concaténer les syllabes

  select Sound lasynth_1
  for k from 2 to nbSyllabes + nbPauses
    plus Sound lasynth_'k'
  endfor
  Concatenate
  Rename... phraseDuree

endproc

procedure changer_pitch ()
  # select Sound phraseDuree_int
  select Sound phraseDuree
  To Manipulation... timeStep minPitch maxPitch

 #Changer le pitch
```

```
 select Sound 'fichier$'
 To Manipulation... timeStep minPitch maxPitch
 select Manipulation 'fichier$'
 Extract pitch tier
 Shift times by... -'dureeDebut'
 select Manipulation phraseDuree
 plus PitchTier 'fichier$'
 Replace pitch tier
 select Manipulation phraseDuree
 Get resynthesis (overlap-add)
 Rename... phraseDuree

endproc

procedure duree ()

 #Ajuster la durée des syllabes et des pauses
 offset_pauses = 0
 for m from 1 to nbSyllabes + nbPauses
            #choisi tout pour pouvoir sélectionner les sounds un par un
            select all

  #Trouve la durée de chaque syllabe initiale
  #Manipule la durée de la syllabe porteuse
            #+1 pour skipper la phrase initiale
  nom_son$ = selected$ ("Sound", ('m')+1)
            #récupère le nom du sound prochain dans la liste complete
            appendInfo (nom_son$, newline$)
            #choisi le sound suivant
            select Sound 'nom_son$'
  dureeInitiale = Get duration

            if nom_son$ = "pause"
           durationFactor = dureeInitiale/1
                     offset_pauses = offset_pauses + 1
                     nom_pauses$ = "pause" + "'offset_pauses'"
                     select Sound 'nom_pauses$'
                     To Manipulation... timeStep minPitch maxPitch
                     Create DurationTier... 'nom_pauses$' 0 'dureeInitiale'
                     Add point... 0 'durationFactor'
                     select Manipulation 'nom_pauses$'
                     plus DurationTier 'nom_pauses$'
                     Replace duration tier

           #Resynthétiser avec la durée changée
           select Manipulation 'nom_pauses$'
           Get resynthesis (overlap-add)
           Rename... lasynth 'm'
            else
           durationFactor = dureeInitiale/dureePorteuse
           indice_la = m - offset_pauses
                     nomManip$ = "la" + "'indice_la'" + "_"
                     select Sound 'nomManip$'
                     To Manipulation... timeStep minPitch maxPitch
                     Create DurationTier... 'm' 0 'dureeInitiale'
                     Add point... 0 'durationFactor'
                     select Manipulation 'nomManip$'
                     plus DurationTier 'm'
                     Replace duration tier

           #Resynthétiser avec la durée changée
```

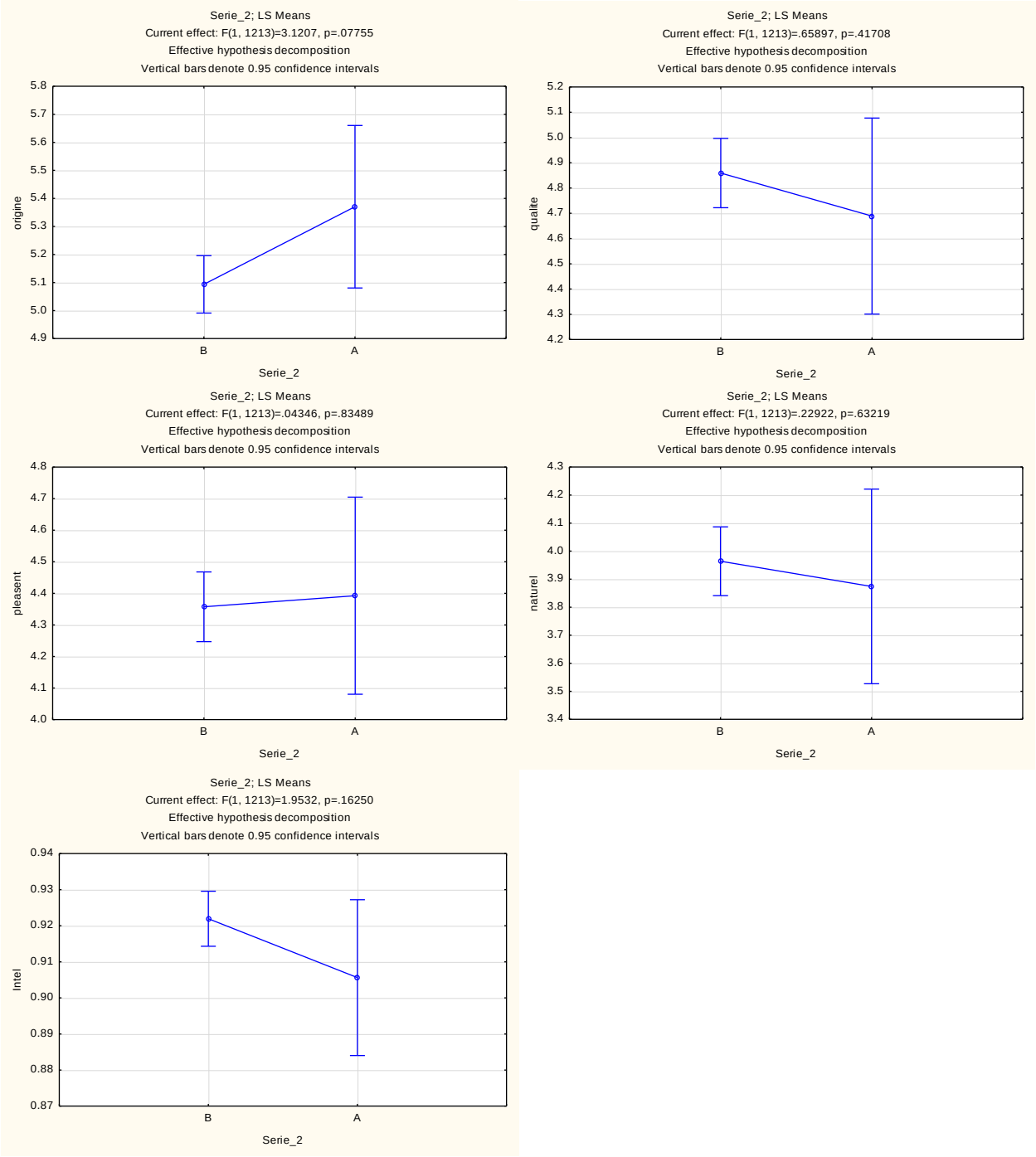
```
            select Manipulation 'nomManip$'
            Get resynthesis (overlap-add)
            Rename... lasynth 'm'
             endif
  endfor
endproc
```

APPENDICE H
ANOVAS DES SÉRIES VS L'ENSEMBLE DES RÉPONSES

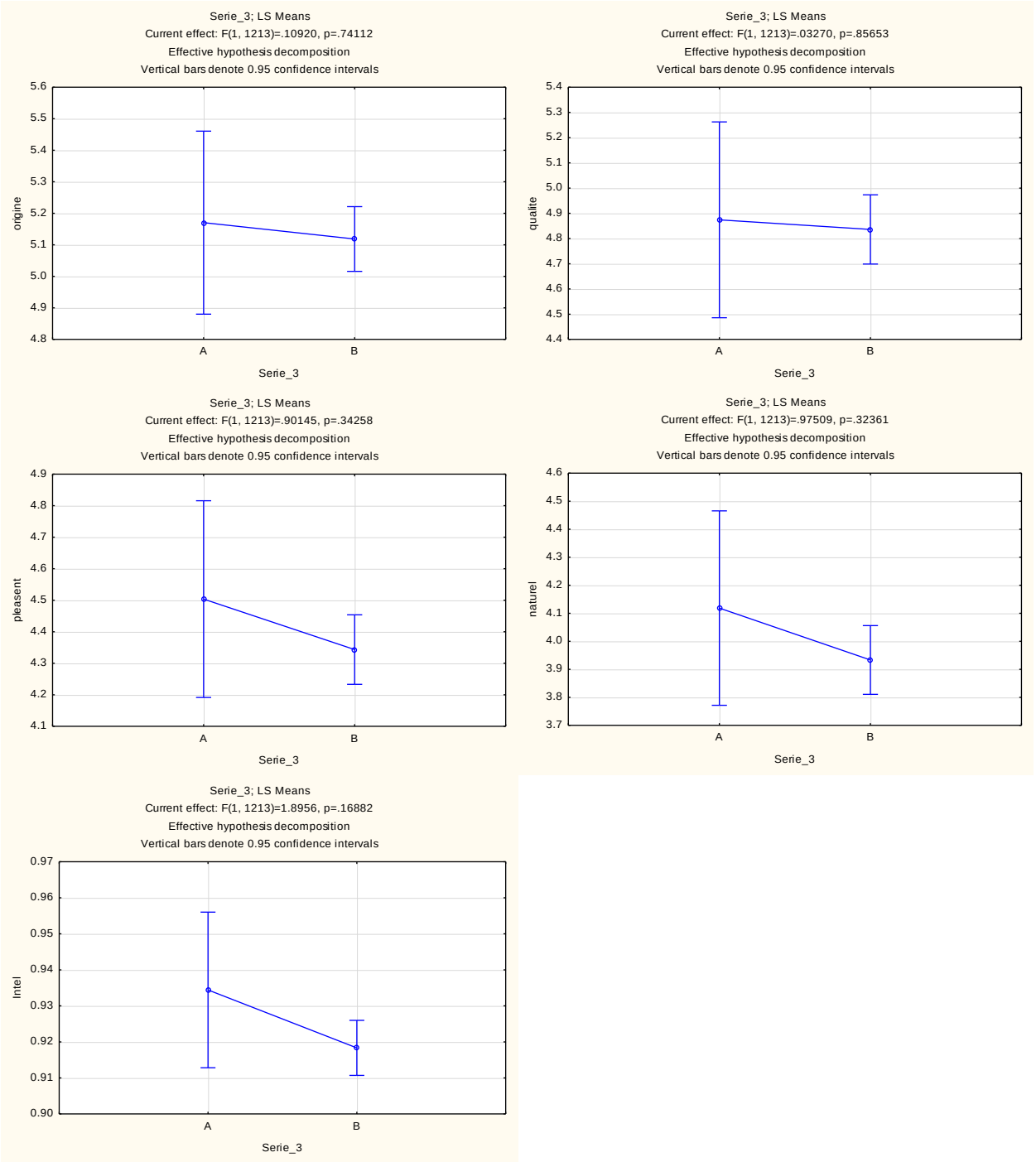
Appendice H-1 : ANOVAs série f0 moy (A) vs l'ensemble des réponses (B) pour les cinq questions



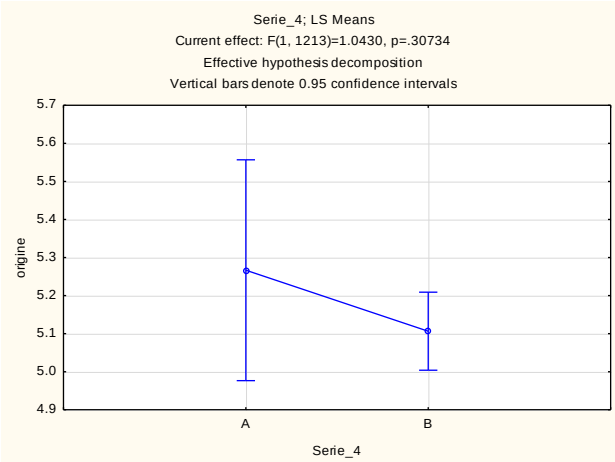
Appendice H-2 : ANOVAs série registre (A) vs l'ensemble des réponses (B) pour les cinq questions

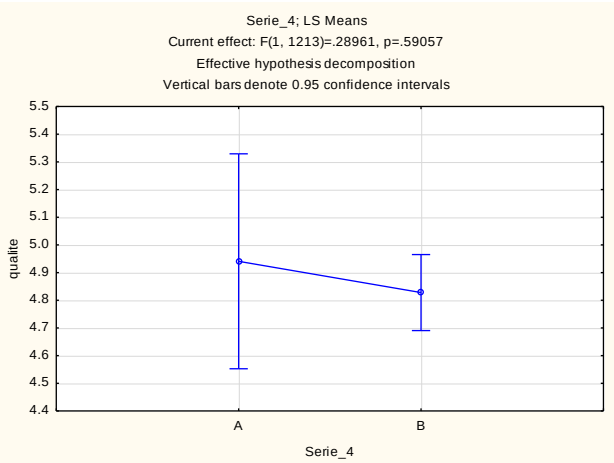


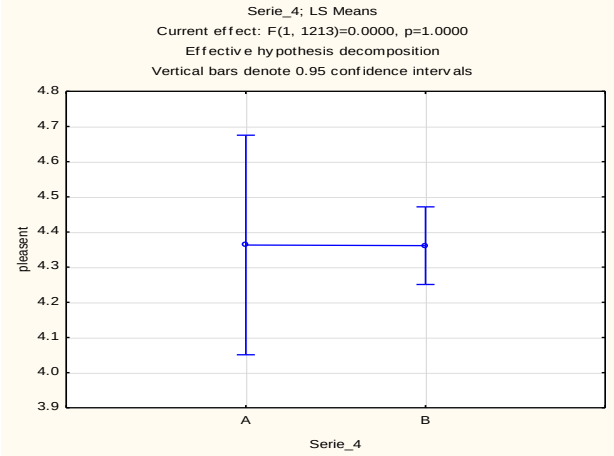
Appendice H-3 ANOVAs série syldur TH/TB (A) vs l'ensemble des réponses (B) pour les cinq questions

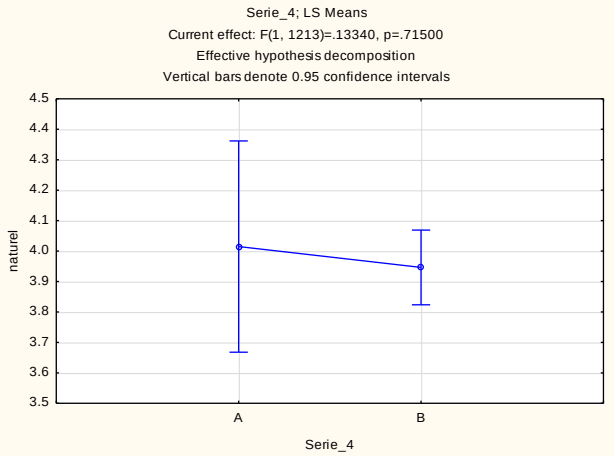


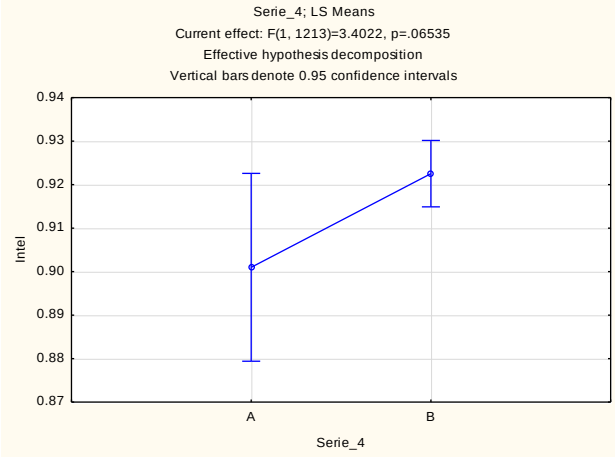
Appendice H-4 ANOVAs série F0 moy TH/TB (A) vs l'ensemble des réponses (B) pour les cinq questions



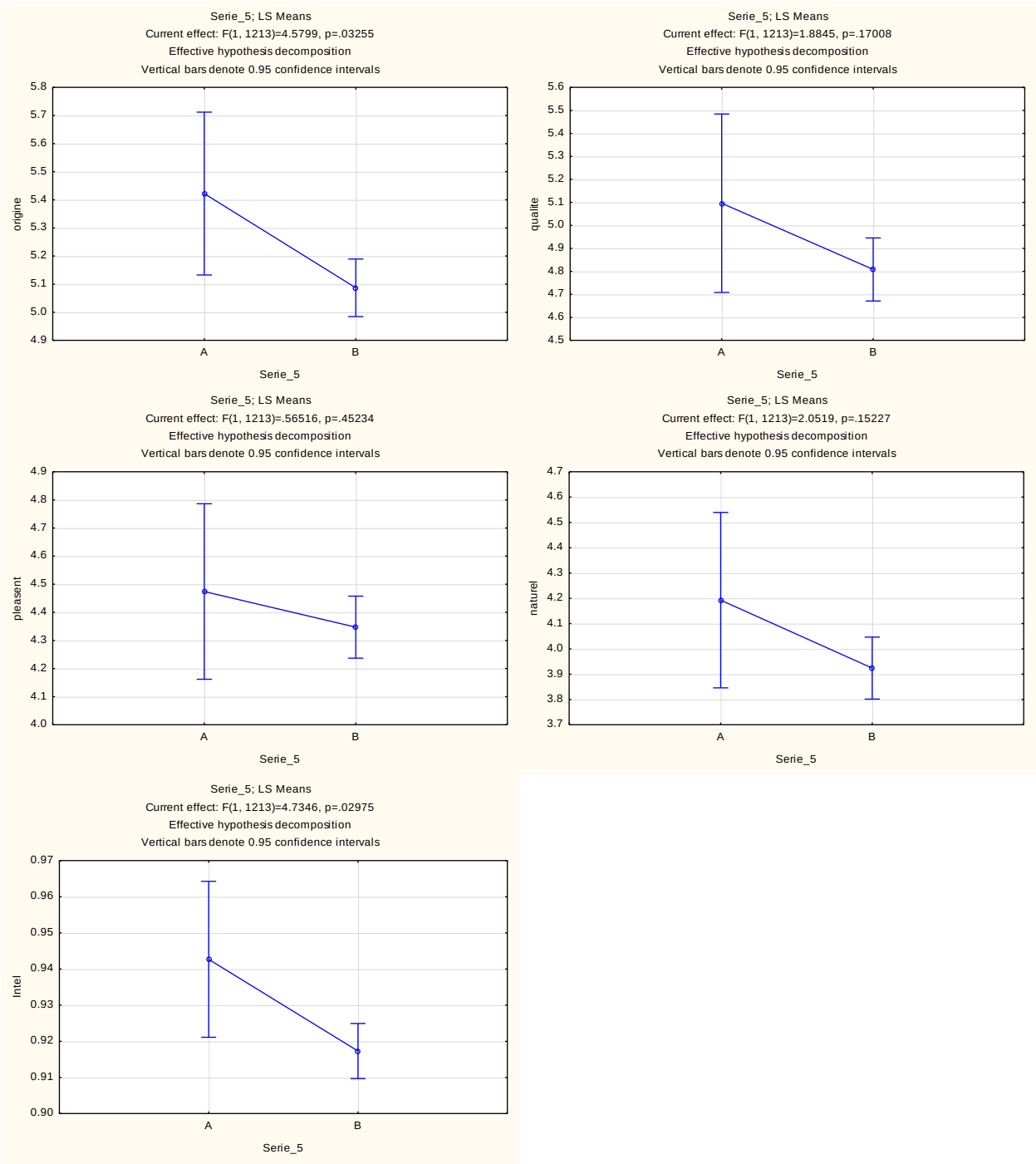




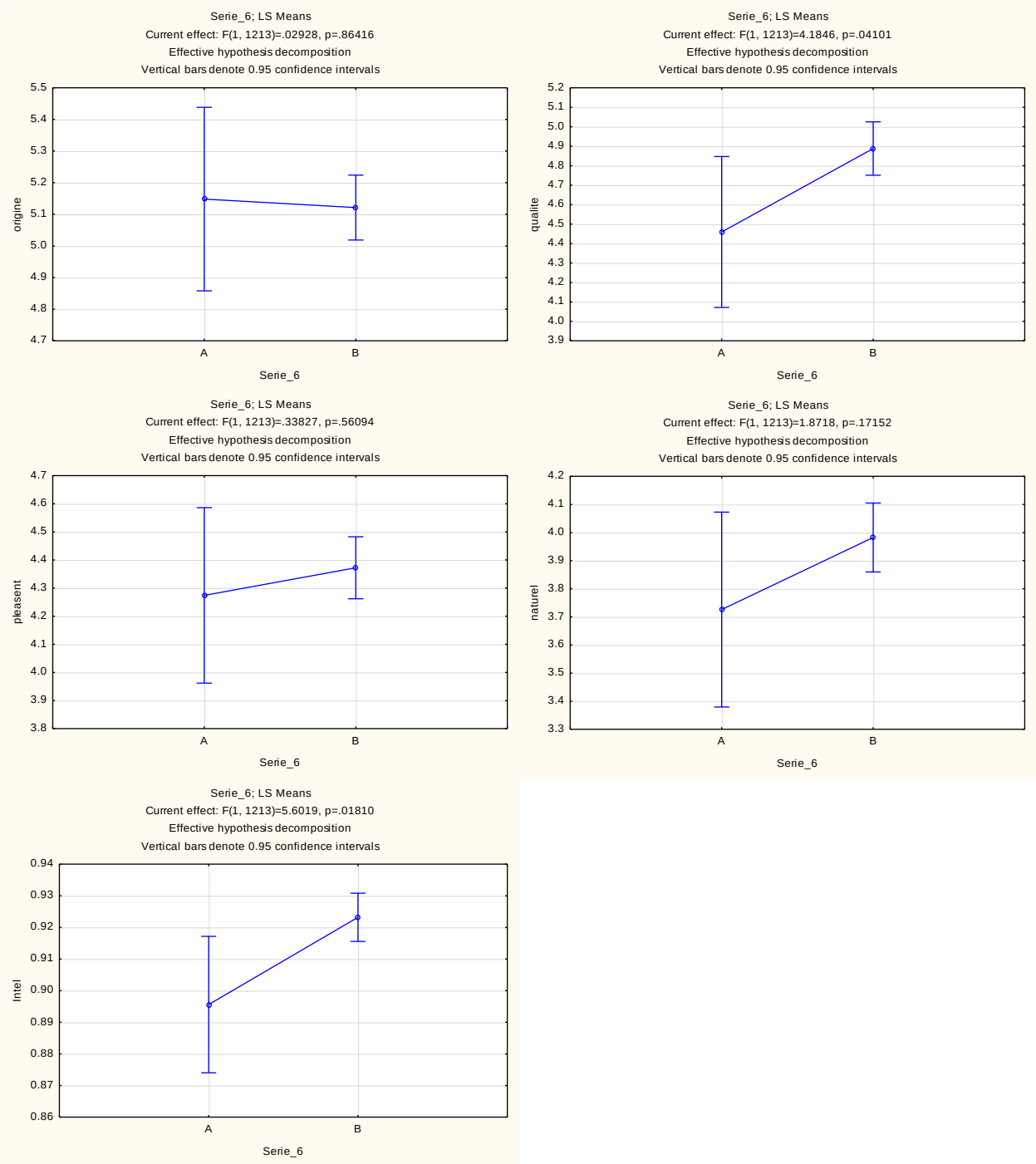




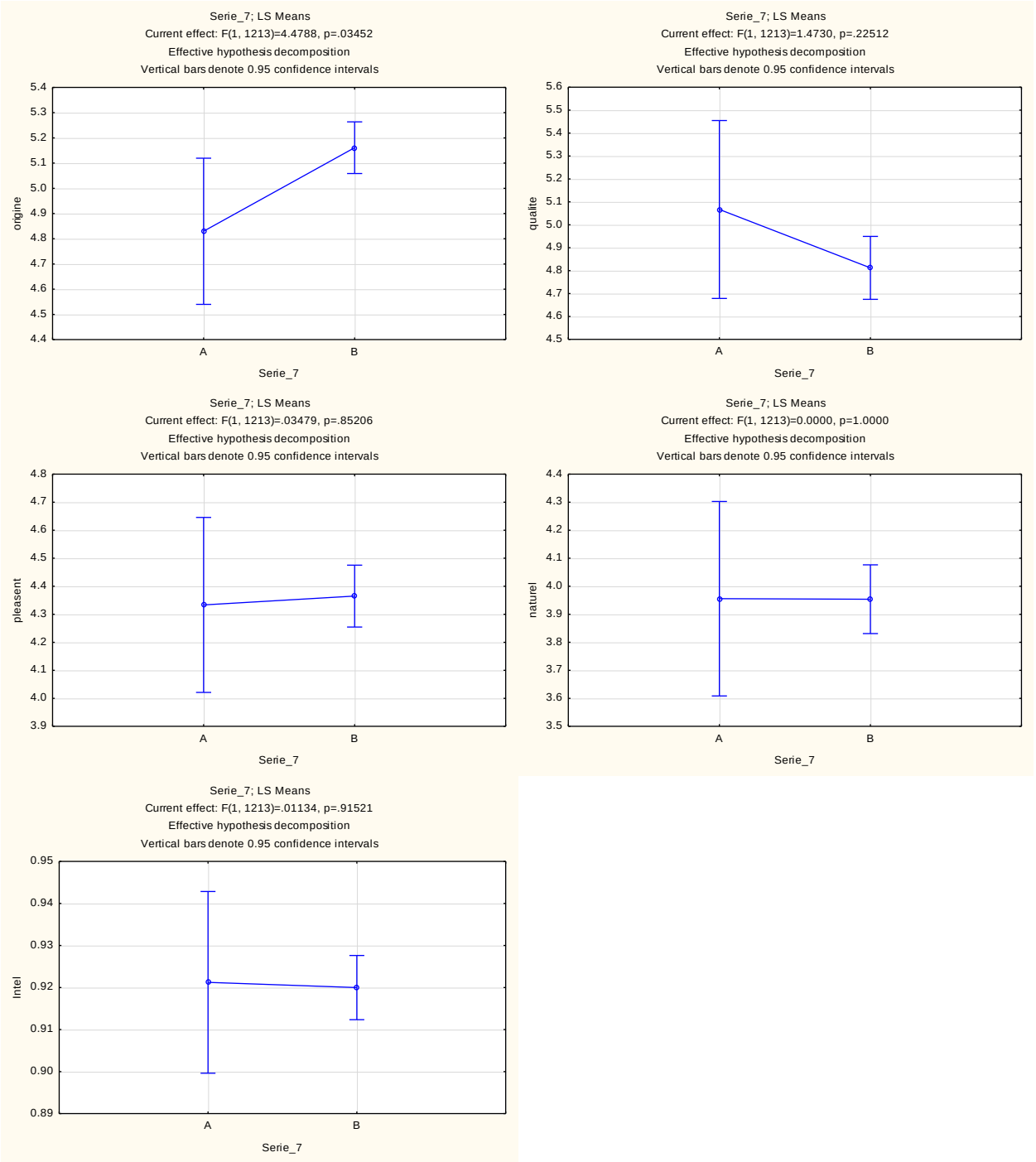
Appendice H-5 ANOVAs série duréeV (A) vs l'ensemble des réponses (B) pour les cinq questions



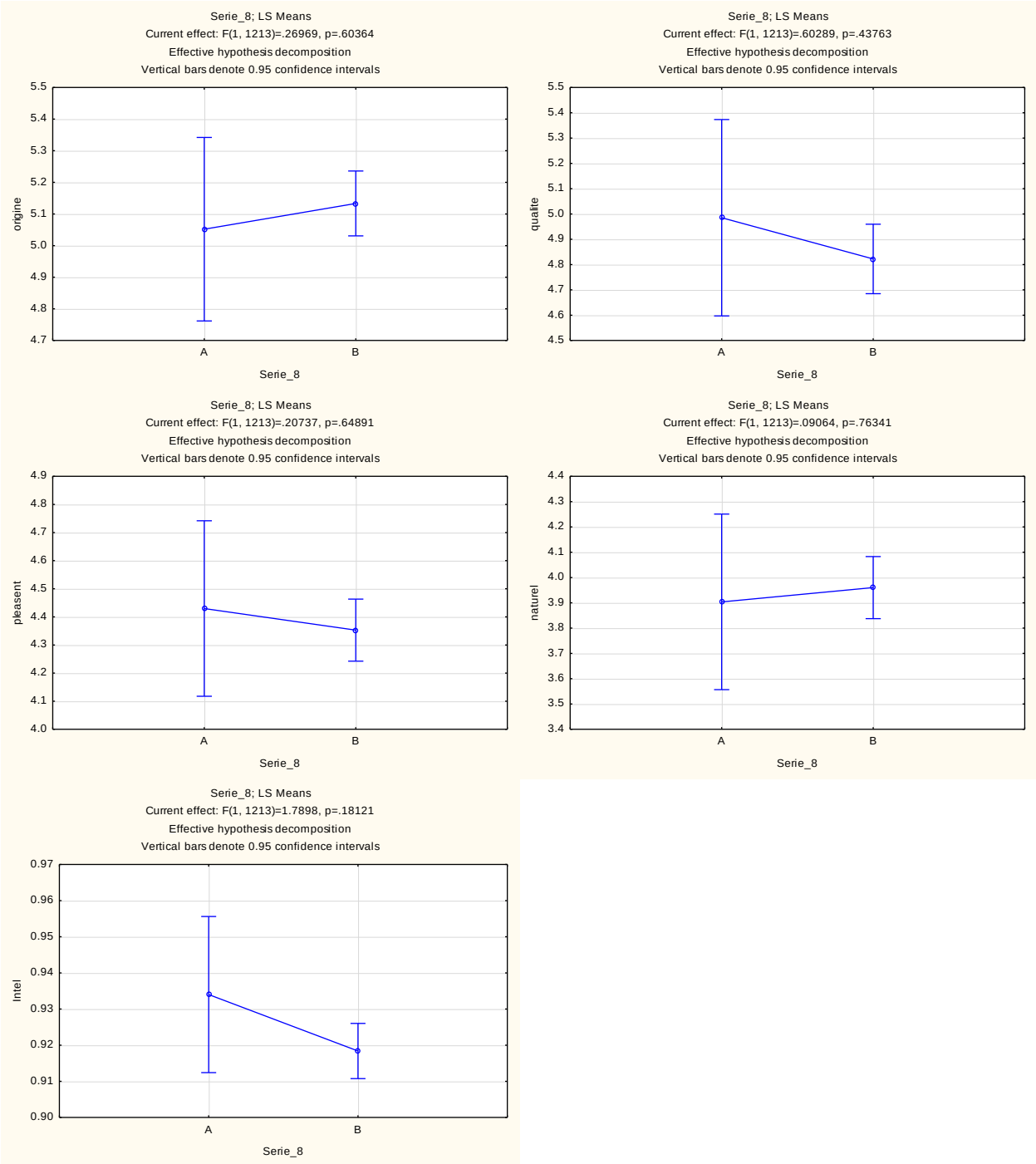
Appendice H-6 ANOVAs série durée (A) vs l'ensemble des réponses (B) pour les cinq questions



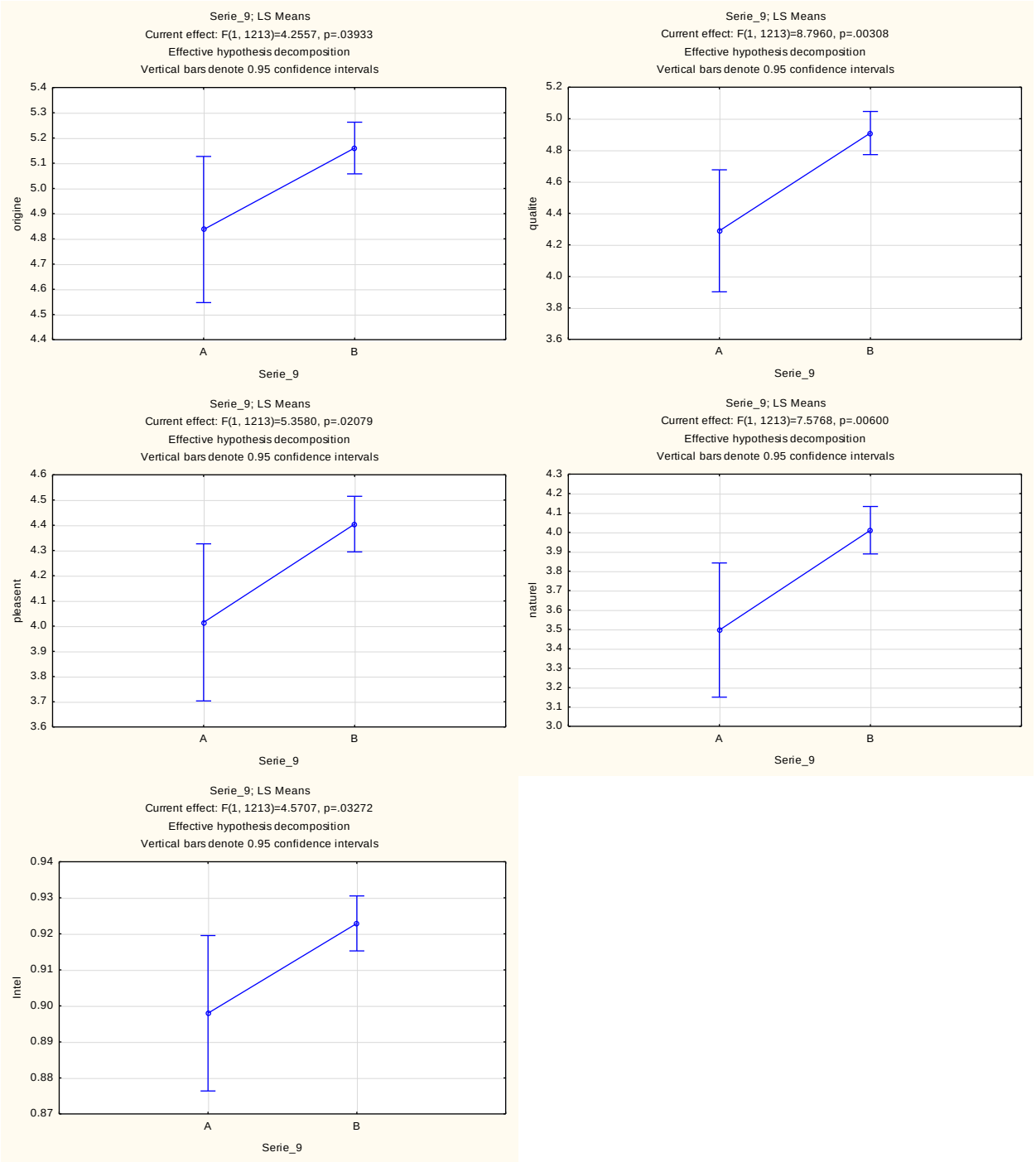
Appendice H-7 ANOVAs série F0 (A) vs l'ensemble des réponses (B) pour les cinq questions



Appendice H-8 ANOVAs série F0 et durée (A) vs l'ensemble des réponses (B) pour les cinq questions

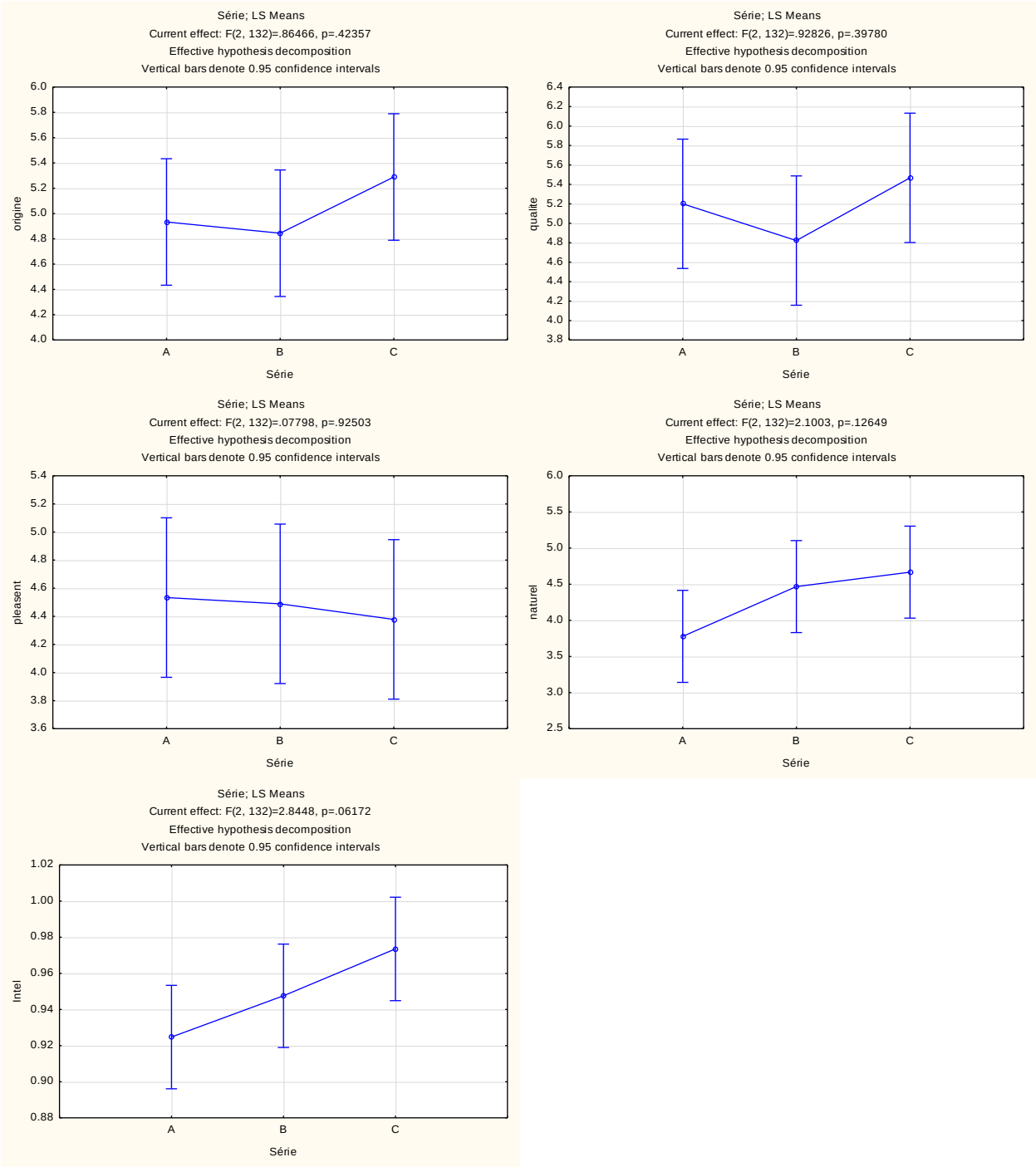


Appendice H-9 ANOVAs série F0 et durée inversées (A) vs l'ensemble des réponses (B) pour les cinq questions

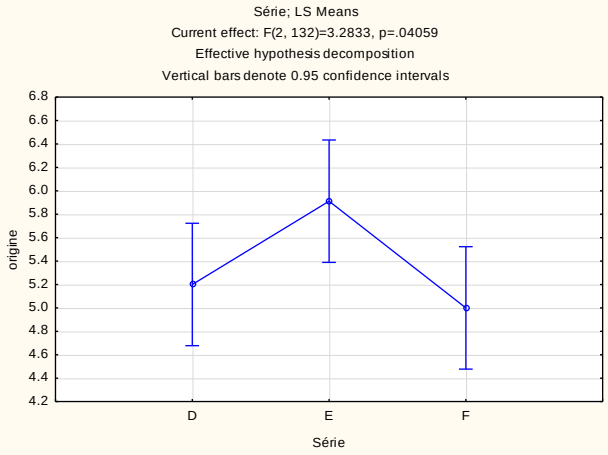


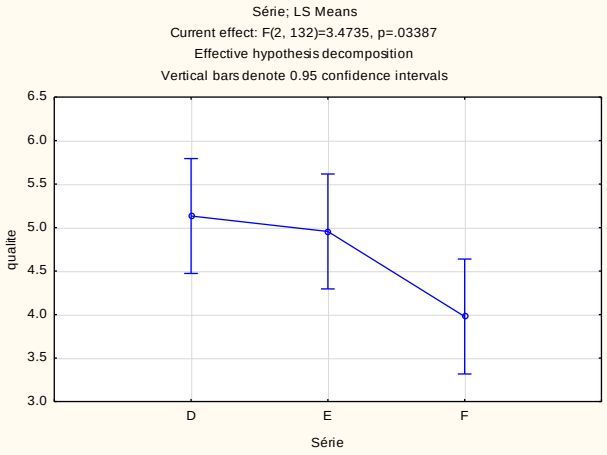
APPENDICE I
ANOVAS DES COHÉRENCES INTERNES DES SÉRIES

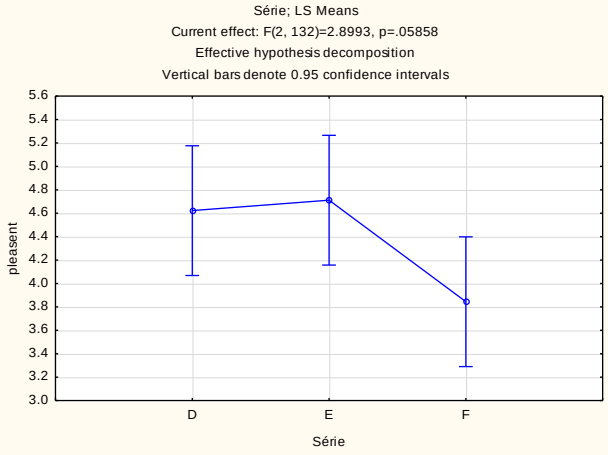
Appendice I-2 ANOVAs série register – coherence interne

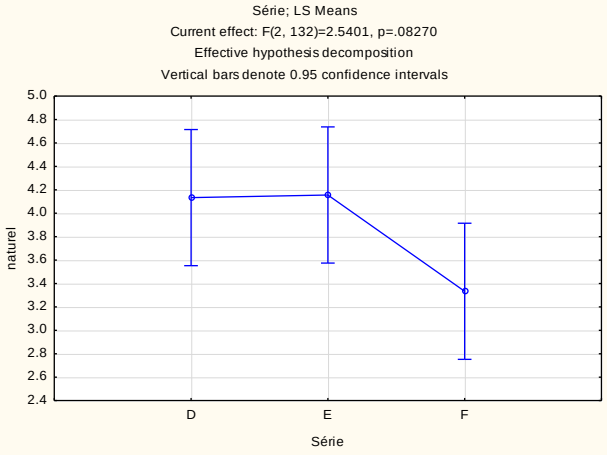


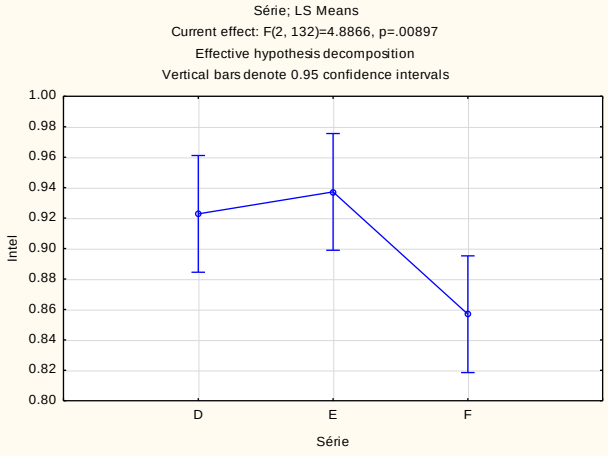
Appendice I- 2 ANOVAs série registre – cohérence interne



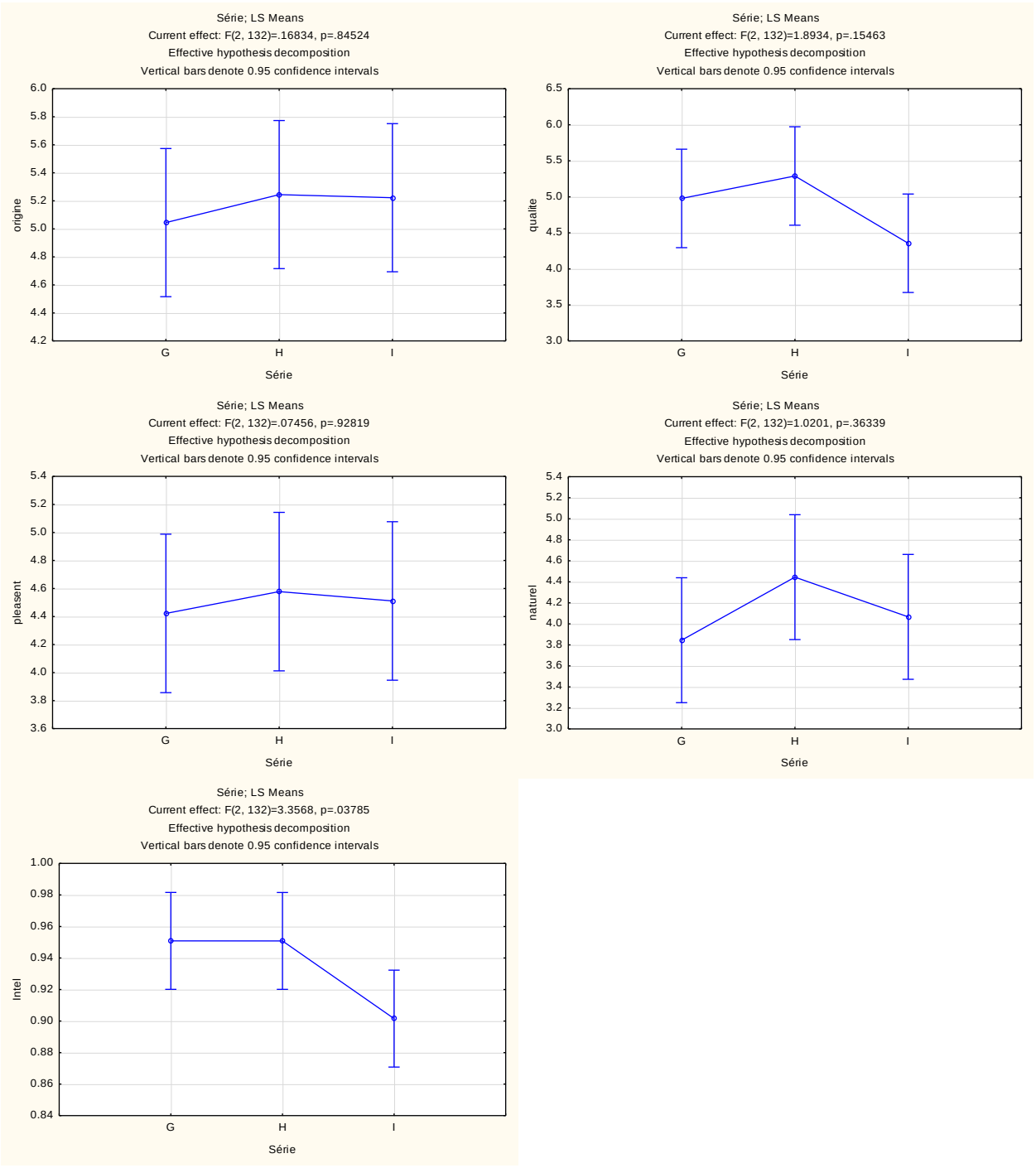




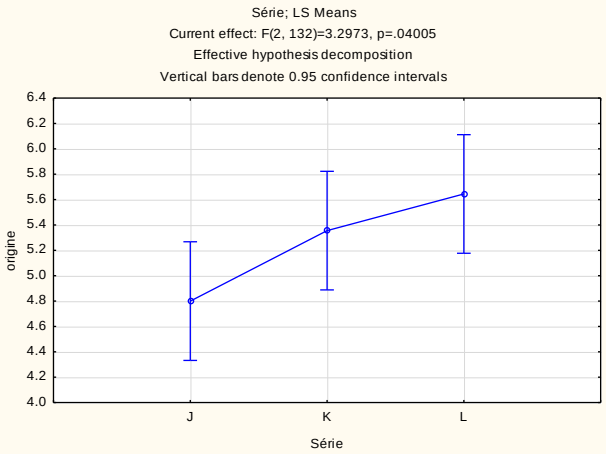


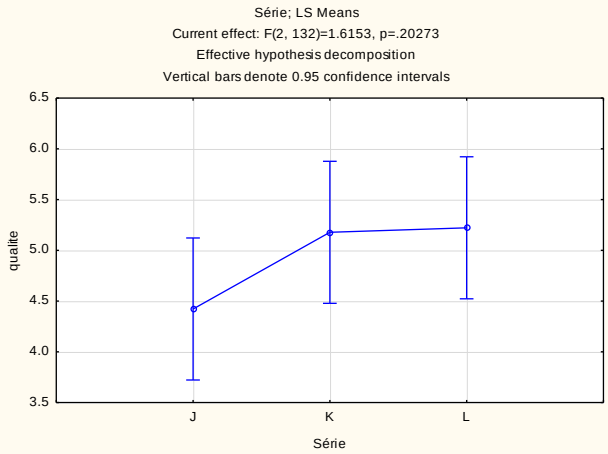


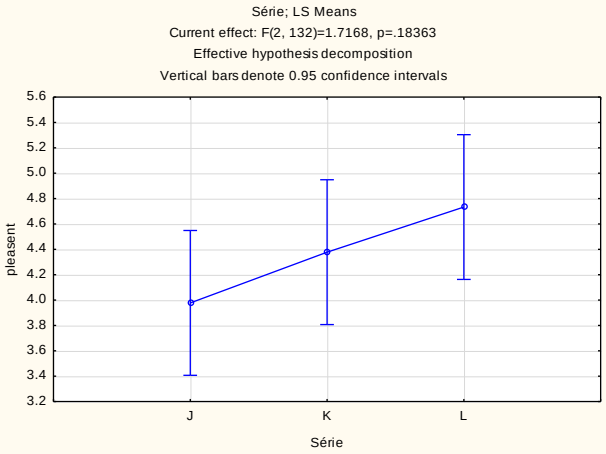
Appendice I - 3 ANOVAs syldur TH/TB – cohérence interne

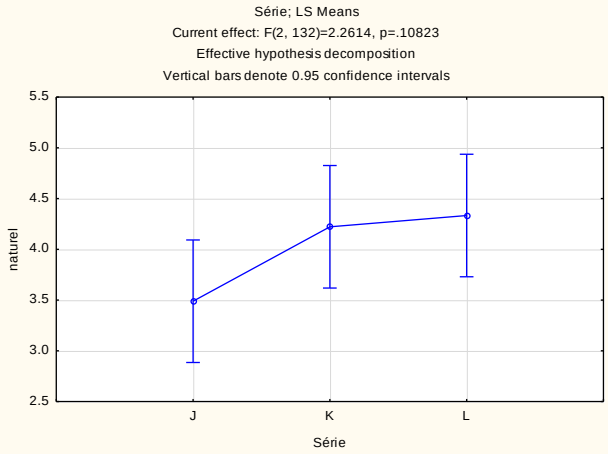


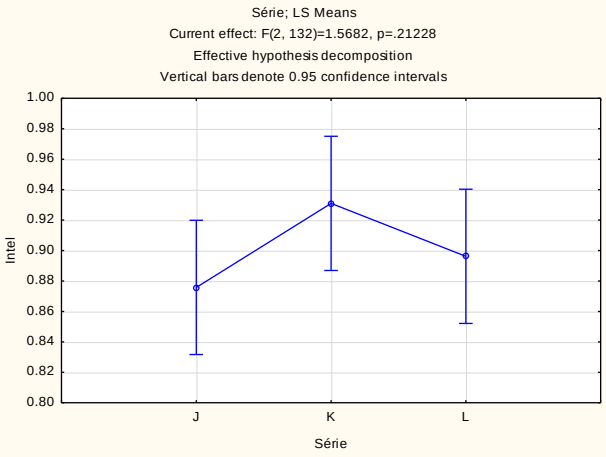
Appendice I - 4 ANOVAs série F0 TH/TB – cohérence interne



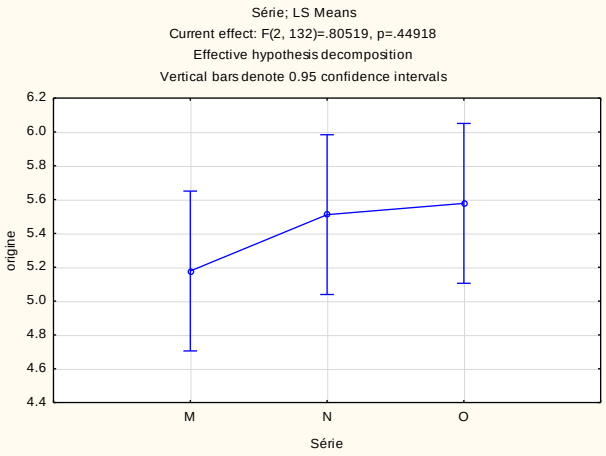


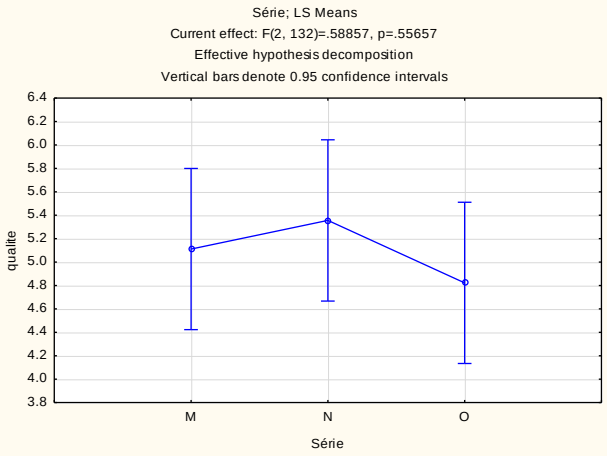


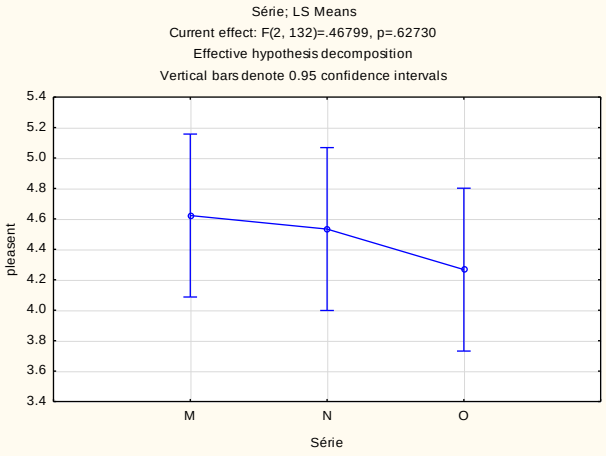


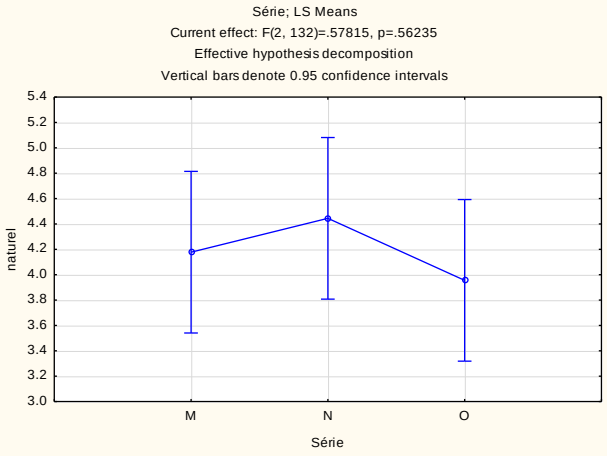


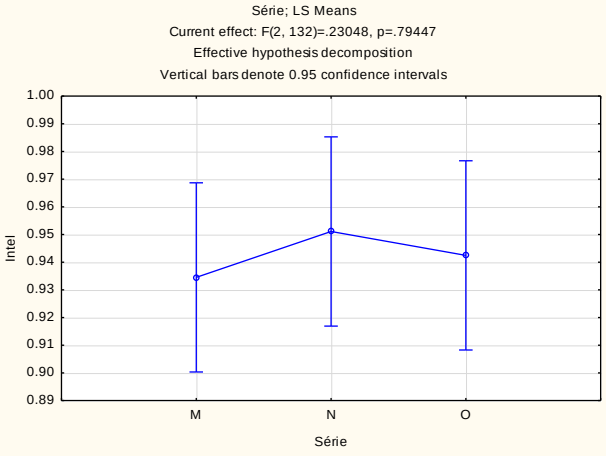
Appendice I - 5 ANOVAs série dureeV – cohérence interne



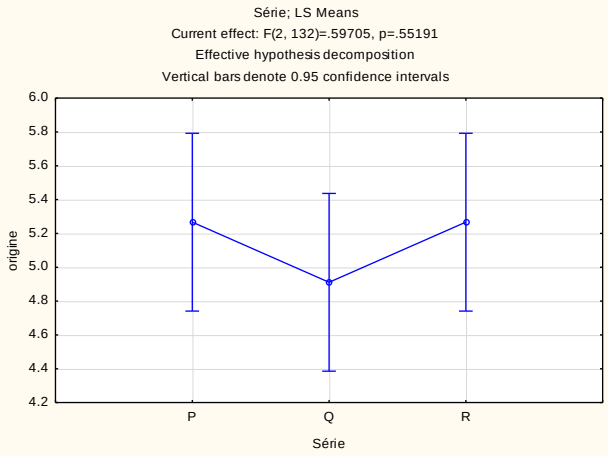


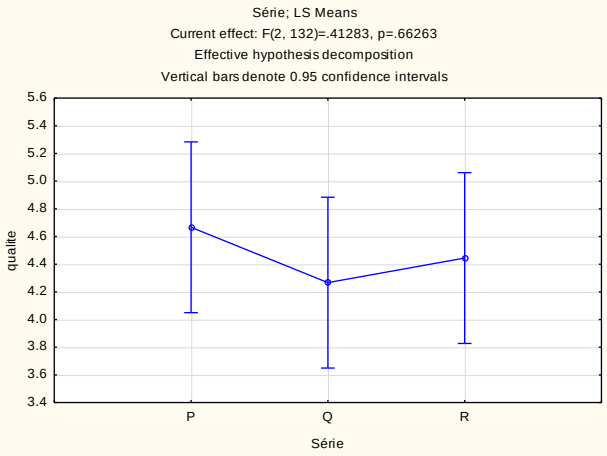


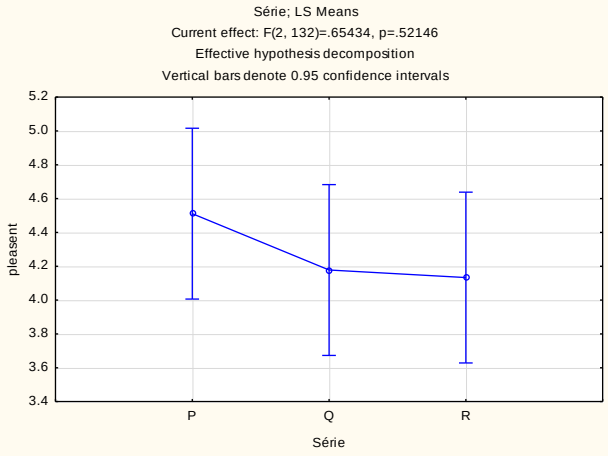


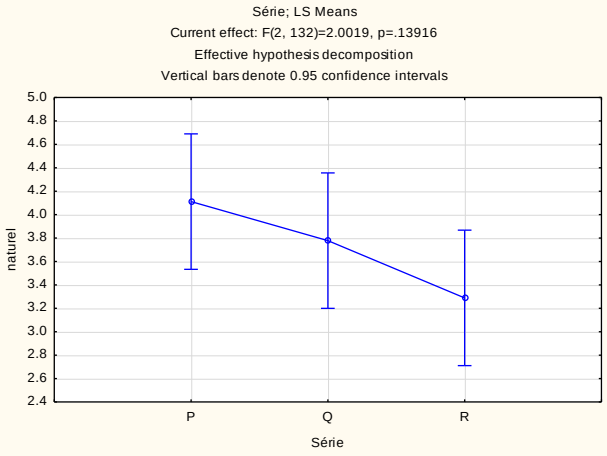


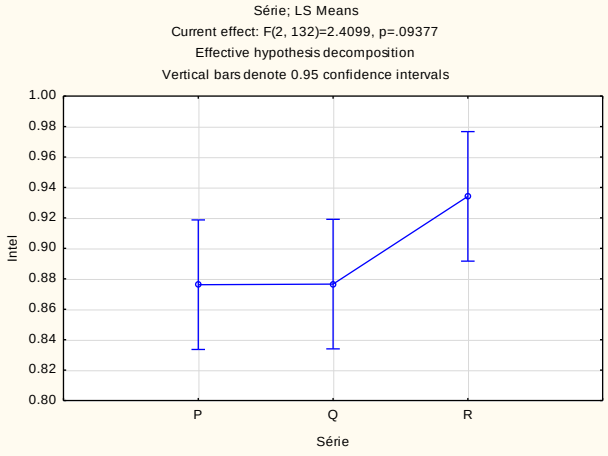
Appendice I - 6 ANOVAs série durée – cohérence interne



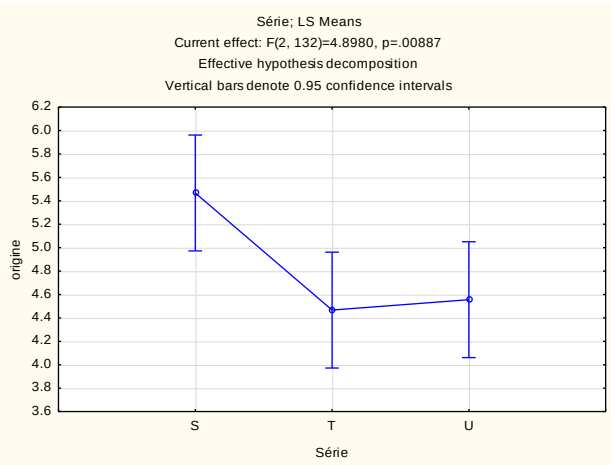


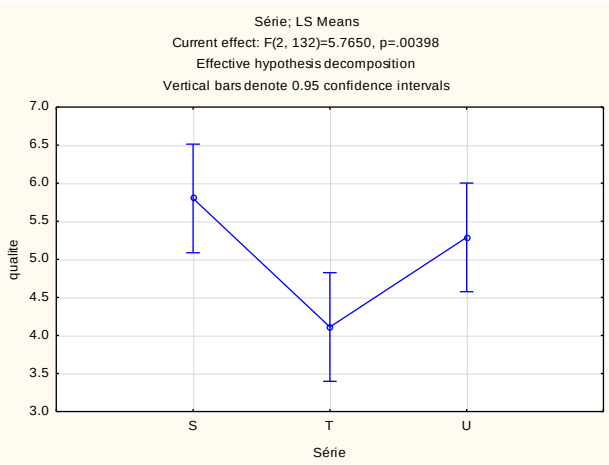


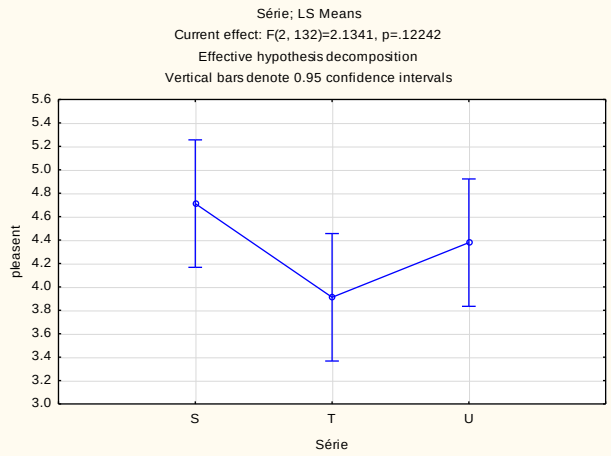


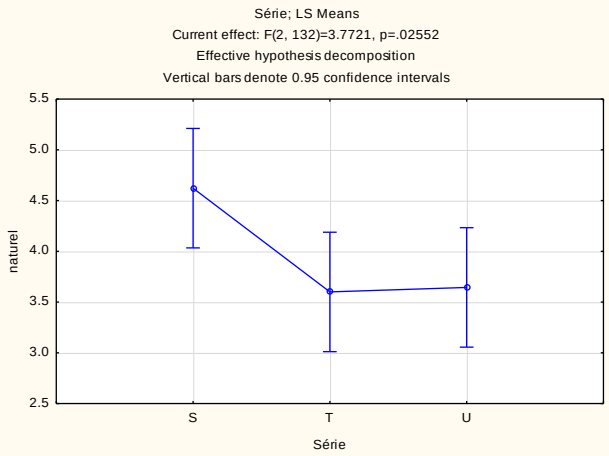


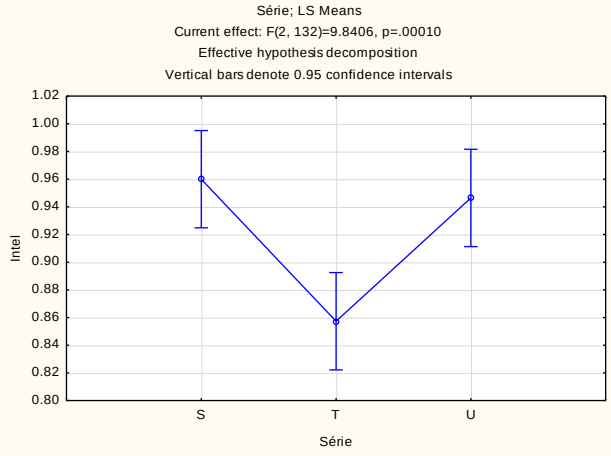
Appendice I - 7 ANOVAs série F0 – cohérence interne



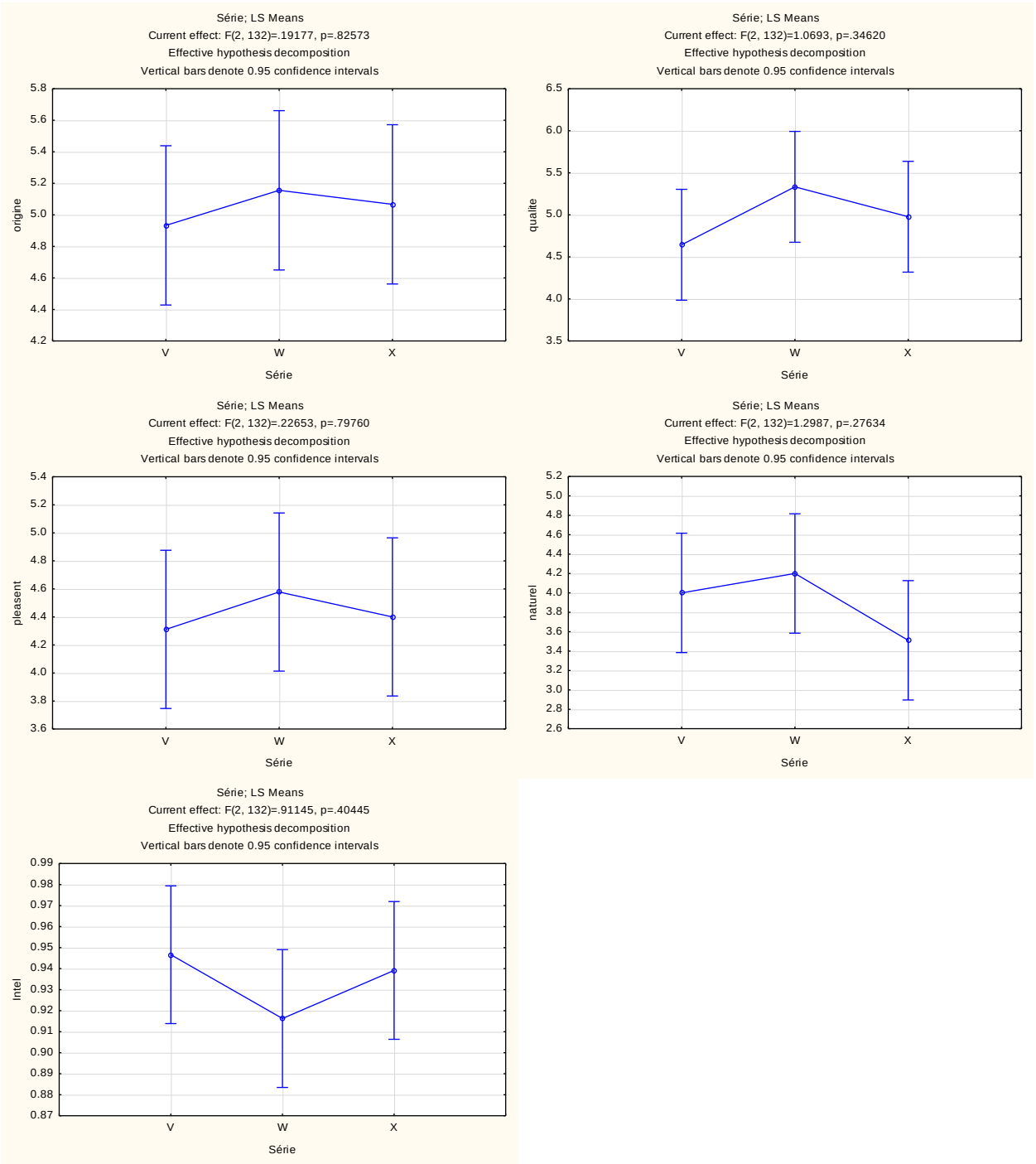




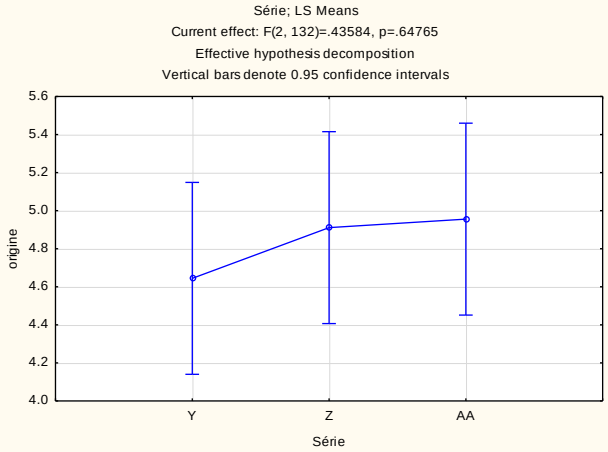


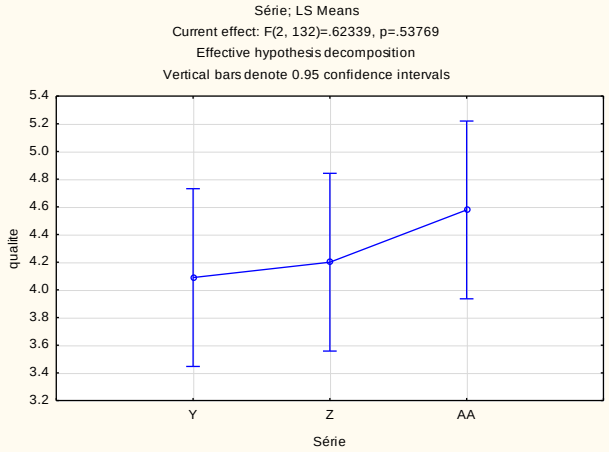


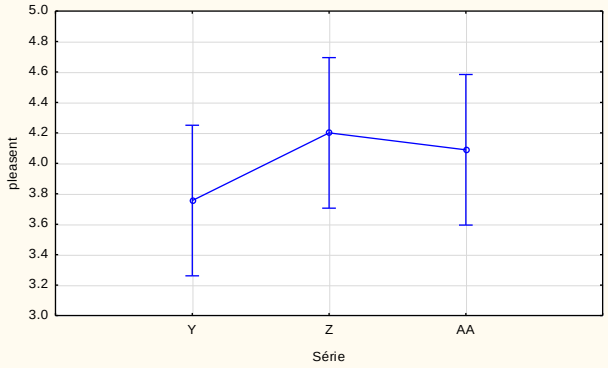
Appendice I - 8 ANOVAs série durée et F0 – cohérence interne

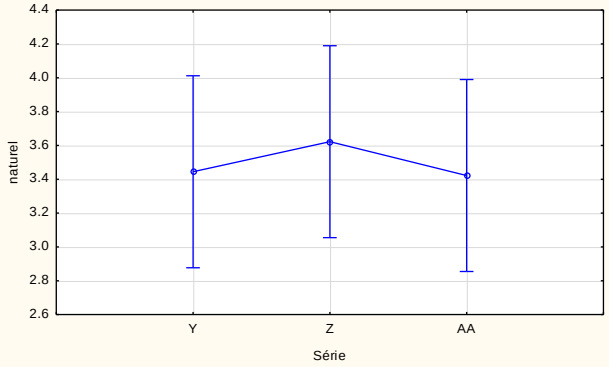


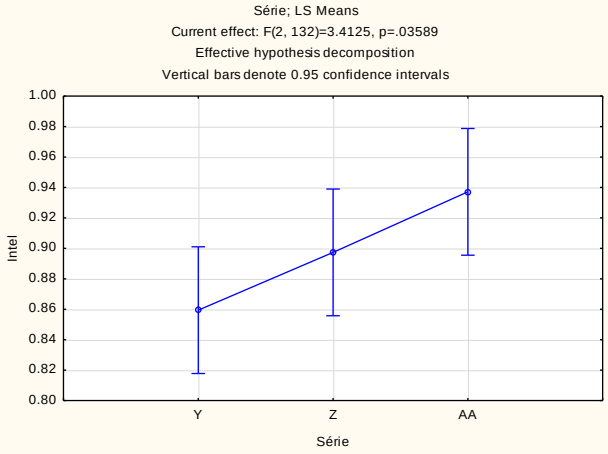
Appendice I - 9 ANOVAs série durée et F0 inverse – cohérence interne











RÉFÉRENCES

Artaud, M.-C. and Philippe Martin. 1968. “Répartition de l’énergie articulatoire en français canadien et en français standard.” Dans *Recherches sur la structure phonique du français canadien1*, sous la dir. de Pierre Léon, p. 145–160. Studia Pho. Ottawa : Didier.

Bagein, M, Thierry Dutoit, Fabrice Malfrère, Vincent Pagel, A Ruelle, N Tounsi and D Wynsberghe. 2000. “The EULER Project: an Open, Generic, Multi-lingual and Multi-Platform Text-To-Speech System.” Dans *Proceedings of ProRISC 2000*, p. 193–197.

Barra-Chicote, Roberto, Junichi Yamagishi, Simon King, Juan Manuel Montero and Javier Macias-Guarasa. 2010. “Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech.” *Speech Communication*, vol. 52, no 5, p. 394–404. En ligne. <http://linkinghub.elsevier.com/retrieve/pii/S0167639309001824>. Consulté le 5 September 2013.

Benoît, C, M Grice and V Hazan. 1996. “The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences.” *Speech Communication*. En ligne. <http://www.sciencedirect.com/science/article/pii/016763939600026X>. Consulté le 30 September 2014.

Beukelman, D and P Mirenda. 2005. “Augmentative and alternative communication: Supporting children and adults with complex communication needs.” En ligne. <http://www.citeulike.org/group/14233/article/9686963>. Consulté le 8 September 2014.

Black, A and K Tokuda. 2005. “The Blizzard Challenge 2005: Evaluating corpus-based speech synthesis on common databases.” *Proceedings of Interspeech*. En ligne. <http://festvox.org/blizzard/blizzard_cop.pdf>. Consulté le 30 September 2014.

Boersma, Paul and David Weenink. 2012. “Praat: doing phonetics by computer.” En ligne. <http://www.praat.org/>.

Boudreault, Marcel. 1967. “Rythme et mélodie: étude instrumentale comparative entre sujets québécois et français.” Dans *Étude de linguistique franco-canadienne*, sous la dir. de J.-D. Gendron and G. Straka, p. 69–88. Paris/Québec : Klincksieck/Presses de l’Université Laval.

Boula de Mareüil, Philippe, Christophe D’Alessandro, Alexander Raake, Gérard Bailly, Marie-Neige Garcia and Michel Morel. 2006. “A joint intelligibility evaluation of French text-to-speech synthesis systems: the EvaSy SUS/ACR campaign.” *LREC*. En ligne. <http://hal.archives-ouvertes.fr/hal-00103571/>. Consulté le 30 September 2014.

Brousseau, Martin. 1992. “Effets de l’utilisation de la variété québécoise sur l'intelligibilité de la parole de synthèse.” *Actes des 5èmes Journées de linguistique*. Université Laval.

Brown, P and S Levinson. 1979. “Social structure, groups and interaction ’.” En ligne. <http://pubman.mpdl.mpg.de/pubman/item/escidoc:66768:7/component/escidoc:532190/1979_Social_structure_groups.pdf>. Consulté le 2 October 2014.

Bush, Clara N. 1967. “Some Acoustic Parameters of Speech and Their Relationships to the Perception of Dialect Diferences.” *TESOL Quaterly*, vol. 1, no 3, p. 20–30.

Clopper, C. G. and a. R. Bradlow. 2008a. “Perception of Dialect Variation in Noise: Intelligibility and Classification.” *Language and Speech*, vol. 51, no 3, p. 175–198. En ligne. <http://las.sagepub.com/cgi/doi/10.1177/0023830908098539>. Consulté le 27 January 2014.

Clopper, CG and AR Bradlow. 2008b. “Perception of dialect variation in noise: Intelligibility and classification.” *Language and speech*. En ligne. <http://las.sagepub.com/content/51/3/175.short>. Consulté le 17 September 2014.

Côté-Giroux, Patricia, Natacha Trudeau, Christine Valiquette, Ann Sutton, Elsa Chan and Catherine Hébert. 2011. “Évaluation de neuf synthèses vocales françaises basée sur l’intelligibilité et l'appréciation.” *Canadian Journal of Speech-Language Pathology and Audiology*, vol. 35, no 4, p. 300–311.

Di Cristo, Albert. 1998. “Intonation in French.” Dans *Intonations Systems a Survey of Twenty Languages*, sous la dir. de D Hirst and A Di Cristo, p. 195–218. Cambridge University Press.

———. 2000. “Vers une modélisation de l’accentuation du français (seconde partie).” *Journal of French Language Studies*, vol. 10, no 1, p. 27–44.

Dutoit, Thierry, Laurent Couvreur, Fabrice Malfrère, Vincent Pagel and Christophe Ris. 2002. “Synthèse Vocale et Reconnaisance de la Parole : Droites Gauches et Mondes Parallèles.” Dans *Actes du 6è Congrès Français d’Acoustique*. Lille, France.

Dutoit, Thierry, Vincent Pagel, Nicolas Pierret, François Bataille and Olivier Van der Vreken. 1996. “The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes.” Dans *Proc. ICSLP’96*, p. 1393–1396. Philadelphia.

Fujisaki, Hiroya. 1995. “Prosody, Models, and Spontaneous Speech.” Dans *Computing Prosody : Computational Models for Processing Spontaneous*

Speech, sous la dir. de Yoshinori Sagisaka, Nick Campbell, and Norio Higuchi, p. 27–42. New York : Springer.

Gendron, J.-D. 1966. *Tendances phonétiques du français parlé au Canada*. Québec/Paris : Presses de l'Université Laval/Klincksieck.

Goldman, Jean-philippe. 2001. "De l ' analyse syntaxique à la synthèse de la parole dans le système FipsVox : phonétisation et génération de la prosodie." *Cahiers de Linguistique Française*, no 23, p. 211–234.

Holder, M. 1968. "Étude sur l'intonation comparée de la phrase énonciative en français canadien et en français standard." Dans *Recherches sur la structure phonique du français canadien*, sous la dir. de Pierre Léon, p. 175–1991. Studia Pho. Ottawa : Didier.

Huang, B. H. and S.-a. Jun. 2011. "The Effect of Age on the Acquisition of Second Language Prosody." *Language and Speech*, vol. 54, no 3, p. 387–414. En ligne. <http://las.sagepub.com/cgi/doi/10.1177/0023830911402599>. Consulté le 2 December 2011.

Jun, Sun-Ah and Cécile Fougeron. 2000. "A Phonological Model of French Intonation." Dans *Intonation : Analysis, Modeling and Technology*, sous la dir. de Antonis Botinis, p. 209–242. Dordrecht : Kluwer academic publishers.

Kalikow, DN, KN Stevens and LL Elliott. 1977. "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability." *The Journal of the Acoustical* …. En ligne. <http://scitation.aip.org/content/asa/journal/jasa/61/5/10.1121/1.381436>. Consulté le 1 October 2014.

Kaminskaïa, Svetlana. 2005. "Une étude comparée de l'intonation de la parole spontantée dans deux dialectes du francais sur deux niveaux prosodiques." University of Western Ontario.

King, Simon. 2010. “A beginners ’ guide to statistical parametric speech synthesis.” *DRAFT*, p. 1–16.

King, Simon and Vasilis Karaiskos. 2013. “The Blizzard Challenge 2013.” Dans *The Blizzard Challenge 2013 Workshop, SSW 7*. Barcelona.

Klatt, DH. 1980. “Software for a cascade/parallel formant synthesizer.” *the Journal of the Acoustical Society of America*. En ligne. <http://scitation.aip.org/content/asa/journal/jasa/67/3/10.1121/1.383940>. Consulté le 1 December 2014.

Kreul, EJ, JC Nixon, KD Kryter and DW Bell. 1968. “A proposed clinical test of speech discrimination.” *Journal of Speech, …*. En ligne. <http://jslhr.pubs.asha.org/article.aspx?articleid=1754604>. Consulté le 1 October 2014.

Lacheret-Dujour, Anne and Frédéric Beaugendre. 1999. *La prosodie du français*. Coll. « CNRS langage ». CNRS.

Lai, Jennifer, David Wood and Michael Considine. 2000. “The effect of task conditions on the comprehensibility of synthetic speech.” Dans *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI ’00*, vol. 2, p. 321–328. The Hague, Netherlands : ACM Press. En ligne. <http://portal.acm.org/citation.cfm?doid=332040.332451>.

Lanchantin, P. 2010. “A HMM-based speech synthesis system using a new glottal source and vocal-tract separation method.” *Acoustics Speech and …*. En ligne. <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5495550>. Consulté le 8 September 2014.

Landercy, Albert and Raymond Renard. 1977. *Éléments de phonétique*. Bruxelles : Didier.

Laures, Jacqueline S and Kate Bunton. 2003. “Perceptual effects of a flattened fundamental frequency at the sentence level under different listening conditions.” *Journal of Communication Disorders*, vol. 36, no 6, p. 449–464. En ligne. <http://www.sciencedirect.com/science/article/pii/S0021992403000327>. Consulté le 3 October 2014.

Laures, Jacqueline S. and Gary Weismer. 1999. “The Effects of a Flattened Fundamental Frequency on Intelligibility at the Sentence Level.” *Journal of Speech Language and Hearing Research*, vol. 42, no 5, p. 1148. En ligne. <http://jslhr.pubs.asha.org.proxy.bibliotheques.uqam.ca:2048/article.aspx?articleid=1781167>. Consulté le 30 September 2014.

Martin, Pierre. 1996. *Éléments de phonétique avec application au français*. Québec : Les Presses de l’Université Laval. En ligne. <http://www.phonetique.uqam.ca/upload/files/ORA1531/phontique combinatoire.pdf>. Consulté le 11 September 2014.

Mayo, Catherine, Robert a.J. Clark and Simon King. 2011. “Listeners’ weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis.” *Speech Communication*, vol. 53, no 3, p. 311–326. En ligne. <http://linkinghub.elsevier.com/retrieve/pii/S0167639310001627>. Consulté le 28 November 2012.

Ménard, Lucie. 1998. “Perception et reconnaissance des accents québécois et français : identification de marqueurs prosodiques.” Université Laval.

Miller, SE, RS Schlauch and PJ Watson. 2010. “The effects of fundamental frequency contour manipulations on speech intelligibility in background noisea).” *The Journal of the Acoustical …*. En ligne. <http://scitation.aip.org/content/asa/journal/jasa/128/1/10.1121/1.3397384>. Consulté le 3 October 2014.

Monaghan, Alex. 2002. “Prosody in Synthetic Speech Problems, Solutions and Challenges.” Dans *Improvements in Speech Synthesis*, sous la dir. de Eric Keller, vol. 4, p. 89–92. John Wiley & Sons.

Nusbaum, Howard C., Alexander L. Francis and Anne S. Henly. 1995. “Measuring the naturalness of synthetic speech.” *International Journal of Speech Technology*, vol. 1, no 1, p. 7–19. En ligne. <http://www.springerlink.com/index/10.1007/BF02277176>.

O’Shaughnessy, Douglas. 2007. “Modern Methods of Speech Synthesis.” *Ieee Circuits And Systems Magazine*, vol. 3, p. 6–23.

Obin, Nicolas, Pierre Lanchantin, Mathieu Avanzi, Anne Lacheret-Dujour and Xavier Rodet. 2010. “Toward Improved HMM-Based Speech Synthesis Using High-Level Syntactical Features.” Dans *Speech Prosody 2010*, p. 100133. Chicago.

Ohala, John J and J.B Gilbert. 1981. “Listeners’ ability to identify languages by their prosody.” Dans *Problèmes de Prosodie, vol. 2 : Expérimentations, modèles et fonctions*, sous la dir. de Pierre Léon and M Rossi, p. 123–131. Ottawa : Didier.

Ouellet, Marise. 1992. “Systématique des durées segmentales dans les syllabes en français de France et du Québec.” Université de Montréal.

Paradis, Claude, Martin Brousseau and Jean Dolbec. 1993. “Variétés linguistiques et intelligibilité : enjeux sociolinguistiques pour la synthèse de parole.” *Revue québécoise de linguistique*, vol. 22, no 2, p. 13–36.

Paris, Carol R., Margaret H. Thomas, Richard D. Gilson and J. Peter Kincaid. 2000. “Linguistic Cues and Memory for Synthetic and Natural Speech.” *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 42, no 3, p. 421–431. En ligne. <http://hfs.sagepub.com/cgi/doi/10.1518/001872000779698132>. Consulté le 16 November 2012.

Pasdeloup, Valérie. 1990. “Multi-Style Prosodic Model for Frnch Text-To-Speech Synthesis.” Dans *The ESCA Workshop on Speech Synthesis*, p. 193–196. Autrans, France.

Peters, J, P Gilles, P Auer and M Selting. 2002. “Identification of Regional Varieties by Intonational Cues An Experimental Study on Hamburg and Berlin German.” *Language and speech*. En ligne. <http://las.sagepub.com/content/45/2/115.short>. Consulté le 17 September 2014.

Petrone, Caterina. 2010. “At the interface between phonetics and pragmatics : Non-local f 0 effects on the perception of Cosenza Italian tunes.” Dans *Speech Prosody 2010*, p. 100966. Chicago.

Ratcliff, Ann, Sue Coughlin and Mark Lehman. 2002. “Factors Influencing Ratings of Speech Naturalness in Augmentative and Alternative Communication.” *Augmentative and alternative communication*, vol. 18, no March, p. 11–19.

Robert, P, A Rey and J Rey-Debove. 1993. *Le nouveau petit Robert*. En ligne. <http://doxa.u-pec.fr/help/visiteguideeRobert2009.2.pdf>. Consulté le 1 December 2014.

Robinson, L. 1968. “Étude du rythme syllabique en français canadien et en français standard.” Dans *Recherches sur la structure phonique du français canadien*, sous la dir. de Pierre Léon, p. 161–174. Studia Pho. Ottawa : Didier.

Van Santen, Jan P.H., Richard W. Sproat, Joseph P. Olive and Julia Hirschberg. 1997. *Progress in Speech Synthesis*. Springer.

Spitzer, SM, JM Liss and SL Mattys. 2007. “Acoustic cues to lexical segmentation: A study of resynthesized speech.” *... Journal of the Acoustical Society of ...*. En ligne. <http://scitation.aip.org/content/asa/journal/jasa/122/6/10.1121/1.2801545>. Consulté le 3 October 2014.

Szmidt, Y. 1968. “Étude de la phrase interrogative en français canadien et en français standard.” Dans *Recherches sur la structure phonique du français canadien*, sous la dir. de Pierre Léon, p. 175–191. Studia Pho. Ottawa : Didier.

Tamura, Masatsune, Norbert Braunschweiler, Takehiko Kagoshima and Masami Akamine. 2010. “Unit Selection Speech Synthesis Using Multiple Speech Units at Non-adjacent Segments for Prosody and Waveform Generation.” Dans *ICASSP 2010*, p. 4802–4805.

Taylor, Paul. 2009. *Text-to-Speech Synthesis*. Cambridge : Cambridge University Press.

Thibault, Linda. 1999. “Variations phonétiques et tonales en français québécois lu et spontané.” Université du Québec à Montréal.

Trofimovich, Pavel and Wendy Baker. 2006. “Learning Second Language Suprasegmentals : Effect of L2 Experience on Prosody and Fluency Characteristics of L2 Speech.” *Studies in Second Language Acquisition*, vol. 28, p. 1–30.

Vinay, J. P. 1955. “Aperçu des études de phonétique canadienne.” Dans *Études sur le parler français au Canada*, p. 61–82. Québec : Presses universitaires Laval.

Viswanathan, Mahesh and Madhubalan Viswanathan. 2005. “Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale.” *Computer Speech & Language*, vol. 19, no 1, p. 55–83. En ligne. <http://www.sciencedirect.com/science/article/pii/S0885230803000676>. Consulté le 30 September 2014.

Yergeau, Eric, Martine Poirier, Marc Couture and Yves Poulin. 2013. “Régression logistique.” *Université de Sherbrooke*. En ligne. <http://spss.espaceweb.usherbrooke.ca/pages/stat-inferentielles/regression-logistique.php>.

Zen, H, K Tokuda and A Black. 2009. “Statistical Parametric Speech Synthesis.” *Speech Communication*, vol. 51, no 11, p. 1039–1064.