DOCTORAT AIX-MARSEILLE UNIVERSITE
Délivré par l'UNIVERSITE DE PROVENCE

N° attribué à la bibliothèque

Analyse et Interprétation de l'Empan Temporel des Variations Prosodiques en Français et en Anglais

THÈSE

présentée et soutenue publiquement le 29 janvier 2010

pour l'obtention du grade de

Docteur d'Aix-Marseille Université

(spécialité Sciences du Langage – Phonétique anglaise)

par

Céline De Looze

Composition du jury

Président : Mariapaola D'Imperio, Professeur, Laboratoire Parole et Langage, UMR 6057 CNRS - Université de Provence

Rapporteurs : Marc Swerts, Professeur, Tilburg University - Pays Bas
Jacqueline Vaissière, Professeur, Laboratoire de Phonétique et Phonologie, UMR 7018 - Université Paris 3

Examinateur : Nick Campbell, Professeur, Trinity College, Dublin - Irlande

Directeur : Daniel Hirst, Directeur de recherche, Laboratoire Parole et Langage, UMR 6057 CNRS - Université de Provence

---

Formation Doctorale Cognition, Langage, Education – ED356

Mis en page avec la classe thlpl.

A mon lit douillet, ma famille.

.

Remerciements

Je tiens à remercier en premier lieu Daniel Hirst qui a dirigé ce travail. Tout au long de ces quatre années, il a su être disponible pour répondre à mes multiples questions et j'ai particulièrement apprécié son aptitude à expliquer et ré-expliquer les choses, toujours griffonnées au coin d'une feuille, sans jamais donner l'impression de la mauvaise question. Vrai père scientifique, il a été rassurant et enrichissant de grandir auprès de lui. Sa gentillesse, son humour et son flegme légendaire ont fait que cette thèse n'aurait pu être plus agréable et qu'il m'a été difficile d'y mettre fin. Pour tout cela, je lui suis très reconnaissante.

Je remercie aussi sincèrement Nick Campbell, MariaPaola d'Imperio, Marc Swertsz et Jacqueline Vaissière d'avoir accepté d'évaluer ce travail.

Mes remerciements vont tout particulièrement à mon italienne (allemande ?...) de soeur, Caterina Petrone, pour avoir pris le temps de relire ce travail, pour ses nombreux conseils et son affection. Ils vont aussi à BraBrachef (Stéphane Rauzy) pour m'avoir montré que la programmation ne tourne pas qu'autour de Praat... Je le remercie pour les heures qu'il a passé à rendre le complexe si accessible, pour sa disponibilité et son amitié. Ils reviennent enfin à Robert Espesser sans qui R ne serait qu'une foule de questions ou de réponses... Je le remercie pour son aide, bien qu'il ne veuille jamais le reconnaître, inestimable.

Une thèse ne peut se faire sans ceux qui baignent dans la même galère et qui comprennent si bien. Je remercie par là le doctorants-nucleus du Laboratoire Parole et Langage. Je pense aussi bien à ceux qui s'y trouvent encore que ceux qui ont aujourd'hui sorti la tête de l'eau. Un merci bien aimant à mes plus proches, Cécilou Petitjean, mon fwèw Manu Makasso, re-Kathia, Amandinette Michelas, Annchou Tortel, Mes boucles d'argent Aubanel, Prince JeanPhiPhi Prost, Paulinette Péri, Céline Dou, Siiiiiil Cho, pour leur amitié, leur affection et leurs encouragements. Un grand merci à la CIES-team (re-Cécilou, re-Céline, Greg, Cola, Renaud et Xav) et au bon vieux temps.

Je remercie aussi la plancha du roi Philippe et de ses acolytes qui fait du Laboratoire Parole et Langage un cadre privilégié et très accueillant pour y effectuer une thèse. Mes remerciements vont directement à ceux qui sont toujours là... Cyril Deniaud, Sébastien Bermond, Alain Ghio, Thierry Legou, Isabelle Vincent et Joëlle Lavaud.

Mère, belle-mère et tante ont relu attentivement ce manuscrit. Je les remercie tendrement pour le temps qu'elles y ont consacré et leur patience à corriger mes anglicismes. Un merci tout particulier à ma mère pour ses commentaires avisés sur les formules utilisées dans ce manuscrit (...), pour m'avoir veillée lors de la toute dernière nuit, que tout thésard sait, plus que difficile.

Je remercie aussi mon père et son aptitude à rendre les choses toujours si faciles... la 405 Peugeot de mon frère de s'être tapé 850 bornes juste pour être là... la porte de mes grands-parents toujours ouverte... le mac de Nono sans qui Momel n'aurait pu tourner aussi bien... le soutien de toute une famille... la patience de mes amis face à mes absences...

Je remercie tout particulièrement Michel P. auprès de qui j'ai aimé grandir et sans qui cette thèse n'aurait pas eu le goût de la tranquillité.

Et un merci inestimable, mis en exergue, en fin de texte, à Jérôme qui, avec beaucoup de patience, a subi mes absences et rendu la rédaction de cette thèse légère par ses encouragements et son amour.

Et pour tous ceux qui comprendront : hé ! Wantchogalantcho... pinco !

Table des matières

Introduction

1 Prosodie et variations prosodiques

1.1 Problématique de l'empan temporel

Depuis quelques années maintenant, de nombreuses études menées en prosodie se sont attachées à décrire les variations prosodiques de la parole, d'un point de vue acoustique, perceptif, formel, fonctionnel, ou cognitif, sous le couvert de divers cadres théoriques qu'impliquent ces différents niveaux d'analyse, ou encore, couplées avec d'autres champs d'investigation, par la relation que la prosodie entretient avec les autres composantes du langage, que sont la sémantique, la syntaxe ou encore la pragmatique. Pourtant, malgré la disparité et l'aspect pluridisciplinaire des travaux menés en prosodie, une problématique sous-jacente et commune à ces multiples disciplines est celle de l'empan temporel des variations prosodiques de la parole. En effet, que les auteurs s'intéressent à l'analyse formelle de la prosodie ou à celle de sa fonctionnalité, une difficulté récurrente rencontrée est celle de la délimitation du domaine des variations prosodiques. Leur empan temporel, à plus ou moins long terme, rend leur analyse séparée difficile, de par le chevauchement qui en résulte.

Nous utilisons le terme d'*empan temporel* comme synonyme de *domaine* et le définissons comme un intervalle de temps délimité par des bornes. Comme White (2002), nous distinguons les termes d'empan temporel (ou de domaine) du terme de *locus*. Le domaine est l'unité à partir de laquelle opère un phénomène prosodique, le locus est la localisation exacte des effets de ce phénomène au sein du domaine. Par exemple, dans le cas d'une remise à niveau de registre, le locus du changement se situe au niveau de la syllabe mais il délimite aussi la frontière de

domaines plus larges, tels que l'énoncé. Nous reviendrons sur ce point dans le premier chapitre de cette thèse.

L'étude d'un phénomène prosodique spécifique, comme nous le verrons au cours de cette introduction et du premier chapitre, se fait donc pour beaucoup sans la considération des variations prosodiques interférentes, rendant de ce fait l'analyse du phénomène prosodique caduque. Les variations prosodiques émergent en effet d'un ensemble de sous-systèmes, « partiellement autonomes » et « interactifs » (Di Cristo, 2005) qui exige leur distinction et la prise en compte de leur interaction.

Di Cristo (2005) reconnaît dans cet ensemble de sous-sytèmes trois ordres structurants, l'ordre de structuration tonale, l'ordre de structuration métrique et l'ordre de structuration temporelle. Ces trois ordres de structuration sont eux-mêmes subdivisés, selon qu'ils s'appliquent au niveau lexical ou au niveau post-lexical ou supra-lexical (organisation d'unités de rang supérieur au lexique ; Hirst et Di Cristo (1998)). Ainsi, au niveau lexical, les ordres de structuration tonale, de structuration métrique et de structuration temporelle assument, respectivement, l'organisation (1) des contrastes de tons, (2) d'accentuation lexicale et (3) des oppositions de quantité. Au niveau post-lexical (intonatif et discursif), l'ordre de structuration tonale, de structuration métrique et de structuration temporelle gouvernent, respectivement, (1) l'intonation, (2) l'accentuation post-lexicale et le rythme, ainsi que (3) les pauses structurales, les effets d'allongements et les variations de tempo (Di Cristo, 2005).

Les variations des tons lexicaux et d'accent, de quantité dont la dimension est qualifiée de syntagmatique (Di Cristo et al., 2004) sont ainsi envisagées comme des variations dites locales, à court terme, et l'unité pour laquelle elles sont définies ne dépasse généralement pas le mot (niveau morpho-lexical). Les variations de registre, de tempo [1] et d'amplitude [2], elles, sont dites à plus long terme et occupent des domaines de niveau supérieur (discursif), tels que le syntagme, la phrase, ou l'énoncé. Elles sont qualifiées d'« orthogonales », vecteurs superposant à la dimension syntagmatique (Di Cristo et al., 2004). Il est important de mentionner que cette liste n'est pas exhaustive et qu'il n'existe pas de franc consensus dans la littérature quant à la détermination de l'empan temporel des variations prosodiques à court et à long terme. Si les auteurs s'entendent sur le fait que les variations prosodiques se font localement ou globalement, il n'existe pas de stricte frontière entre ce qui est référé au terme d'événement local et ce qui est référé au terme d'événement global. Collier et Cohen (1990) expliquent, par exemple, que les variations locales opèrent sur quelques syllabes seulement alors que les variations globales s'étendent sur des extraits de parole plus longs tels que la proposition ou l'énoncé. Par ailleurs,

1. Une définition précise de ce que nous entendons pas « registre » et « tempo » sera donnée dans le premier chapitre.

2. Notons que nous introduisons dans le concept de prosodie défini par Di Cristo (2005) les notions de registre et d'amplitude.

un même patron prosodique peut être à plus ou moins long terme, selon l'intention du locuteur. L'élévation de la voix par exemple peut s'étendre sur un ensemble de phrases comme se restreindre au domaine du mot. Laver et Trudgill (1979) et Laver et Hanson (1981) proposent de qualifier ces caractéristiques de moyen terme de par la variation de leur empan temporel. Les caractéristiques identificatrices de l'individu sont elles qualifiées de long terme, permanentes ou quasi-permanentes, de par leur nature organique, fonction de l'appareil vocal du locuteur. Pour sa part, Ladd (1996) préfère le terme d'indice paralinguistique, plutôt que de parler de variations à moyen ou à long terme, dès lors que les variations de ces indices viennent se greffer aux indices linguistiques modifiant leur réalisation acoustique sans pour autant altérer leur identité. En effet, selon l'auteur, la différence entre les éléments linguistiques et paralinguistiques ne réside pas dans le domaine de variations. Ils ne doivent pas être différenciés par l'utilisation de « à plus ou moins long terme » car ils pouvent être liés à des parties individuelles du message, localisées ou au contraire, groupées, globalisées. Il est cependant de fait que la problématique d'empan temporel n'est toujours pas résolue à ce jour et pose de nombreuses difficultés dans l'analyse automatique, l'exploitation et la comparaison des études menées en prosodie.

1.2 De la nécessité d'une distinction de l'empan temporel des variations prosodiques

La notion d'empan temporel, notamment l'empan temporel des variations prosodiques à long terme [3], à notre connaissance, n'a pas fait l'objet d'études approfondies en prosodie. Dans de nombreux travaux, les spécialistes s'intéressent davantage à la description des phénomènes intonatifs et rythmiques des langues naturelles et considèrent généralement les variables prosodiques à long terme comme invariantes. Ils décrivent, pour beaucoup, les patrons intonatifs et rythmiques des langues naturelles sans considérer les variations de registre, de tempo ou d'intensité.

1.2.1 Chevauchement et interaction

Si ces études et les outils développés pour l'analyse automatique et la modélisation de l'intonation et du rythme de la parole sont satisfaisants pour l'étude de la parole lue, comme c'est souvent le cas pour la parole de laboratoire, il n'en est pas de même pour l'étude de la parole spontanée. En effet, il sera très tôt reconnu que, en parole spontanée, l'interaction et le chevauchement des empans à court terme et à long terme rendent leur séparation difficile.

3. Nous référons au terme de « variations prosodiques à long terme » les variations de registre, de tempo et d'intensité.

Bolinger (1951) expliquera notamment que l'objection, qui pourrait être portée à la plupart des systèmes scalaires dédiés à l'annotation des patrons intonatifs, repose sur leur incapacité à distinguer les changements locaux (correspondant à une distinction phonologique) des changements plus globaux, ie. de registre (dus à des facteurs extralinguistiques ou au discours). Dans un système comme celui de Trager et Smith (1957), une petite chute (/31/ ou /21/) dans un registre étendu ne peut être distinguée d'une grande chute (/41/) dans un registre étroit. Un tel système ne peut également distinguer deux chutes isolées de type /43/ et /21/, ou /42/ et /31/, à moins de spécifier la hauteur du registre du locuteur. De même, Hirst (2006) souligne la difficulté, dès lors que l'on s'intéresse à la structure rythmique d'un énoncé, à distinguer un phonème court dans un tempo ralenti d'un phonème long dans un tempo accéléré, une distinction, qui, non établie, peut rendre un énoncé polysémique. Comment distinguer, en effet, l'énoncé [ilabatyləʃjɛ̃] (il a battu le chien) dans un tempo ralenti, de l'énoncé [ilaabatyləʃjɛ̃] (il a abattu le chien) dans un tempo accéléré ? Les variations à long terme doivent donc être spécifiées pour une meilleure analyse et interprétation des variations à plus court terme.

1.2.2 Fonctionnalité

Il est d'autant plus nécessaire de distinguer les variations prosodiques à long terme des variations prosodiques à court terme qu'elles assument des fonctions différentes. En effet, les variations de tons, de quantité, d'accent lexical peuvent participer, dans certaines langues, à l'identité lexicale d'un mot. En vietnamien, par exemple, /ma/ prononcé avec un ton haut signifie « fantôme », avec une montée « mère », avec une montée glottalisée « cheval », avec un ton descendant « maïs », avec un ton descendant et montant « tombeau » et avec une descente glottalisée « jeune plant de riz ». En finnois, c'est par la durée des phonèmes que l'on distingue les mots « derrière », « cheminée », « fardeau » prononcés respectivement [taka], [takka] et [taaka] (Hirst, 2006). En anglais, la position de l'accent lexical participe à la distinction des mots /'record/ (disque) et /re'cord/ (enregistrer).

Les variations prosodiques à plus long terme revêtent de fonctions multiples, participant à l'hétérogénéité fonctionnelle de la prosodie. De par cette multiplicité, il en va d'un foisonnement de termes référant aux fonctions prosodiques et d'une difficulté de définir exactement les diverses significations véhiculées par les contrastes prosodiques. Parce que l'interprétation de la forme prosodique est dépendante du contexte, des significations différentes résultent pour un même patron prosodique lorsque les contextes divergent ou, inversement, une même signification malgré des patrons prosodiques différents (Di Cristo, 2004). Pour notre part, malgré les difficultés soulignées, nous retiendrons, pour faciliter la lecture de ce contexte scientifique, les fonctions dites linguistiques, extra-linguistiques et paralinguistiques, bien que, comme Trouvain (2004),

nous ne pensons pas qu'il existe de franche limite entre ces trois catégories [4].

Fonctions linguistiques

Les variations de registre, de tempo et de force de voix assument des fonctions linguistiques, elles informent d'une part de l'organisation informationnelle du discours, d'autre part de sa dimension hiérarchique et de l'organisation relationnelle des unités linguistiques. Nous entendons par organisation informationnelle du discours la façon dont l'information est mise en perspective, ie. la façon dont un élément est mis en avant, focalisé, par rapport au reste de la chaîne linguistique (Lacheret-Dujour, 1999). La fonction informationnelle ou emphatique des variations prosodiques est en effet celle de la construction d'une échelle focale des énoncés permettant la signalisation et la distinction d'un nouveau sujet vs. d'un sujet connu, d'un topique vs. d'un commentaire, d'un thème vs. d'un rhème, d'aspects de focalisation et de proéminence, etc. (Bolinger, 1972; Laver, 1991; Ladd, 1996; Rossi, 1999; Xu, 1999; Di Cristo, 2000; Mertens, Auchlin, Goldman, & Grobet, 2001; Fon, 2002; Verhoeven, 2002; Di Cristo, 2004; Trouvain, 2004; Gussenhoven, 2005; Tseng, Chang, & Su, 2005; Den Ouden, Noordman, & Terken, 2008). Nous entendons par dimension hiérarchique et organisation relationnelle la façon dont la structure des énoncés et du discours est signalée, permettant ainsi leur interprétation sémantique et pragmatique. Cette fonction structurale ou grammaticale et délimitative des variations prosodiques permet la hiérarchisation des différents éléments du discours, de leurs liens, ainsi que la démarcation de ces différents éléments en unités linguistiques par la signalisation de frontières (Ayers, 1994; Hirschberg & Grosz, 1992b; Swerts, 1997; Mertens et al., 2001; Patterson, 2000; Vaissière, 2005; Di Cristo, 2004).

Dimension hiérarchique et organisation relationnelle

Ainsi, il apparaît, dans de nombreuses langues, que les variations de registre, de tempo (vitesse d'articulation ou débit d'élocution et pauses) et d'intensité participent à la structuration du discours. En lecture de texte, les phrases situées en début de paragraphe ont généralement un registre plus étendu et plus élevé (Lehiste, 1975; Brazil, 1980; G. Brown, Currie, & Kenworthy, 1980; Bruce, 1982; Menn & Boyce, 1982; G. Brown, 1983; Thorsen, 1985; Sluijter & Terken, 1992; Ayers, 1994; Nicolas & Hirst, 1995; Swerts, 1997; Clark, 1999; Ouden, Noordman, & Terken, 2009), une intensité plus élevée (G. Brown, 1983), une vitesse d'articulation plus ou moins lente selon les études (Brubaker, 1972; Butterworth, 1975; Miller, Grosjean, & Lomanto, 1984; Ayers, 1994; Beinum & Donzel, 1996; Fougeron & Jun, 1998; Hirose & Kawanami, 1998; Dankovicova, 1999; Ouden et al., 2009) et sont précédées de pauses plus longues (Goldman-Eisler, 1961; Grosjean & Deschamps, 1972a; Grosjean & Collins, 1979; Lehiste, n.d.; Duez, 1982; G. Brown, 1983; Silverman, 1987; Ayers, 1994; Swerts, 1997; Fon, 2002) que les phrases

4. Notons que nous n'aborderons pas dans cette thèse, par soucis de cadre, la fonction d'assistance à l'encodage et au décodage de la parole des variations prosodiques (Di Cristo, 2000; Dubois, 1994).

en milieu de paragraphe ou en fin de paragraphe. Dans son étude, menée à partir du corpus *Boston radio News Corpus*, Clark (1999) rapportera que la position des groupes tonals au sein de prédicats, en anglais, est signalée par des variations de hauteur de registre. La hauteur moyenne des groupes tonals initiaux est en effet de 200 Hz alors que celle des groupes médians est de 160-170 Hz. Les groupes médians ont un statut similaire, la hauteur du registre pour chacun de ces groupes étant sensiblement équivalente. Enfin, les groupes médians se distinguent aussi des groupes finaux par leur hauteur. Les travaux de Nicolas et Hirst (1995) confirmeront le statut « distinctif » des groupes initiaux pour le français ; les unités intonatives en début de paragraphe ont un registre plus étendu que les unités médianes et finales. Il est à noter que, pour ces deux études, la différence de hauteur et d'étendue du registre entre les groupes initiaux et les groupes médians est plus importante que celle entre les groupes médians et finaux. La déclinaison (« downtrend »), ou supra-déclinaison (« supra-declination »), selon l'empan temporel considéré dans les études, signale ainsi la segmentation du discours (Bruce, 1982; Vaissière, 1983; Thorsen, 1985; Ladd, 1988; Sluijter & Terken, 1992, 1993; Ayers, 1994; J. Pijper & Sanderman, 1994; Swerts, Bouwhuis, & Collier, 1994; Nicolas & Hirst, 1995) [5].

Le degré d'expansion et les différences de hauteur du registre, à des frontières de segments du discours, s'avèrent donc localiser, en anglais et en français, et dans d'autres langues d'ailleurs, la place des frontières, et refléter la relation des segments dans la structure du discours (Ayers, 1994; Fon, 2002). Plus exactement, en lecture de texte ou en parole spontanée, les variations de registre communiqueraient la structure du topique dans le discours. Une remise à niveau mélodique (« pitch reset ») (Brazil, 1980; G. Brown et al., 1980; Yule, 1980; Menn & Boyce, 1982; Terken & Collier, 1989; Nakajima & Allen, 1992, 1993; Swerts & Geluykens, 1993; J. Pijper & Sanderman, 1994; Oliveira, n.d.; Kong, 2004; Wang & Xu, 2006; Ouden et al., 2009) ou un élargissement de l'étendue du registre (Hirschberg & Pierrehumbert, 1986; Grosz & Hirschberg, 1992) ou encore la combinaison de ces deux indices acoustiques (Bruce, 1995) permettraient ainsi au locuteur d'introduire un nouveau topique [6]. Un changement de topique est donc indiqué par une discontinuité mélodique (Menn & Boyce, 1982; Grosz & Hirschberg, 1992). En revanche, les unités traitant d'un même topique sont marquées par un registre identique, dont l'étendue diminue et la hauteur s'abaisse progressivement au cours du temps (Brazil, 1980; G. Brown et al., 1980; Yule, 1980; Bruce, 1982; Hirschberg & Pierrehumbert, 1986, 1986; Silverman, 1987; Nakajima & Allen, 1993; Swerts & Geluykens, 1993). Sur un empan plus local, les auteurs rapportent enfin qu'un abaissement final (*final lowering*) indiquerait la clotûre de topiques majeurs (Hirschberg & Pierrehumbert, 1986; Nakajima & Allen, 1992). Plus le registre est bas et étroit, plus le degré de finalité d'un énoncé est élevé

5. Bien que nous focalisons notre intérêt sur la fonctionnalité de la déclinaison, nous ne remettons pas pour autant qu'elle résulte de l'influence de facteurs physiologiques, e.g. variations de la pression sous-glottique, ajustements laryngés, etc. (Honda, 2004).

6. Nous adoptons ici la définition de Lambrecht (1996, p118), pour qui le « topique » équivaut à « ce dont on parle ».

(Hirschberg & Pierrehumbert, 1986).

Les variations du débit d'élocution, du nombre et de la durée des pauses participent également à la signalisation de la structure du discours et révèlent des changements ou des continuations de topicalisation. Les segments initiaux du discours, introduisant un nouveau topique, sont généralement précédés de pauses plus longues que les segments médians et finaux, notamment en anglais et en français (Lehiste, n.d.; G. Brown et al., 1980; Silverman, 1987; Swerts & Geluykens, 1993; J. Pijper & Sanderman, 1994; Fon, 2002; C. Smith, 2005; Ouden et al., 2009). Aucun consensus n'a encore été établi quant aux variations du débit d'élocution. Brubaker (1972), et Butterworth (1975) rapportent qu'en anglais, les segments initiaux, après une frontière de paragraphe, sont articulés plus lentement que ceux positionnés au sein ou en fin de paragraphe, pour lesquels le débit d'élocution s'accélère. En revanche, dans les études de Dankovicova (1999), Ayers (1994) et Beinum et Donzel (1996), les unités finales du discours sont plus lentes que les unités initiales. Dankovicova (1999) conclut en effet de son étude que la vitesse d'articulation, en anglais et en tchèque, ne varie pas arbitrairement mais de façon structurée : un *rallentando* s'exercerait en effet sur l'unité intonative. C. Smith (2005), quant à elle, explique qu'en anglais et en français, la signalisation de la topicalisation se fait par un ralentissement du débit d'élocution avant et après un changement de topique, une tendance également rapportée en japonais (Hirose & Kawanami, 1998)). Fougeron et Jun (1998) et Miller et al. (1984) proposent plutôt que la vitesse d'articulation, en français et en anglais, n'augmenterait pas ou ne diminuerait pas graduellement, mais changerait plusieurs fois au cours de l'énoncé.

Les variations de registre, de tempo et d'intensité permettent aussi de signaler des clauses parenthétiques (Kutik, Cooper, & Boyce, 1983; Fagyal, 2002) et des citations ; en anglais, les clauses parenthétiques sont révélées par un registre bas et étroit, les citations, par un registre haut et élargi (Grosz & Hirschberg, 1992; Hirschberg & Grosz, 1992a; Arons, 1994). Elles peuvent aussi constituer les indices de possibles tours de parole (*turn-taking*) (Ayers, 1994; Batliner, Kießling, Kompe, Niemann, & Nöth, 1997) et sont utilisées pour désambiguïser des structures syntaxiques complexes (Silverman, 1987; Batliner et al., 1997; Mayer, Jasinskaja, & Kölsch, 2006).

Les études de perception de Butterworth (1980), Umeda (1982), Grosz et Hirschberg (1992), J. Pijper et Sanderman (1994), Swerts et al. (1994) et Silverman (1987) confirment d'ailleurs l'importance des variations prosodiques, comme indices acoustiques, dans la segmentation du discours. Silverman (1987) rapporte en effet, qu'en anglais, les auditeurs se fient à 70.4% du temps à la structure prosodique de l'énoncé pour délimiter, dans un texte lu, des paragraphes. Dans l'étude de Grosz et Hirschberg (1992), ils repèrent les clauses parenthétiques dans 89.2% des cas, lorsqu'elles sont énoncées dans un registre bas, les citations à 88.5%, lorsqu'elles sont énoncées dans un registre haut et à 86.4%, lorsqu'ils se fient aux pauses et au débit d'élocution.

Organisation informationnelle

Les variations de registre, de tempo et d'intensité révèlent donc à la fois la structure hiérarchique et l'organisation informationnelle du discours. Outre la signalisation de frontières majeures et parallèlement de continuation ou changement de topique, elles sont associées aux éléments porteurs d'information « nouvelle » (*new information*) et permettent ainsi la distinction de ces derniers avec les éléments porteurs d'information « donnée » (*given information*) (Halliday, 1967; Chafe, 1974), éléments « inférables » ou encore « évoqués » (E. Prince, 1981) [7]. En effet, dans de nombreuses langues, les éléments porteurs d'une lourde charge d'information sémantique, non connue, nouvelle, sont signalés par un ralentissement de la vitesse d'articulation et un registre légèrement élevé. En revanche, toute information prévisible, inférable de par le contexte, est généralement marquée par un débit d'élocution plus rapide (G. Brown, 1983; Fowler & Housum, 1987; Eefting, 1991; Koopmans-Van Beinum, 1992; Beinum & Donzel, 1996; Batliner et al., 1997; Koiso, Shimojima, & Katagiri, 1998). De telles variations permettent aux locuteurs d'attirer l'attention des auditeurs sur un élément en particulier et ainsi de les informer de la façon dont ils doivent interpréter le message. L'étude de Fowler et Housum (1987) montrera notamment que si la durée d'un élément porteur d'information nouvelle est écourtée, les auditeurs, privés de tout contexte, ont des difficultés à le reconnaître. Si, inversement, l'information est inférable, la perception des auditeurs n'est pas affectée. Si le caractère « nouveau » de l'information est signalé par des variations de la vitesse d'articulation, on constate que les locuteurs utilisent ces stratégies pour la mise en avant d'un élément par rapport au reste de la chaîne linguistique, ie. l'emphase ou focus [8]. Dans de nombreuses langues, en effet, l'emphase est marquée par une augmentation de la durée, ie. un ralentissement du débit d'élocution et une augmentation de la durée des pauses qui la précèdent (Cooper, Eady, & Mueller, 1985; Eady & Cooper, 1986; Ericson & Lehiste, 1995; Strangert, 2003; Arvaniti & Garding, 2007) ainsi que par une élévation de la hauteur et/ou par une expansion de l'étendue du registre sur l'élément focalisé (G. Brown, 1983; Eady & Cooper, 1986; Ladd & Morton, 1997; Man, 2002; Xu, Xu, & Sun, 2004; Xu & Xu, 2005; Wang & Xu, 2006; Arvaniti & Garding,

7. Les termes d'information « nouvelle » vs. « donnée », « inférable », « évoquée », ont posé des difficultés dans la définition même de ce qui les qualifie. Halliday (1985, p40) définit une information « nouvelle »comme toute information ne pouvant être inférée du contexte et l'oppose ainsi au terme d'information « donnée », défini comme toute information pouvant être inférée du contexte ; chez Chafe (1974), la notion d'information « donnée »repose sur la supposition que fait le locuteur sur la connaissance qu'a son auditeur de l'élément donné. E. Prince (1981) renverra la conception d'« information donnée » définie par Halliday (1967) au terme de « prévisible » (*givenness : predictability/recoverability*, p226), celle de Chafe (1974) au terme de « saillant » (*givenness : saliency*, p228) et distinguera pour sa part ce qui est « nouveau » (*new*), de ce qui est « inférable » (*inferrable*) et de ce qui est « évoqué » (*evoqued*). Pour ce travail, nous pensons qu'une définition fine de ces termes n'est pas nécessaire et considérons donc les termes de « donné », « inférable » et « évoqué » synonymes. Ils renvoient globalement aux définitions proposées par Halliday (1967), Chafe (1974) et E. Prince (1981).

8. Dans ce travail, nous distinguons ce qui est « emphatique » de ce qui est « non emphatique » (Di Cristo, 1998) ; le terme d'« emphase » y est synonyme de « focus », bien que nous admettions l'interchangeabilité de ces termes discutable ; cf. Hirst et Di Cristo (1998).

2007). En effet, l'étude d' Arvaniti et Garding (2007) rapporte que le focus est caractérisé par une élévation de la hauteur du registre (les pics et les creux s'élevant) tandis que celles de Liberman et Pierrehumbert (1984), A. Rietveld et Gussenhoven (1985), Ladd et Morton (1997), Gussenhoven et Rietveld (2000), Man (2002), Xu et al. (2004), Xu et Xu (2005), Wang et Xu (2006) et Chen et Gussenhoven (2008) argumentent plutôt en faveur d'une expansion du registre comme indice acoustique de l'emphase (les pics s'élevant, les creux s'abaissant). Les auteurs constatent aussi une accélération de la vitesse d'articulation ainsi qu'un rétrécissement de l'étendue du registre et un abaissement de sa hauteur au niveau des éléments post-focalisés (Cooper et al., 1985; Eady & Cooper, 1986; Ericson & Lehiste, 1995; Di Cristo, 1998; Xu et al., 2004; Xu & Xu, 2005; Wang & Xu, 2006; Patil et al., 2008). Le débit d'élocution et le registre des régions pré-focales restent relativement stables ; il apparaît donc que la décélération du débit ainsi que l'élévation et l'expansion du registre sont restreints à l'élément focalisé. Parce que les séquences pré-focales restent inchangées, Xu et Xu (2005) concluront en une « asymétrie radicale » autour de l'élément focalisé. En effet, nous pouvons distinguer trois zones de variations de registre et de tempo qui permettent l'encodage de l'emphase : (1) une région pré-focale neutre, (2) une région focale pour laquelle le registre est étendu et élevé et la vitesse d'articulation ralentie, et (3) une région post-focale pour laquelle le registre est abaissé et rétréci, la vitesse d'articulation plus rapide. Cependant, cette construction schématique a été remise en question en français par Delais-Roussarie, Rialland, Doetjes, et Marandin (2002) qui pensent que les séquences postfocales ont un registre variable en fonction de leur réalisation métrique et de l'information ou topicalisation qu'elles communiquent.

Fonctions extra-linguistiques

Les variations à long terme revêtent également des fonctions extralinguistiques, elles caractérisent en effet à la fois l'individualité du locuteur et le style discursif qu'il pratique (Di Cristo, 2000; Trouvain, 2004). Qualifiée d'informations diagnostiques, marqueurs d'identité ou encore marqueurs attributifs (Lienard, 1989; Di Cristo, 2000; Verhoeven, 2002), la fonction identificatrice laisse en effet présager le sexe, l'âge, l'aspect physique, l'état de santé, la personnalité ou encore le statut socioprofessionnel, l'origine géographique et ethnique du locuteur.

Sexe

Les différences de tessiture et les variations de registre, résultantes des caractéristiques physiologiques des locuteurs (e.g. la dimension et la masse volumique des cordes vocales, la configuration de l'appareil vocal), permettent de toute évidence la distinction des voix masculines et féminines. Sur un ensemble de 534 individus âgés de 18 à 36 ans, Hollien, Dew, et Philips (1971) rapporteront, par exemple, que les voix d'hommes se situent entre 78 à 698 Hz alors que celles des femmes sont plus hautes et couvrent un champ plus large, de 139 à 1108 Hz. Les variations de registre, ie. la gamme tonale effectivement utilisée en parole, identifieraient également le

sexe du locuteur. L'étendue et la hauteur du registre seraient soumises à plus de variations chez les femmes que chez les hommes (Takefuta, Jancosek, & Brunt, 1972; McConnell-Ginet, 1978; P. Smith, 1979; Graddol, 1986; Johns-Lewis, 1986; Huber, 1989). Cependant, ce marqueur, parfois stéréotype, a été remis en question dans de nombreuses études. Les femmes n'utiseraient pas une gamme tonale plus étendue et plus variante que les hommes (Henton, 1995; Fitzsimons, Sheahan, & Staunton, 2001). Elles chercheraient plutôt, en effet, dans une société androcentrique, à se « défaire » des caractéristiques des voix de femmes, jugées, quand « trop hautes » et « trop variantes », excessivement émotives (Demers, 2000). De nombreux auteurs se sont également intéressés au tempo comme indice différenciateur homme/femme. Les études de Byrd (1992), Whiteside et Hodgson (2000), Fitzsimons et al. (2001), Verhoeven, De Pauw, et Kloots (2004), Binnenpoorte, Bael, Os, et Boves (2005), Doherty et Lee (2009) et Jacewicz, Fox, O'Neill, et Salmons (2009) mentionnent, chez des sujets adultes, que les hommes utilisent un débit plus rapide que les femmes, une différence estimée à 6% dans les études de Byrd (1992), Verhoeven et al. (2004) et de Jacewicz et al. (2009). Cette différence serait due à une différence du débit d'élocution et non du nombre de pauses. A contrario, l'étude de Saint-Bonnet et Boe (1977) révèle une autre tendance : les femmes ont un débit plus rapide que celui des hommes, une différence estimée à 5%. Dans cette étude, la différence de tempo est attribuée à une réduction du temps de pause, plus élevée chez les sujets femmes (une différence de 16%). Les études de Lass et Sandusky (1971), Malecot, Johnston, et Kizziar (1972), Tsao et Weismer (1997) et Robb, Maclagan, et Chen (2004), quant à elles, ne révèlent aucun effet du sexe du locuteur sur le débit de parole ou tempo.

Age

Les variations prosodiques à long terme seraient aussi un indice de l'âge d'un individu. Outre la distinction des voix d'enfants et d'adultes, la tessiture caractériserait les voix de ces derniers, selon leur âge, plus ou moins avancé. Dans les études de Hollien et Shipp (1972), Helfrich (1979), Laver et Trudgill (1979), Linville (1987), Russell, Penny, et Pemberton (1995) et Hollien, Hollien, et Jong (1997), les individus d'une vingtaine d'années ont en effet une tessiture plus élevée que ceux d'une quarantaine d'années et l'étendue de leur tessiture est signalée plus étroite à un âge avancé, par la perte des valeurs fréquentielles hautes et basses. Verhoeven et al. (2004), Malecot et al. (1972), O'Neill (2008) et Jacewicz et al. (2009) montrent également que les jeunes adultes ont un débit d'élocution plus rapide que leurs pairs plus âgés, ces différences étant notamment expliquées par une diminution du contrôle neurologique de la production de la parole, en lien avec le vieillissement. Les études de Kowal, O'Connell, et Sabin (1975), Walker, Archibald, Cherniak, et Fish (1992), Whiteside et Hodgson (2000) et Colletta, Pellenq, et Rousset (2005) montrent aussi que, chez l'enfant, le débit varie en fonction de l'âge. Plus un enfant avance en âge, plus son débit est rapide.

Aspect physique

Les variables prosodiques ont également été étudiées dans le cadre d'un éventuel lien entre tessiture et aspect physique d'un individu, à savoir son poids et sa taille. L'appareil vocal d'un locuteur serait en effet corrélé avec ses caractéristiques physiques ; un homme bien bâti, par exemple, serait caractérisé par une voix grave (Laver & Trudgill, 1979), et c'est sur cet indice, entre autres, que les auditeurs seraient capables de le décrire (Lass & Davis, 1976; Laver & Trudgill, 1979; Evans, Neave, & Wakelin, 2006). A l'inverse, les études de Künzel (1989) et Dommelen et Moxness (1995) montrent qu'il n'existe pas d'effet de l'aspect physique d'un individu sur la hauteur moyenne de la tessiture. Les différences rapportées répondraient plutôt du code biologique fréquentiel proposé par Ohala (1984) selon lequel la voix de l'être fort, large ou dominant est plus grave que celle de l'être faible, petit ou subordonné.

Etat de santé

Les variations de registre et de débit de parole ou tempo donneraient aussi des indications sur l'état de santé d'un individu, ie. révèleraient anormalités physiologiques, addiction aux drogues, fatigue, ménopause, etc.[9]. Sobell et Sobell (1972), Johnson, Pisoni, et Bernacki (1990), Cooney, McGuigan, Murphy, et Conroy (1998), et Hollien et al. (2001) signalent, par exemple, dans leur étude, que, sous l'emprise de l'alcool, la vitesse d'articulation est plus lente, les pauses sont plus longues et plus nombreuses. Dans l'étude de Braun et Kunzel (2003), en revanche, le débit d'élocution est plus rapide. La hauteur moyenne et l'étendue du registre varieraient également, bien que les résultats des différentes études ne convergent pas vers une même tendance : un abaissement du registre dans les études de Klingholz et al. (1988) et Johnson et al. (1990), une élévation du registre dans celles de Hollien et al. (2001) et Braun et Kunzel (2003), une expansion du registre dans celles de Klingholz et al. (1988) et Braun et Kunzel (2003). L'étude de Cooney et al. (1998), quant à elle, ne révèle aucun effet.

Personnalité

Les variations de registre et de tempo seraient aussi marqueurs de personnalité. Elles renseigneraient en effet sur le comportement, le tempérament, l'émotivité et le mental d'un individu (Mairesse, Walker, Mehl, & Moore, 2007). A partir des grandes dimensions de la personnalité (*the Big five*) que sont l'extraversion, le neuroticisme, l'agréabilité, le caractère consciencieux et l'ouverture à l'expérience, les auteurs rapportent qu'un registre haut et étendu renvoie à une personnalité positive, ie. une personne agréable, consciencieuse, modeste, polie, qui a confiance en soi, qui fait preuve d'assertivité, etc. Inversement, un registre étroit et bas renvoie à une personne désagréable, ennuyeuse. Un registre haut chez les hommes, relativement identique à celui d'une femme, marque une certaine faiblesse, un manque d'assurance. Un débit d'élocution soutenu révèle également le caractère extraverti, dynamique, ambitieux, consciencieux

9. (Sobell & Sobell, 1972; Stoicheff, 1981; Klingholz, Penning, & Liebhardt, 1988; Laver, 1991; Whitmore & Fisher, 1996; Hollien, Dejong, Martin, Schwartz, & Liljegren, 2001; Verhoeven, 2002)

et agréable d'un individu, un débit plus lent dénote politesse, professionnalisme (B. Brown, Strong, & Rencher, 1973, 1974; B. Smith, Brown, Strong, & Rencher, 1975; Apple, Streeter, & Krauss, 1979; Scherer, 1979; Loveday, 1981; Frick, 1985; Van Bezooijen, 1995). Il est à noter que les caractéristiques personnelles attribuées aux locuteurs en fonction du débit d'élocution dépendent du degré d'accélération ou de décélération. En effet, B. Smith et al. (1975) expliqueront que si certains traits sont « jugés » de façon linéaire (e.g. plus le débit est rapide, plus le caractère est dit consciencieux, et, inversement, plus le débit est lent, moins le caractère est dit consciencieux), d'autres relèvent d'une « relation inversée en U » (*inverted U-relationship*), ie. un tempo normal caractériserait un individu agréable alors qu'un tempo trop rapide ou trop lent dénoterait d'une personne désagréable.

Statut socioprofessionnel, origine géographique et origine ethnique

Les variations de registre et de tempo seraient aussi révélatrices du statut socioprofessionnel d'un individu. Demers (2000) et Verhoeven et al. (2004) rapportent notamment que les hommes de classe ouvrière utilisent un registre relativement plus haut que les hommes de statut socioprofessionnel élevé. Ces études corroboreraient ainsi le code biologique de la fréquence fondamentale » (*frequency code*) de Ohala (1983) selon lequel la voix du « socialement fort/dominant » est plus grave que celle du « socialement faible/subordonné ».

Les variations de registre et de tempo révèlent aussi l'origine géographique d'un individu. Demers (2000), par exemple, rapporte que les hommes québécois de son étude ont une voix plus basse et un registre plus réduit que les locuteurs hommes français, résultant, selon lui, d'une influence nord-américaine. Par contre, il ne trouve aucune différence de registre entre les locutrices femmes québécoises et françaises. I. Mennen, Schaeffler, et Docherty (2007), quant à eux, montreront que les femmes anglaises (accent standard du sud de l'Angleterre) qui ont participé à leur étude, ont un registre plus étendu que les femmes allemandes (Accent standard du Nord de l'Allemagne). Les débits d'élocution diffèrent également en fonction des langues. Dans les études de Grosjean et Deschamps (1975), Iivonen, Niemi, et Paananen (1995), Fackrell, Vereecken, Buhmann, Martens, et Coile (2000) et Dellwo et Wagner (2002), les locuteurs français ont un débit plus rapide que celui des Anglais et des Allemands, les Anglais un débit également plus rapide que celui des Allemands et des Finnois. En outre, des différences de débit inter-langues et dialectales sont rapportées dans les études de Byrd (1992), Robb et al. (2004), Pellegrino, Farinas, et Rouas (2004), Rouas, Farinas, et Pellegrino (2004), Verhoeven et al. (2004), O'Neill (2008) et Doherty et Lee (2009). Rouas et al. (2004) estiment que ces différences résultent des inventaires phonémiques des langues étudiées tandis que Osser et Peng (1964) ne montrent aucun effet de la langue sur le débit d'élocution. Hirst (2006) propose que les différences de débit perçues entre langues résultent d'une asymétrie entre perception et production. Osser et Peng (1964), Grosjean (1977) et de Pfitzinger et Tamashima (2006) ont également étudié l'effet de la langue maternelle sur la perception qu'ont

les auditeurs d'une langue étrangère. En général, les auditeurs sur-estiment le débit de la langue étrangère, perçue comme plus rapide que la langue maternelle, un effet expliqué, entre autres, par la connaissance, la maîtrise que les auditeurs ont de cette langue.

L'origine ethnique, enfin, est marquée par une différence de hauteur de voix, des voix rapportées plus graves pour la communauté noire que pour la communauté blanche (Hudson & Holbrook, 1982), plus aigües pour les Japonaises que pour les Caucasiennes Américaines notamment (Yamazawa & Hollien, 1992). On pourrait cependant suggérer que ces différences sont stéréotypées et qu'elles peuvent être expliquées plutôt, ici aussi, par le code biologique fréquentiel proposé par Ohala (1984).

Style discursif

De nombreux auteurs se sont penchés sur les indices prosodiques qui participeraient à la caractérisation de différents modes ou styles de parole [10]. Ils se sont notamment intéressés aux dichotomies parole préparée/spontanée, parole formelle/informelle, monologale/ conversationnelle, etc.

Il apparaît, en anglais et en français, que les paroles spontanée et dialogique sont caractérisées par un registre plus haut et plus étendu que les paroles préparée et monologale (Graddol, 1986; Daly & Zue, 1992; Gut, 2007). Pourtant, dans son étude, Johns-Lewis (1986) rapporte que les locuteurs utilisent un registre plus bas et plus réduit en conversation qu'en monologue joué (*acted monologue*). Cette caractéristique, selon l'auteur, révèle en fait le type d'audience à laquelle s'adressent les locuteurs. En entretien, le registre serait plus bas et plus étroit de par le caractère intimiste de la conversation ; en revanche, le registre serait plus haut et plus étendu en monologue joué, car le locuteur « joue » le monologue, comme s'il s'adressait aux auditeurs d'une radio. Les variations de registre ne révèlent donc pas seulement le caractère spontané vs. préparé de la parole mais également son caractère situationnel [11]. Ainsi, deux paroles monologales peuvent être distinguées par des variations de registre selon le caractère situationnel qu'elles induisent. D. Crystal et Davy (1969) montreront, par exemple, dans leur étude, que les commentaires funéraires ont un registre plus réduit que les commentaires de criquet, pour lesquels l'apogée de l'action se traduit par un registre étendu et un tempo accéléré.

Le caractère plus ou moins exalté, passionné du locuteur mais aussi le caractère plus ou moins formel de la parole font qu'un même type de parole peut être caractérisé par des registres différents. Par exemple, en anglais, les informations radiophoniques ont un registre plus étroit

10. Nous ne rapportons ici que les études menées sur l'anglais et le français ; cf., entre autres, Beinum et Donzel (1996) pour le néerlandais ; Delgado-Martins et Freitas (1991) pour le portugais ; Fónagy et Magdics (1960) pour le hongrois ; Hirose et Kawanami (1998) pour le japonais ; Fon (1999) pour le mandarin.

11. Nous parlerons ainsi dans cette thèse de « parole authentique » ; définie ci-après en section 2.

que les présentations de concert ou les commentaires sportifs (Bhatt & Léon, 1991) ; le registre est plus étendu et plus haut en parole informelle qu'en parole formelle (Sityaev, Webster, Braunschweiler, Buchholz, & Knill, 2007).

Ayers (1994), quant à elle, rapporte que les différences d'étendue et de hauteur du registre ne sont pas fonction des types de production mais plutôt du locuteur. De plus, ce qui différencie, selon l'auteur, la parole lue de la parole spontanée, ce sont leur organisation informationnelle et leur dimension hiérarchique. La parole lue est en effet marquée par des structures de topique hiérarchiques claires alors que ces mêmes structures, en parole spontanée, sont interrompues par des faux-départs, des reprises et l'influence de possibles tours de parole. La distinction entre parole lue et parole spontanée n'est donc pas systématiquement évidente. Les indices prosodiques, à eux seuls, ne permettent pas cette distinction, ils y participent en combinaison avec d'autres (e.g. hésitations, reprises, qualité de voix, réduction vocalique, etc.). Ayers (1994) montre d'ailleurs, à partir d'un test de perception, que les auditeurs sont capables de distinguer correctement parole lue et parole spontanée entre 64% et 74%, un score bien moins élevé que celui de Blaauw (1991) (82%). L'auteur explique que, parce que la tâche de lecture, dans son étude, est calquée sur la conversation spontanée des participants, la syntaxe y est donc plus typique de la parole conversationnelle et, par conséquent, rend difficile la dichotomie de ces deux modes de parole.

Les variables temporelles participent aussi à la distinction des modes de production de la parole. En anglais, par exemple, la vitesse d'articulation (mesurée en termes de durée de voyelles, syllabes par seconde, etc.) est rapportée plus lente et plus variable en parole conversationnelle qu'en parole lue (Ayers, 1994; Hirschberg, 2000). En revanche, les études de Goldman-Eisler (1968), Grosjean et Deschamps (1972b) et Silverman, Blaauw, Spitz, et Pitrelli (1992) révèlent que le débit d'élocution reste plutôt constante et que ce sont le nombre, la durée, le type et la distribution des pauses qui varient. En effet, les paroles conversationnelles et spontanées, outre les faux-départs, les hésitations, les reprises, que nous avons mentionnés précédemment, sont marquées par des pauses remplies et des pauses silencieuses plus longues et plus nombreuses qu'en parole lue, qui peuvent se trouver au sein de constituants syntaxiques, alors qu'en parole lue, elles sont restreintes aux frontières grammaticales majeures. Les problèmes de planification rencontrés en parole spontanée augmenteraient en effet le nombre de pauses remplies et silencieuses et résulteraient en des portions interpausales plus courtes. En revanche, le procédé d'abstraction impliqué lors de la formulation de sens n'aurait, selon Goldman-Eisler (1961), aucun effet sur la vitesse d'articulation.

De plus et pareillement aux variations de registre, les différences temporelles reflètent non seulement la dichotomie de ces deux modalités (parole lue vs. parole spontanée) mais également les sous-ensembles qu'elles représentent. Duez (1982) rapporte par exemple, qu'en français, les pauses remplies sont plus fréquentes et plus longues lors d'interviews que lors de discours

politiques ; la durée totale des pauses silencieuses des discours politiques est en revanche 50% supérieure (plus fréquentes et plus longues) à celle des interviews. Cela peut être dû au fait que l'auditeur sait qu'il ne sera pas interrompu. Grosjean et Deschamps (1972b) rapportent, en anglais et en français, que la description oralisée d'images se distingue des interviews radiophoniques par des pauses plus courtes et moins nombreuses. En revanche, dans l'étude de Goldman-Eisler (1968), les pauses ne varient pas en fonction du type de production. Selon Grosjean et Deschamps (1972b), cette différence résulterait du niveau cognitif de l'opération verbale requise pour chacun de ces modes :

> L'élément de base qui permettra [aux différents modes de production] de se confondre ou non est en grande partie le degré d'aisance de l'encodage qui est lui-même directement lié à la tâche linguistique requise du locuteur. Demander à un sujet de parler de son travail ou de son principal centre d'intérêt est forcément moins contraignant pour lui au niveau de l'encodage que de lui demander de participer à une discussion ou de l'interviewer sur un sujet qui lui est moins familier, voire même qui le laisse indifférent. Et la contrainte sera aussi forte sinon plus si on lui demande de faire un reportage sur un événement qui se déroule devant lui ou de décrire un dessin humoristique sans mentionner l'interprétation de celui-ci (Grosjean & Deschamps, 1972b, p192).

L'interview et la description d'images, dans l'étude de Goldman-Eisler (1968), relèvent donc d'un même effort cognitif, ce qui n'est pas le cas dans l'étude de Grosjean et Deschamps (1972b).

Ainsi, les variations temporelles ne reflètent pas seulement le type de parole mais l'effort cognitif qu'il requiert. Les variations du débit d'élocution et/ou le nombre, la durée et la distribution des pauses révèleraient en effet des difficultés d'accès au lexique, des retards dans la construction syntaxique et des problèmes de planification sémantique (Miller et al., 1984). La parole spontanée requiert donc un effort cognitif plus important que la parole préparée, ce qui expliquerait que le tempo global soit généralement plus lent. Elle se distinguerait également de la parole préparée par des alternances de « phases coulantes » (*fluent* ; plus d'effort articulatoire et moins de pauses) et de « phases hésitantes » (moins d'effort articulatoire ; plus de pauses), qui reflèteraient un système de hiérarchie selon lequel le tempo est plus lent à une haute planification de sens et plus rapide lorsque cette planification a été effectuée (Fon, 1999).

Nous retiendrons de ces études que les variations de registre et de tempo participent en effet à la distinction des styles de parole mais qu'ils ne permettent pas leur dichotomisation en parole lue/ préparée vs. spontanée. Si l'on veut décrire les indices prosodiques participant à la distinction de ces modes, il faut prendre en considération plusieurs aspects, tels que la familiarité entre les participants d'une conversation, leur nombre, leur statut social, l'effet que

cherche à avoir un locuteur sur son auditoire (Eskenazi, 1993; Barry, 1991). Nous verrons d'ailleurs, dans le chapitre 3 de cette thèse, qu'une description multidimensionnelle des types de discours est nécessaire à la compréhension des multiples fonctionnalités que revêtent les variations prosodiques à long terme.

Fonctions para-linguistiques

La participation individuelle ou la combinaison des caractéristiques prosodiques telles que le registre, le tempo, et l'intensité assume aussi des fonctions paralinguistiques ou paragrammaticales (Ladd, Silverman, Tolkmitt, Bergmann, & Scherer, 1985; K. A. Fant G. & Nord, 1990; Di Cristo, 2000; Patterson, 2000; Mozziconacci, 2002). On parle ainsi dans la littérature d'« informations affectives » (*affective information*, Laver et Trudgill (1979)), de « marqueurs affectifs » (*affective markers*, Abercrombie (1967)) ou encore de « marqueurs d'état » (*state markers*, Verhoeven (2002)) pour décrire le rôle joué par les variations prosodiques dans la signalisation de l'état psycho-physiologique du locuteur. Ces variations révèlent en effet des changements de l'état émotionnel du locuteur, telles que la peur, la surprise, la colère, la joie, l'ennui (C. Williams & Stevens, 1972; Cosmides, 1983; Ladd et al., 1985; Ladd, 1996; Trouvain, 2004) et outre la fonction communicative des émotions primaires (joie, colère, etc .) ou affect (Scherer, 1979; Mozziconacci, 1998), les variations à long terme participent à la communication des émotions socialisées ou attitudes, telles que le doute ou la surprise (Ladd, 1996; Fujisaki, 1991; Mozziconacci & Hermes, 2000; Di Cristo, 2000). Elles permettent aussi de déceler les intentions du locuteur, de son état conversationnel (agression, apaisement, solidarité, condescendance) (Laver & Trudgill, 1979; Scherer, 1979; Fujisaki, 1991; Ladd, 1996; Shriberg, Ladd, Terken, Int, & Park, 1996; Di Cristo, 2000; Patterson, 2000; Mozziconacci & Hermes, 2000; Wichmann, 2000; Di Cristo, 2004) [12].

Les variations de registre, de tempo et d'intensité permettent ainsi de distinguer les émotions fortes des émotions modérées, les émotions positives des émotions négatives. Parmi les émotions fortes et positives, les auteurs comptent généralement l'excitation, la surprise et la joie, parmi les émotions fortes et négatives, la colère, la peur, le mépris et l'indignation. La satisfaction et l'intérêt sont regroupés sous l'étiquette d'émotions modérées positives, la contrariété, l'ennui et la tristesse, celle d'émotions modérées négatives (Cowie, Douglas-Cowie, & Romano, 1999; Mozziconacci & Hermes, 2000; Wichmann, 2000). Nous proposons une schématisation de ces catégorisations dans la figure 1. Dans la littérature, il est généralement rapporté qu'un registre étendu et haut, un débit de parole ou tempo accéléré et une plus forte amplitude

12. Nous présentons, dans cette section, les émotions dites « primaires » que sont la joie, la colère, etc., et la façon dont elles sont révélées par les variations prosodiques à long terme, bien que nous considérions fragile l'association du terme global d'« émotions » à celui d'« émotions primaires » uniquement. cf. les discussions de Murray et Arnott (1993), Scherer et Wallbott (1994), Wichmann (2000), Cowie et Cornelius (2003), Scherer (2003), Schroder (2003) et Caelen-Haumont (2005) à ce sujet.

sont associés à des émotions fortes, qu'elles soient positives ou négatives, alors qu'un registre compressé et bas, un débit plus ralenti et une amplitude plus faible sont plutôt associés à des émotions modérées, qu'elles soient positives ou négatives également (Abercrombie, 1967; C. Williams & Stevens, 1972; Laver & Trudgill, 1979; Brazil, 1980; Cosmides, 1983; Scherer, Ladd, & Silverman, 1984; Frick, 1985; Murray & Arnott, 1993; Banse & Scherer, 1996; Cowie et al., 1999; Montero, Gutiérrez-Arriola, Colás, Enríquez, & Pardo, 1999; Paeschke, Kienast, & Sendlmeier, 1999; Mozziconacci & Hermes, 2000; Mozziconacci, 2000; Paeschke & Sendlmeier, 2000; Wichmann, 2000; Breitenstein, Van Lancker, & Daum, 2001; Trouvain, 2004; Biersack & Kempe, 2005; Caelen-Haumont, 2005; Burkhardt et al., 2006).

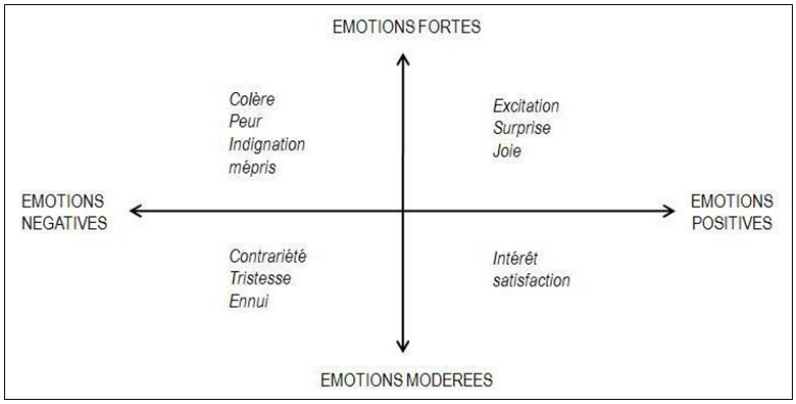


FIGURE 1 – Schématisation des catégorisations des émotions selon des degrés d'intensité et de valence.

De ce constat naît une problématique : comment distinguer les émotions fortes négatives des émotions fortes positives ; ou encore quel indice participe à la distinction des émotions modérées positives et des émotions modérées négatives ? Ladd et al. (1985) répondra que la combinaison des modifications de registre et de qualité de voix peuvent permettre une telle distinction ; les variations de registre indiqueraient le degré d'intensité d'une réaction émotionnelle tandis que la qualité de voix reflèterait le caractère positif ou négatif de la réaction. Wichmann (2000), quant à elle, pense que les variations de registre combinées à des indications d'orientation positive ou négative non verbales participent à la distinction des émotions positives et négatives. Un registre étroit, combiné à un sourire, révèlerait ainsi la satisfaction du locuteur ; un registre étroit, combiné avec des indications d'orientation négative, reflèterait son ennui, voire sa tristesse. De nombreuses études ont également montré que les contours intonatifs participeraient à l'expression des émotions (Cosmides, 1983; C. Williams & Stevens, 1972; Ladd, Scherer, & Silverman, 1986; Mozziconacci, 1998; Mozziconacci & Hermes, 1999; Mozziconacci, 2000; Banziger & Scherer, 2005).

Plusieurs indices verbaux (registre, tempo, contour, qualité de voix, etc.) et non-verbaux (posture, geste, expression du visage, etc.) révèlent ainsi de l'état émotionnel d'un individu. Cependant, même si les auteurs rapportent des tendances similaires dans les signaux des émotions, les individus peuvent cependant utiliser des stratégies différentes. Les études de Gibbon (1998), Mozziconacci (2000), Breitenstein et al. (2001), Ibrakhim (2004), Beller, Schwarz, Hueber, et Rodet (2006), entre autres, révèlent en effet des différences idiosyncratiques et inter-culturelles, tant dans la communication des émotions que dans leur perception, d'où la difficulté de leur modélisation et de leur reconnaissance. Dans l'étude de Mozziconacci (2000), par exemple, les locuteurs utilisent des stratégies différentes pour l'expression de la colère : deux locuteurs ralentissent leur débit d'élocution quand le troisième l'accélère. La peur et l'indignation sont également exprimées de manière différente : un débit lent pour un locuteur, un débit rapide pour les deux autres. Dans l'étude de Beller et al. (2006), la tristesse est exprimée par un registre haut, contrairement à ce qui est rapporté dans de nombreuses études. Les auteurs concluent que cette différence idiosyncratique soulève la fragilité de toute généralisation. Les études de Breitenstein et al. (2001) et de Graham, Hamblin, et Feldstein (2001), quant à elles, rapportent des différences majeures entre auditeurs natifs et non natifs d'une langue dans leur capacité à identifier les émotions de la parole. Cette capacité chez les non-natifs dépendrait de la connaissance qu'ils ont de la langue étrangère. Plus un individu maîtrise une langue étrangère, mieux il perçoit les émotions de cette langue.

Globalement, on peut donc dire des variations prosodiques à long terme qu'elles ont des fonctions linguistiques, extra-linguistiques et para-linguistiques. Elles signalent la structure informationnelle, l'organisation hiérarchique et relationnelle du discours, l'âge, le physique, l'état de santé, l'origine socio-professionnelle et géographique d'un individu, le style de parole qu'il adopte ou encore son état émotionnel. Si les études rapportées divergent sur certains points, elles ne peuvent pour autant être contestées. Elles ne font que rappeler le caractère à la fois universel et spécifique des variations prosodiques et de leurs fonctionnalités. Les variations prosodiques se manifestent en effet universellement, ie. elles dépendent des caractéristiques biologiques et physiologiques d'un individu ; elles s'avèrent aussi différentes, ie. elles relèvent de facteurs socio-culturels, particuliers à chaque langue, à chaque communauté, et en ce qui concerne l'individu, de facteurs idiosyncratiques (Abercrombie, 1967; Tickle, 2000; Xu & Xu, 2005).

Les divergences soulevées dans la littérature révèlent également la difficulté de neutraliser les facteurs interférents dans l'étude d'un phénomène spécifique et ainsi confirment l'existence de la multiplicité de fonctions que revêtent les variations prosodiques. En effet, étudier le débit de parole en fonction de l'âge d'un locuteur par exemple, requiert que les variables interférentes (état émotionnel, personnalité du locuteur, style de parole utilisé, etc.) aient été au préalable neutralisées.

Elles rappellent encore la difficulté de rassembler et de confronter des études portant sur un même sujet, lorsque celles-ci manquent d'utiliser une même méthodologie, de mêmes mesures, de mêmes domaines, et se font à partir d'échantillons de données différents (e.g. langues, individus). Nous avons rapporté par exemple de la littérature les fonctions que revêtent les variations prosodiques à long terme que sont le registre, le tempo et l'intensité. Or, lorsqu'on regarde de plus près l'empan temporel utilisé dans l'étude de ces variations, on s'aperçoit que ces variations dites à long terme sont évaluées sur un empan temporel parfois équivalent à celui du mot. Si certains auteurs considèrent l'empan temporel des variations de registre, de tempo et d'intensité « très localement » (e.g. au niveau du mot), quand d'autres les évaluent « plus globalement » (e.g. au-delà de l'énoncé), on comprend dès lors les divergences rapportées dans la littérature.

A partir de ces exemples, il apparaît clairement la nécessité de distinguer les variations prosodiques à court et à long terme. Leur interaction, leur chevauchement et les diverses fonctions qu'elles assument exigent en effet leur étude séparée. Ostendorf (1993) le souligne d'ailleurs, « La prosodie peut opérer à des niveaux multiples (e.g., le mot, le syntagme, la phrase, le paragraphe), rendant la modélisation computationnelle de la prosodie particulièrement difficile » (notre traduction ; (Ostendorf, 1993, p315)). Il en résultera, selon nous, une meilleure analyse des variations à plus court terme et une amélioration des systèmes d'analyse automatique et de modélisation de l'intonation et du rythme de la parole.

2 Objectifs

La problématique de l'empan temporel des variations prosodiques se veut donc au coeur de cette thèse. Parce qu'elle est sous-jacente et commune aux travaux menés en prosodie, et ce, malgré la disparité et l'aspect pluriel de ces disciplines, elle nécessite, selon nous, une attention tout à fait particulière.

On distingue, d'un côté, les multiples travaux menés sur l'intonation, d'un autre côté, les études, bien moins foisonnantes, portant sur le registre. Hormis quelques rares travaux, la plupart des études conduites ont en effet observé indépendamment chacun de ces phénomènes, si bien que de nombreux outils d'analyse automatique de l'intonation des langues ne prennent pas en compte les variations de registre, considérées communément stables. L'analyse et la modélisation du rythme de la parole ont en revanche très tôt amené les auteurs à prendre en considération les effets de tempo (Klatt, 1976) mais l'interaction de ces différents effets temporels est loin d'être résolue.

Ce travail se veut donc de soulever la problématique de l'empan temporel des variations prosodiques, par l'étude des variations à long terme que sont le registre et le tempo. Nous

proposons en effet une étude de l'empan temporel de leurs variations et des fonctions qu'elles revêtent, notamment la façon dont elles renseignent sur l'identité du locuteur ou encore la façon dont elles indiquent la structure intentionnelle du discours. Cette étude a par ailleurs nécessité l'élaboration d'outils de détection automatique des variations de registre et de tempo, que nous présenterons dans le corps de nos chapitres expérimentaux. Nous verrons aussi que de tels outils peuvent être implémentés dans des outils d'analyse automatique de l'intonation (e.g. INTSINT ; Hirst (2007)) ou du rythme de la parole, et qu'ils permettent une meilleure qualité d'analyse automatique et de synthèse des variations à plus court terme.

Nous travaillons donc à partir de deux paramètres acoustiques : ie. la fréquence fondamentale et la durée, afin d'analyser, de détecter et de prédire automatiquement les variations de registre et de tempo, et concentrons nos travaux de recherche sur les variations prosodiques de deux langues, l'anglais et le français. Il semble en effet intéressant d'étudier deux langues dont les structures prosodiques divergent et de voir si de telles différences se retrouvent à des niveaux plus globaux.

Nos analyses se feront à partir de 4 corpus : les corpus PFC et CID pour l'étude du français parlé et les corpus PAC et AIX-MARSEC pour l'étude de l'anglais parlé [13]. Les corpus choisis pour cette étude marquent notre intérêt pour la parole « authentique », ie. toute parole produite dans le but de communiquer du sens (Hirst, Bouzon, & Auran, 2007). La parole authentique ne doit être pour autant confondue ou synonyme de parole spontanée, nous classons en effet plutôt cette dernière comme l'une de ses composantes, ni ne doit être strictement opposée à la parole dite de laboratoire, qui, bien que dans de nombreuses études est dépourvue de fonction communicative, peut, selon le protocole expérimental choisi, être qualifiée comme parole communicative. Ce qui caractérise la parole authentique, c'est la volonté du locuteur à faire passer du sens, de l'information à son auditeur. Ainsi, un monologue préparé et oralisé à la radio, par exemple, est considéré comme parole authentique, comme l'est une conversation à bâtons rompus, entre deux locuteurs, où le locuteur cherche à se faire comprendre de son interlocuteur. Au contraire, une lecture oralisée de phrases ou de textes répondant à une simple tâche de lecture et non de communication, est, elle, considérée comme non authentique. La tâche demandée au locuteur n'est pas de communiquer une information dans le but de se faire comprendre, elle n'est que pure tâche de lecture. Nous pensons en effet que l'étude des faits prosodiques, notamment à long terme, est plus pertinente sur des corpus de parole authentique car leurs variations pour cette catégorie sont nombreuses et endossent des fonctions multiples. Pour autant, afin que nos algorithmes puissent être évalués pour tout type de parole, nous avons choisi des corpus représentatifs de parole authentique et non authentique, bien que la plupart des enregistrements relèvent de la parole authentique. Les corpus ont également été sélectionnés pour les différents types de production qu'ils représentent et pour la possibilité

13. Les corpus seront présentés en détails dans nos chapitres expérimentaux

d'une analyse comparable de la fréquence fondamentale, de la durée et de l'intensité, que leur nature permet.

Nous proposons à présent de définir le cadre théorique et formel de ce travail, avant de passer au premier chapitre de cette thèse, dans lequel nous définirons plus précisément les termes de registre et de tempo et relèverons les problématiques sous-jacentes à l'étude de leurs variations.

3 Cadre théorique et formel

3.1 Théories autosegmentale et métrique

Bien que nous nous prévalions de tout cadre théorique pour que notre travail puisse être interprété à la lumière d'approches théoriques différentes, nous nous axons dans le cadre de la phonologie autosegmentale-métrique (pour référence Goldsmith (1976), Bruce (1977), Pierrehumbert (1980), Goldsmith (1990), Goldsmith (1999), Ladd (1996), Gussenhoven (2002), Gussenhoven (2004), Gussenhoven (2005), Gussenhoven (2007)), en particulier dans la version développée par Hirst et Di Cristo (INTSINT, ProZed - Hirst, Di Cristo, et Espesser (2000), Hirst et al. (2007), Hirst (2007)). Les approches autosegmentale et métrique, bien que très fortement liées de par leur complémentarité dans l'étude de la représentation prosodique des langues naturelles, s'attachent à décrire des phénomènes prosodiques spécifiques : les approches autosegmentales se consacrent à la description phonologique de l'intonation, les approches métriques à la description formelle de l'accentuation et du rythme de la parole. Cependant, elles ne se veulent pas particularisées, la théorie métrique intégrant des concepts d'organisation tonale, l'approche autosegmentale associant tons et proéminences. Nous verrons d'ailleurs au cours du premier chapitre (section 1.4) la façon dont certains auteurs ont pu s'inspirer des théories métriques pour une représentation phonologique du registre.

3.1.1 Approche autosegmentale

Nous chercherons tout d'abord, dans notre travail, à évaluer la place accordée aux variations de registre dans la description des patrons intonatifs de l'anglais et du français, sous le couvert d'approches dites autosegmentales (Woo, 1969; Leben, 1973; Goldsmith, 1976; Bruce, 1977; Clements & Ford, 1979; Pierrehumbert, 1980; Liberman & Pierrehumbert, 1984; Pierrehumbert & Beckman, 1988; Goldsmith, 1990; Pierrehumbert & Hirschberg, 1990; Ladd, 1996; Gussenhoven, 2002). La représentation autosegmentale, contrairement aux représentations génératives traditionnelles (*The Sound Pattern of English*, Chomsky et Halle (1968)), admet une représentation multidimensionelle où les tons deviennent des entités indépendantes, représen-

tés à un niveau autonome de la chaîne phonémique. Ainsi, sur une première rangée (*tier*) [14], sont représentés les éléments ou unités porteuses de tons (*tone bearing units*), sur une couche supérieure, indépendante, les tons figurés comme une chaîne d'éléments catégoriels, dits autosegments. Cette représentation à plusieurs niveaux, à l'origine avancée par Leben (1973), permet l'association des tons et des unités porteuses de tons par l'utilisation de lignes ou de branches d'association (*association lines*), ajoutées ou supprimées en fonction des spécifications que requiert la structure arborescente. Les tons dynamiques (*contour or dynamic tones*) montants et descendants n'y sont plus considérés comme une seule unité phonologique atomique (Goldsmith, 1976), contrairement à l'approche britannique (O'Connor & Arnold, 1961; D. Crystal, 1969; Halliday, 1970; A. Cruttenden, 1997), mais plutôt comme une séquence de tons statiques (*level tones*) distincts haut et bas, associés aux unités porteuses de tons. Les tons montants et descendants ne sont donc plus décrits en termes de traits ([montée] ou [chute]) mais comme la concaténation de deux tons, ie. un ton bas suivi d'un ton haut pour le contour montant, et un ton haut suivi d'un ton bas pour le contour descendant.

Si les tenants de l'approche autosegmentale s'entendent globalement sur une telle représentation [15], l'une des questions toujours débattue est celle du statut d'une dimension orthogonale de l'organisation tonale. Quelle place, en effet, donner aux phénomènes d'abaissement globaux (déclinaison ; *declination* et variations de hauteur et d'étendue) ou locaux (catathèse ; *downstep* ou abaissement final ; *final lowering*) et aux variations de registre ? C'est pourquoi nous nous attacherons à regarder comment les uns et les autres, à travers diverses propositions, ont tenté de répondre à cette problématique.

3.1.2 Approche métrique

Théorie de l'accentuation et de l'organisation rythmique d'un énoncé, la théorie métrique est à l'origine proposée par Liberman (1975), puis affinée par Liberman et Prince (1977), dans leur étude sur les patrons accentuels de l'anglais. Les auteurs réinterprètent le principe de subordination accentuelle (« stress subordination principle », Chomsky et Halle (1968)) qui fait des proéminences prosodiques des traits paradigmatiques associés ou subordonnés aux segments, et exposent une organisation hiérarchique des patrons accentuels intrinsèques à la langue étudiée, par l'intégration de divers niveaux d'accentuation. L'accentuation y est représentée comme une organisation hiérarchique de positions fortes/ faibles relatives (Liberman

14. Terme emprunté à Cho (2009).

15. Regard à l'origine posé par Woo (1969), dans ses travaux sur la prosodie et la phonologie des langues à tons, puis repris par de nombreux auteurs, tels que Pierrehumbert (1980), Liberman et Pierrehumbert (1984), Pierrehumbert et Beckman (1988)), Gussenhoven (2002), Gussenhoven (2004), Gussenhoven (2005), Gussenhoven (2007), Ladd (1996), Hirst et Di Cristo (1998), pour n'en citer que quelques uns, résultant en une littérature prolifique de propositions de représentations formelles de l'intonation.

& Prince, 1977). La notion de proéminence relative entre constituants d'un énoncé est en effet la marque de la théorie métrique et c'est notamment en ce sens qu'elle se distingue des théories traditionelles. La notion de relation entre constituants et de force relative entre des unités voisines apporte ainsi un renouveau dans l'organisation des segments, par le fait qu'ils sont regroupés phonologiquement en unités plus larges. Dans la version proposée par Selkirk (1980), ces unités sont spécifiquement identifiées comme des syllabes, elles mêmes regroupées en pieds métriques, les pieds, en mots phonologiques ou prosodiques, les mots, en syntagmes phonologiques, etc. La théorie métrique, comme le souligne (Selkirk, 1980), n'est donc pas seulement une théorie de la proéminence mais une théorie de la structure hiérarchique des constituants phonologiques et des règles qui régissent leurs délimitation et emboîtement. Cet aspect nous amène à aborder la notion de constituance prosodique.

3.2 Phonologie des domaines : constituance prosodique

Très tôt les auteurs ont cherché à se dégager des structures syntaxiques et à proposer une structure hiérarchique de constituants prosodiques indépendante. On peut noter que, dans les travaux de Liberman et Prince (1977), si la hiérarchie des constituants phonologiques est indépendante de la structure syntaxique en-deçà du mot, au-delà, les structures arborescentes syntaxiques et phonologiques sont identiques. D'autres auteurs suggèrent plutôt qu'une hiérarchie des constituants phonologiques est nécessaire au-delà du mot (Pike, 1945; Selkirk, 1980; Grosjean & Dommergue, 1983; Nespor & Vogel, 1983; Selkirk, 1984; Beckman & Pierrehumbert, 1986; Ladd et al., 1986; Ladd & Campbell, 1991; Di Cristo & Hirst, 1993; Hirst & Di Cristo, 1998; Jun & Fougeron, 2000). L'idée d'une structure hiérarchique prosodique universelle et autonome, indépendante, du moins séparée, de la structure syntaxique est ainsi envisagée. La théorie de la phonologie prosodique admet ainsi une structure emboîtée de domaines : la more, la syllabe, le pied, le mot prosodique, le syntagme phonologique, le syntagme intonatif et l'énoncé (de la base au sommet de la structure).

Si les auteurs s'entendent aujourd'hui sur la notion et les termes de hiérarchie et constituants prosodiques, en revanche, les approches qu'ils choisissent pour les décrire, le nombre de niveaux qu'ils reconnaissent à cette hiérarchie et ce à quoi réfère les constituants prosodiques ne font pas consensus. On peut distinguer d'un côté les approches « prosodico-syntaxiques » (Nespor & Vogel, 1983; Selkirk, 1984) qui définissent les constituants prosodiques à partir d'aspects syntaxiques, d'un autre côté, les approches « stricto-senso prosodiques » (Beckman & Pierrehumbert, 1986; Di Cristo & Hirst, 1993; Hirst & Di Cristo, 1998; Jun & Fougeron, 2000) qui caractérisent les constituants en s'appuyant uniquement sur des critères prosodiques, relatifs à l'organisation tonale, métrique et temporelle des énoncés. Ces diverses approches font qu'il est parfois difficile aujourd'hui d'établir des liens entre les divers domaines proposés, du fait

qu'un même terme peut être utilisé pour des niveaux de constituances différents ou, qu'inversement des mêmes niveaux de constituance peuvent être désignés par des termes différents. Cette confusion reflète clairement une incertitude et une absence d'accord théorique entre les divers niveaux et le type de constituants que l'on reconnaît aux hiérarchies prosodiques. Bien que nous ayons donné l'exemple d'une structure emboîtée de 7 niveaux afin de rendre notre explication de hiérarchie de domaines plus claire, il serait faux de dire que les auteurs reconnaissent au-delà du mot, les différents constituants que nous avons nommés. Si un certain consensus semble se dégager pour les niveaux inférieurs au pied et de rang supérieur et égal au syntagme intonatif, il n'en est pas de même pour l'existence et l'interprétation des niveaux intermédiaires tels que le mot prosodique, le syntagme accentuel et intermédiaire.

Nous ne pouvons cependant ici chercher à être exhaustive, toutefois nous pensions nécessaire de soulever ce point. En effet, dans notre travail, en abordant la problématique de l'empan temporel des variations prosodiques, nous touchons directement à la problématique très complexe de la constituance prosodique et ainsi des niveaux qu'on lui reconnaît. Nous verrons qu'au cours de ce travail, pour en faciliter la lecture, nous avons dû parfois rendre synonymes des niveaux de constituance définis à partir de structures hiérarchiques différentes (e.g. Nespor et Vogel (1983), Selkirk (1984) vs. Beckman et Pierrehumbert (1986)) alors que nous ne reconnaissons pas la correspondance de ces constituants directe. L'étude des variations de registre et de tempo que nous proposons ne s'effectue donc pas à travers l'étude de constituants phonologiques ou prosodiques particuliers. Elle s'ancre en revanche dans la notion de structure hiérarchique, où nous abordons et défendons l'idée d'une structure emboîtée des variations de registre et de tempo, dont le nombre de niveaux n'est pas limité.

3.3 Approches formelle et fonctionnelle de la prosodie

Dans le cadre de la théorie autosegmentale-métrique (AM), nous adoptons l'idée d'une compositionnalité du contour, selon laquelle le contour peut être décomposé en unités plus petites. Contrairement donc à la conception holistique (Delattre, 1966; Rossi, 1999) selon laquelle le contour intonatif est un seul *gestalt* et ne peut être analysé en différents éléments, nous optons pour une représentation du contour comme une séquences d'unités, plus particulièrement comme une séquence de tons statiques ou niveaux. Il est à noter que l'idée de compositionnalité n'est pas propre à la théorie AM et qu'on la retrouve en termes de mouvements descendants et montants dans les approches dites configurationnelles, telles que celles de l'école britannique (O'Connor & Arnold, 1961; Halliday, 1970; D. Crystal, 1969; A. Cruttenden, 1997) et de la théorie IPO (Hart & Cohen, 1990) , ou encore en termes de tons ponctuels, décrits sur 4 niveaux, dans la tradition structuraliste américaine (Pike, 1948).

Dans ce travail, nous défendons aussi l'idée d'une stricte séparation des annotations formelle

et fonctionnelle dans les systèmes d'annotation prosodique. En effet, bien que les fonctions et les formes prosodiques semblent être quasi-universelles, la correspondance entre les deux, en revanche, ne l'est pas forcément. Un même patron formel peut être interprété différemment selon la langue ou le dialecte donné, ou inversement, une même fonction peut être exprimée par différentes formes. Par exemple, bien que dans de nombreuses langues une montée finale soit associée à une question, il existe des langues et des dialectes pour lesquels un énoncé déclaratif est aussi associé à une montée finale (e.g. les dialectes urbains de Belfast, Manchester, Liverpool, etc.). La séparation dans les systèmes prosodiques des formes et fonctions prosodiques nous semble donc nécessaire, notamment dès lors que l'on s'intéresse à la technologie de la parole.

Notre approche fonctionnelle se fait à partir d'une annotation manuelle fonctionnelle de la structure du discours (décrite dans le corps de nos chapitres expérimentaux), appliquée à un petit corpus de parole, afin d'amorcer un système d'annotation automatique, permettant la correspondance automatique et directe entre les données acoustiques et l'annotation fonctionnelle (cf. les démarches de Vainio, Hirst, Suni, et De Looze (2009)). Selon nous, une telle approche fonctionnelle confrontée aux données acoustiques permettrait l'extraction et la prédiction automatique de l'information fonctionnelle des données non encore étiquetées, et par ailleurs, la mise en correspondance (*mapping*) automatique entre forme et fonction prosodiques en parole de synthèse.

Nous proposons ainsi de confronter une approche objective des paramètres acoustiques par l'élaboration d'outils de détection automatique des variations de registre et de tempo, à une annotation subjective fonctionnelle, effectuée à partir d'une étude auditive. Une démarche ascendante (*bottom-up*), ie. procédant à partir de données observables, pourra selon nous, outre la prédiction automatique de l'information fonctionnelle des données non encore étiquetées, définir l'empan temporel des variations de registre et de tempo et ainsi asseoir un ou des domaines à partir desquels les variations à plus court terme (cibles tonales, durée segmentale) peuvent être décrites.

Une telle problématique nécessite en amont de comprendre les enjeux qui gravitent autour des variations de registre et de tempo. Ces enjeux, nous proposons de les aborder à travers deux axes distincts, l'un consacré au registre, l'autre au tempo. Nous définirons ainsi ces termes explicitement et présenterons les problématiques qu'ils soulèvent dans la littérature aujourd'hui.

Chapitre 1

Variations de registre et de tempo : les enjeux

1 Registre

1.1 Une définition

Le terme de « registre », nous l'avons vu, se réfère à l'ordre de structuration tonale. De façon générique, il renvoie à l'espace tonal de la voix d'une personne. Dans la littérature francophone et anglophone, de nombreux termes lui sont associés, synonymes : « tessiture », « pitch range », « pitch level », « pitch span », « register », « key », « space », « width », « tonal space », etc.

Parmi les termes listés de la littérature, nous relevons une première distinction, relative aux valeurs tonales comprises dans l'espace tonal de la voix. Soit il est fait référence à l'espace tonal comprenant et délimité par les valeurs extrêmes qu'une personne est capable d'émettre, ce que Laver et Hanson (1981) appellent caractéristiques de voix organiques ; soit il est fait référence à l'espace tonal ajusté lorsque un locuteur parle (bande de hauteur de sons émis sans difficulté qui, généralement, ne prend pas en compte, sans les exclure, les valeurs extrêmes qu'un locuteur est capable d'émettre - Hartmann et Stork (1972)), ie. les caractéristiques de voix phonétiques (Laver & Hanson, 1981). De même, Laver et Trudgill (1979) et Laver et Hanson (1981) distinguent, respectivement, les termes « range » et « span », Möhler et Mayer (1999) « pitch range » et « register », Mertens et al. (2001) « tessiture (tonale) » et « registre », pour faire référence aux caractéristiques de voix organiques et phonétiques d'un locuteur. Dans

ce travail, nous utiliserons les termes de tessiture et de registre selon la définition de Mertens et al. (2001) :

> La tessiture caractéristique de chaque voix se définit par la gamme de hauteur comprise entre deux valeurs extrêmes, minimale et maximale, appelées parfois le plancher et le plafond. [...] Cependant, dans la communication parlée, le locuteur se sert essentiellement d'une plage de hauteur au centre de sa tessiture ; c'est le registre normal (Mertens et al., 2001, p198).

Lorsque les auteurs s'intéressent aux aspects formels et fonctionnels de l'espace tonal de la voix d'un locuteur, ils étudient plutôt les caractéristiques phonétiques de la voix que ses caractéristiques organiques. Ainsi, de nombreux termes, synonymes de registre (Mertens et al., 2001) ont émané de la littérature : « pitch range » (Abercrombie, 1967; Clements, 1990; Hirschberg & Ward, 1992; Ladd et al., 1985; Ladd & Terken, 1995; Shriberg et al., 1996; Clark, 1999; Patterson & Ladd, 1999; Xu, 1999; Patterson, 2000; Portes & Di Cristo, 2003; Gussenhoven, 2004), « register » (Bolinger, 1951; Hollien, 1972; Deinse, 1981; Clements, 1990; T. Rietveld & Vermillion, 2003), « compass » ou « tessitura » (Abercrombie, 1967) et « key » (Brazil, 1980).

Outre le fait que de nombreux termes ont été utilisés pour faire référence à ce que nous appelons désormais « tessiture » et « registre », il faut noter que l'étude de ces phénomènes ne s'est pas faite de façon uniforme, et que leur conceptualisation ne s'est dessinée qu'au cours du temps. Les premières études, portant notamment sur la notion de registre, décrivaient un espace tonal dans lequel varie la fréquence fondamentale. Les auteurs s'intéressaient aux variations de registre, à sa hauteur et/ou à son étendue sans jamais clairement définir les paramètres considérés dans leurs études (Lehtonen, 1978; Abercrombie, 1967; Hollien, 1972; Ladd et al., 1985; Künzel, 1989; Ladd, 1994; Dommelen & Moxness, 1995). Bien que quelques auteurs se soient attachés à distinguer ces deux paramètres pour décrire les phénomènes intonatifs d'une langue (Bolinger, 1951; Jassem, 1971; D. Crystal, 1975; Loveday, 1981), c'est Ladd, en 1995, après avoir démontré que la hauteur et l'étendue du registre étaient deux paramètres indépendants et par soucis de mesurer de façon optimale le registre et ses variations, qui donnera au concept de registre son aspect bidimensionnel :

> There is a clear distinction between overall raising and local emphasis [...]. This could be incorporated into a quantitative model by distinguishing two aspects of what is often loosely called « pitch range », namely the overall level and the width of the space. (Ladd & Terken, 1995, p388).

> (Il existe une nette distinction entre une élévation globale et une emphase locale [...]. Cela pourrait être incorporé dans un modèle quantitatif par la distinction

de deux aspects, repris sous le terme de registre, à savoir la hauteur et l'étendue globales de l'espace tonal.)

En effet, dans cette étude, Ladd et Terken (1995) se sont intéressés aux variations de registre intra- et inter-locuteurs, afin de vérifier l'hypothèse selon laquelle les cibles tonales des contours intonatifs restent toujours invariantes (Bruce, 1977; Liberman & Pierrehumbert, 1984; Van Den Berg, Gussenhoven, & Rietveld, 1992). Dans leur expérience, 16 sujets néerlandais (8 hommes et 8 femmes) avaient pour tâche (1) de lire normalement un énoncé, puis (2) de le lire en élevant la voix, comme pour parler à travers un combiné téléphonique de mauvaise qualité et (3) de mettre en emphase un mot en particulier dans un énoncé. Les auteurs ont mesuré pour ces trois productions (normale, élévation de la voix, emphase locale) des points pré-sélectionnés, ie. la syllabe initiale inaccentuée de l'énoncé, le premier pic accentuel, le ton bas final de l'énoncé dans les affirmations, les creux bas accentuels dans les questions. Les auteurs rapportent ainsi, que, pour l'ensemble des locuteurs, il existe une nette distinction entre une élévation globale de la voix et l'emphase locale. Pour l'élévation, les pics et les creux s'élèvent alors que pour l'emphase seulement les pics sont affectés. Dans l'élévation globale donc, c'est la hauteur qui est principalement affectée alors que pour l'emphase, c'est l'espace tonal qui est étendu. La hauteur et l'étendue du registre varient indépendamment et peuvent donc revêtir des fonctions différentes [16].

Cette bidimensionnalité sera d'ailleurs reprise par de nombreux auteurs : Reissland et Snow (1996) utiliseront les termes « key » et « register », Savino et Grice (2007) les termes « pitch height » et « pitch range », Patterson et Ladd (1999), Patterson (2000) et T. Rietveld et Vermillion (2003) « pitch level » et « pitch span », Fox (2000) « height of the voice » et « pitch range », Portes et Di Cristo (2003) « register level » et « register span », Gussenhoven (2004) « pitch register » et « pitch span », Hirst (2007) et De Looze et Hirst (2008) « key » et « range » comme synonymes respectifs de hauteur et d'étendue.

Nous définissons donc le registre comme l'espace tonal effectivement utilisé dans un énoncé, une bande de hauteur de sons émis sans difficulté qui, généralement ne prend pas en compte, sans les exclure, les valeurs extrêmes qu'un locuteur est capable d'émettre. On distingue deux dimensions au sein du registre : sa hauteur (hauteur globale de voix perçue par un auditeur) et son étendue (espace tonal délimité par les valeurs maximale et minimale émises par un locuteur). Nous empruntons à Patterson (2000), la représentation schématique des modifications de ces deux dimensions (figure 2) et proposons un exemple de variations de registre (figure 3), tiré d'un de nos corpus [17], où la hauteur du registre mesurée à 157 Hz pour le premier énoncé « elles font leur brevet blanc » s'élève à 280 Hz pour le deuxième énoncé « et Laurie, elle est contente alors ? ».

16. Ce point sera développé dans nos chapitres expérimentaux
17. PFC, locuteur 13aAC1lw.

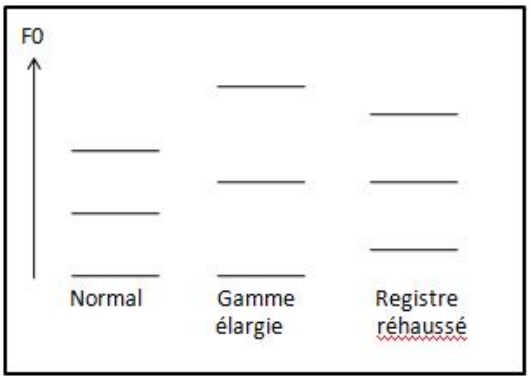


FIGURE 2 – Représentation schématique des variations de hauteur et d'étendue du registre.

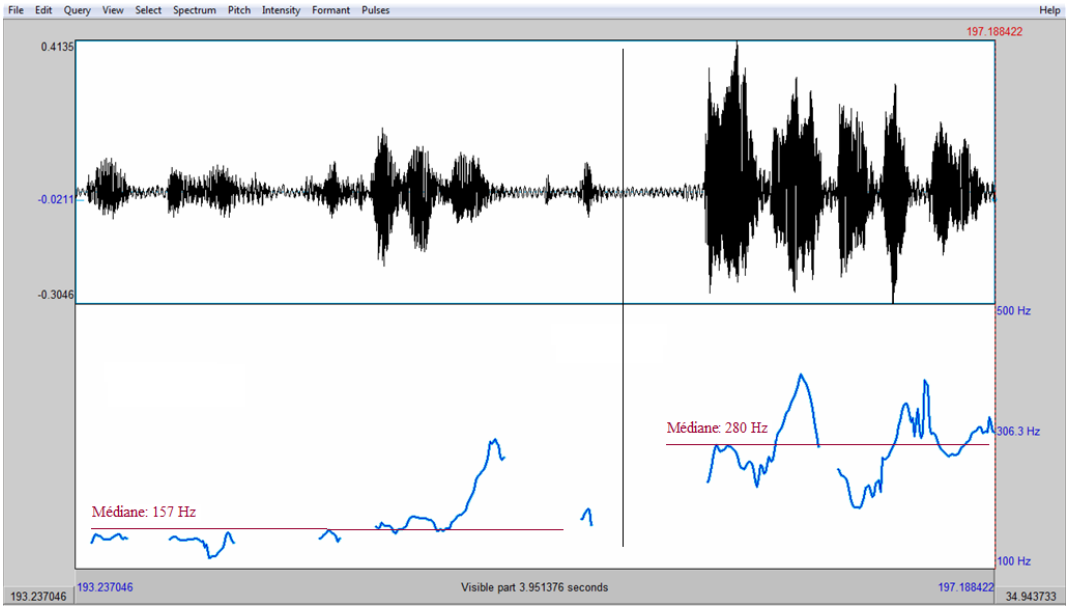


FIGURE 3 – Exemple de changement de hauteur de registre entre les énoncés « elles font leur brevet blanc » et « et Laurie, elle est contente alors ? ». La hauteur est donnée par le calcul de la médiane pour chaque énoncé. Le premier énoncé délimité par une barre verticale noire est estimé à 157 Hz alors que le deuxième est estimé à 280 Hz.

Nous définissons la tessiture comme un intervalle de fréquences délimité par les valeurs extrêmes (hautes et basses) qu'un locuteur est capable d'émettre. Nous appliquons à ce concept l'aspect bidimensionnel défini pour le registre. La tessiture est donc considérée sous deux dimensions : sa hauteur et son étendue. Elle est propre au locuteur, et dépend de facteurs physiologiques. (Lehiste, 1970; Brazil, 1980; Helfrich, 1979; Laver & Trudgill, 1979; Laver &

Hanson, 1981; Linville, 1987; Knowles, Wichmann, & Alderson, 1996; Ladefoged, 2001; Cruttenden, 2001).

Pour notre part, nous nous intéressons à l'espace tonal utilisé dans un énoncé, nous ne reviendrons donc plus sur le concept de tessiture. Le terme de registre à présent défini [18], nous proposons d'aborder les problématiques qu'il soulève. L'une des premières difficultés que pose le registre concerne les descriptions phonologiques de l'intonation des langues naturelles. Comment, en effet, décrire les faits tonals d'une langue lorsque l'espace tonal, dans lequel s'échelonne les cibles tonales, n'est vraisemblablement pas le même pour tous (e.g. voix d'homme vs. de femme) ? Ou encore, comment les décrire, lorsque l'espace tonal varie au cours de l'énoncé pour un même locuteur ?

1.2 Du caractère relatif des variations de registre

Ladd (1996) énonce que deux approches dans la littérature ont cherché à répondre à cette problématique de relativité des faits tonals : l'approche initialisante (Pike, 1948; Jakobson, Fant, & Halle, 1969; D. Crystal, 1969) et l'approche normalisante (Earle, 1975; Liberman & Pierrehumbert, 1984; Ladd et al., 1985; Rose, 1987; Berg, Gussenhoven, & Rietveld, 1992; Ladd & Terken, 1995; Ladd, 1996; Patterson & Ladd, 1999; Patterson, 2000; Hirst, 2007).

A travers une approche initialisante, les auteurs décrivent les patrons intonatifs d'une langue, sans faire référence au registre du locuteur. Une montée est une montée, quelle que soit la hauteur ou l'étendue intrinsèque du registre du locuteur (voix d'homme vs. voix de femme, vs. voix d'enfant) ou quel que soit son état psychologique (ennui vs. colère). Ce qui importe, c'est ce qui précède. Pike (1948) déclare :

> The important feature is the relative height of a syllable in relation to preceding and following syllables. It is even immaterial [...] to know the height of a specific syllable in proportion to the general average pitch which the speaker uses. Rather, one must know the relationship of one specific syllable to the other syllables in the specific context in the particular utterance. A man and a woman may both use the same tonemes, even though they speak on different general levels of pitch. Either of them may retain the same tonemes while lowering or raising the voice in general, since it is the relative pitch of syllables within the immediate context that constitutes the essence of tonemic contrast (Pike, 1948, p4).
>
> (La caractéristique importante est la hauteur relative de la syllabe par rapport à

18. Nous reviendrons dans le cours de ce chapitre sur la définition même du registre à travers les problématiques abordées.

> celles qui la précèdent et qui la suivent. Il n'est même pas nécessaire de connaître la hauteur d'une syllabe en particulier en fonction de la hauteur globale du registre d'un locuteur. Plutôt, l'on doit chercher à connaître la relation entre une syllabe en particulier et les autres syllabes d'un contexte et d'un énoncé spécifiques. Un homme et une femme peuvent utiliser les mêmes tonèmes, bien qu'ils parlent avec des niveaux globaux de hauteur différents. Chacun peut garder les mêmes tonèmes tout en abaissant ou en élevant la voix, puisque c'est la hauteur relative des syllabes au sein d'un même contexte qui constitue l'essence d'un contraste tonémique.)

Ainsi, après avoir établi un point de référence ou un point « initialisant », chaque cible tonale est dérivée de la cible tonale qui la précède. D. Crystal (1969) proposera par exemple que la cible initialisante soit la première syllabe proéminente ou « attaque » de chaque unité tonale (*tone unit*). Une chute ou une montée, dans le bas du registre du locuteur, est ainsi définie comme plus basse que celle qui la précède et, inversement, une chute ou une montée, dans le haut du registre du locuteur, est décrite comme plus haute que sa précédente. L'espace tonal est alors défini par la différence entre la cible tonale initialisante et la cible tonale finale. Cependant, cette représentation « initialisante », ie. en termes de ce qui précède, et négligeant tout repère à un espace tonal de référence, ne peut rendre compte précisément des mouvements mélodiques d'une langue.

L'approche normalisante cherche donc répondre à une telle problématique : les auteurs suggérent plutôt de décrire les patrons tonals d'une langue en termes de points de référence, spécifiques au locuteur, ie. intégrant la notion d'espace tonal (*tonal space*) ou d'échelonnage des cibles tonales (*pitch scaling*) dans leur description. Une montée ou une chute, dans le haut du registre d'un locuteur, n'est donc plus décrite comme plus haute que celle qui précède, mais, comme réalisée dans le haut du registre du locuteur. L'approche normalisante permet ainsi de décrire les patrons tonals d'une langue en normalisant les différences de registre inter- et intra-locuteurs. Une telle procédure apparaît en effet légitimée par le fait que les variations inter-locuteurs et intra-locuteurs n'affectent pas le message linguistique sous-jacent. Quel que soit l'âge, le sexe, l'état de santé du locuteur, le message communiqué est interprété de la même façon ; Liberman et Pierrehumbert (1984) et Ladd et Terken (1995) montreront, par exemple, qu'un même contour, prononcé par différents locuteurs, lorsque normalisé sur une échelle commune, est représenté par de mêmes cibles tonales. Il en est de même pour les variations intra-locuteurs. Que le locuteur élève la hauteur de sa voix lorsqu'il se trouve dans un environnement bruyant ou qu'il s'exprime sous l'effet de la colère, ou, inversement, qu'il l'abaisse lorsqu' il exprime son ennui ou sa tristesse, la relation entre les cibles tonales n'est pas affectée. Cette relation est en effet considérée invariante dans de nombreuses études (Liberman & Pierrehumbert, 1984; Berg et al., 1992; Ladd & Terken, 1995; Shriberg et al., 1996; Ladd, 1996).

Outre la relativité des faits tonals, la considération d'un espace tonal dans les représentations phonologiques de l'intonation soulève une deuxième difficulté, celle de la délimitation entre ce qui relève de la représentation phonologique et ce qui relève de l'implémentation phonétique. En effet, alors que certains considèrent les variations de registre comme purement graduelles, d'autres envisagent une interprétation plus modérée et leur allouent un caractère à la fois graduel et catégoriel. A travers cette problématique conflictuelle, se pose ainsi la question de la place accordée aux variations de registre dans les théories phonologiques de l'intonation et du ou des empans temporels qui leur sont reconnus.

En amont de cette question, nous proposons tout d'abord de relever et de définir les termes satellites qu'on peut rencontrer à la lecture de différents travaux. Autour du concept de registre ou d'espace tonal, et de la problématique du caractère graduel vs. catégoriel qu'on lui reconnaît, de nombreux termes sont en effet utilisés : « composante orthogonale », « échelle verticale » (*vertical scale*), « tone-level frame », « grille » (*grid*), « transform space », « effets d'abaissement » (*downtrends*), « déclinaison » (*declination*), « catathèse » (*downstep*), « abaissement final » (*final lowering*), « élévation du registre » (*global raising* ; *expressing raising*), « abaissement du registre » (*lowering of the voice*), « remise à niveau » (*pitch reset* ; *upstep*). Quelle interprétation admet-on de ces termes ? Quelles relations entretiennent-ils ? Les termes de registre et de « composante orthogonale », par exemple, sont ils synonymes ou emboités ? Les effets d'abaissement doivent-ils être considérés comme des variations de registre ou des effets bien distincts ? La définition que nous avons donnée du registre en 1.1 ne semble pas suffisante en l'état pour répondre à ces questions. C'est à travers la compréhension de ces termes et le lien qu'ils entretiennent qu'elle sera revisitée.

1.3 D'une revisite du terme de registre à travers la conceptualisation de ses « satellites »

Si les auteurs s'appliquent à définir les concepts qu'ils utilisent, ce à quoi ils font référence et le lien qu'entretiennent ces différents concepts ne sont cependant pas envisagés unanimement. Dans la théorie auto-segmentale métrique, dont certains de ces termes sont issus, abondent en effet de nombreuses définitions. Pierrehumbert (1980) définit par exemple la déclinaison (*declination*) comme « a gradual downdrift and narrowing of the *pitch range* ». Clements (1990) regroupe, quant à lui, sous les effets de registre (*register*), la catathèse (*downstep*), la remise à niveau (*upstep*) et l'élévation globale de la voix (*global raising*). De la même façon, Inkelas et Leben (1990) revoient au terme de registre (*register*) les effets de « downdrift » et d'élévation de la voix (*key raising*). Ou encore, Liberman et Pierrehumbert (1984) proposent de distinguer le registre (*overall pitch range*) de la déclinaison (*declination*) bien qu'ils utilisent quasiment le même terme pour illustrer cette dernière : « the *range* of f_0 values [being] narrower

and lower at the end of a phrase than at the beginning » . Chez Di Cristo (2005), c'est le terme de « composante orthogonale » qui regroupe l'ensemble de ces effets : la composante orthogonale à la chaîne des tons relève en effet, dans cette conception, des effets d'abaissement (*downtrends*), tels que la déclinaison (*declination*), la catathèse (*downstep*) et l'abaissement final (*final lowering*) et des effets de registre (*pitch range*), i.e. des modifications du niveau de la hauteur tonale (*register level*) et de la gamme couverte par les variations de la hauteur (*span*) ».

Ladd (1996) proposera donc de définir plus explicitement ces phénomènes et la relation qu'ils entretiennent par la spécification de leur caractère intrinsèque vs. extrinsèque et des fonctions qu'ils revêtent. Le registre, synonyme de « composante orthogonale », est décrit en termes de points de référence (valeurs extrêmes mélodiques), spécifiques au locuteur. Facteur extrinsèque, il renvoie aux modifications d'un espace tonal, ou échelle verticale, dans lequel les facteurs intrinsèques que sont la hauteur relative des cibles tonales sont déterminés. L'auteur distingue au sein de ces modifications, les changements catégoriels locaux, ie. effets de catathèse (*downstep*) qui revêtent de fonctions linguistiques et les changements globaux, ie. élévation/abaissement de la voix (*raising/lowering of the voice*) qui revêtent de fonctions para-linguistiques et extra-linguistiques. Il propose ainsi une classification sous trois dimensions des effets de registre (Ladd, 1996, p271) :

Intrinsec	Extrinsec	
Hvs. M	downstep	Raising the voice
linguistic		Para- or extra-linguistic

FIGURE 4 – Classification des variations de registre telle que proposée par Ladd (1996).

Ce tableau synoptique mériterait d'ailleurs d'être complété par la spécification de l'empan temporel de ces phénomènes, tel que décrit par l'auteur, et par les différentes possibilités de modifications des effets extrinsèques para- et extra-linguistiques. Ci-après, donc :

Intrinsec	Extrinsec	
Hvs. M	downstep	Raising/lowering the voice Expanding/narrowing the voice
linguistic		Para- or extra-linguistic
Categorical local shifts		Overall/global modifications

FIGURE 5 – Classification revisitée des variations de registre.

Quelle conceptualisation retenir de l'ensemble de ces propositions ? Si l'on se penche sur les

définitions, plus ou moins explicites, proposées par les auteurs, on s'aperçoit, en fait, que ces différents phénomènes font tous référence aux variations de l'espace tonal dans lequel s'échelonne les cibles tonales. Et la considération d'un espace tonal donne ainsi lieu à de nombreux termes, « tone level frame » (Clements, 1979), « grid » (Gårding, 1983), « transform space » (Pierrehumbert & Beckman, 1988) ou encore « register » (Connell & Ladd, 1990). S'appuyant sur ce fait et sur la conception de Ladd (1996), nous proposons donc une définition plus élargie du registre. Le terme est en effet, dans notre conception, envisagé synonyme de composante orthogonale à la chaîne des tons, et défini comme l'espace tonal dans lequel s'échelonne les cibles tonales, un espace qui peut être à la fois élevé/ abaissé, étendu/rétréci, déclinant. Les modifications de cet espace sont ainsi traduites par les termes d'abaissement, de catathèse, d'élévation du registre, de remise à niveau, etc., selon le caractère graduel vs. catégoriel qu'on lui reconnaît, des termes qui, par conséquent, se trouvent englobés sous ceux de registre (cf. représentation schématique de la relation qu'entretiennent ces termes en 6).

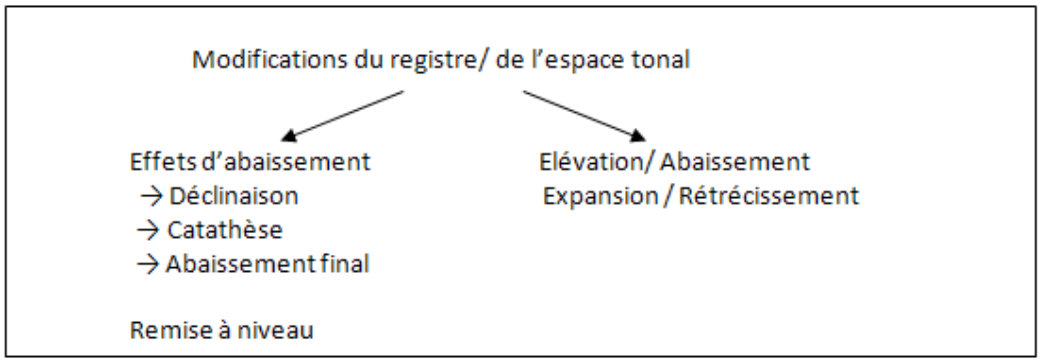


FIGURE 6 – Schématisation des variations de registre.

L'emboitement et les synonymies de ces termes à présent éclaircis, nous proposons de les définir plus explicitement. Nous commençons ainsi par décrire les effets d'abaissement (*downtrends* ; Connell et Ladd (1990)) qui renvoient, dans la littérature anglophone et francophone, aux termes de *declination* / déclinaison, *downstep* / catathèse et *downdrift*. Bien que communément empruntés dans les descriptions phonologiques et phonétiques de l'intonation, ces termes ne sont pas employés de façon systématique et donnent lieu à une certaine nébuleuse qui ne demande qu'à être clarifiée. La difficulté repose clairement sur la nature phonologique vs. phonétique que l'on reconnaît à ces différents effets mais aussi sur la détermination de l'empan temporel qu'on leur accorde. C'est pourquoi nous aborderons ces deux points dans la section 1.4 de ce chapitre.

Le terme de **déclinaison**, à l'origine proposé par Cohen et Hart (1967, p183-4) dans leur étude de l'intonation du néerlandais, renvoie à un abaissement et/ou rétrécissement graduel de

l'espace tonal dans lequel s'échelonne les cibles tonales, au cours de l'énoncé [19]. Parce qu'elle affecte à la fois les sommets et les creux de la f_0, la déclinaison est schématisée à l'aide de lignes déclinantes ou droites de régression, basses et hautes, appelées *statement lines* (Bruce & Gårding, 1978, p224), *Baseline* et *Plateau* (Vaissière, 1983, p55) ou encore *bottom and top of the tonal space* (Ladd, 1992, p330)) (figure 7).

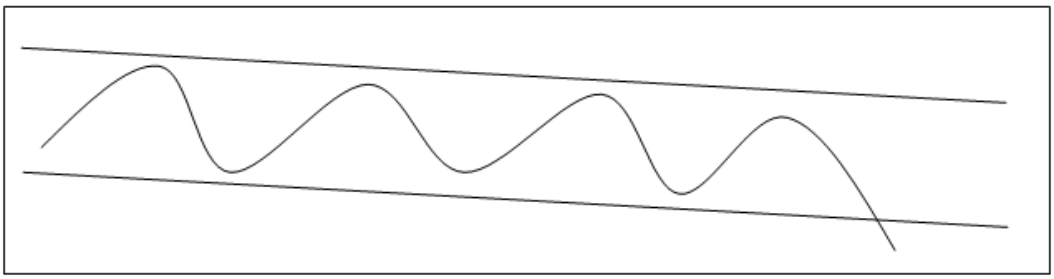


FIGURE 7 – Représentation schématique de la déclinaison.

Cette tendance, universelle, du moins dans les énoncés déclaratifs, est ainsi rapportée dans de nombreuses langues européennes, notamment en français (entre autres, Vaissière (1983)) et en anglais (entre autres, Maeda (1974); Lehiste (1975); Maeda (1976); Pierrehumbert (1980); Sorensen et Cooper (1980); Liberman et Pierrehumbert (1984); Levelt (1989); Hirst (1998)). Caractérisée de temporelle, la déclinaison est attribuée, entre autres, à des contraintes physiologiques telles que la diminution graduelle de la pression sous-glottique ou encore les ajustements laryngés (Maeda, 1976; Honda, 2004). Le terme de déclinaison est donc aujourd'hui strictement réservé à la description des effets ou caractéristiques phonétiques des énoncés, comme fonction linéaire du temps (Pierrehumbert, 1980; Ladd, 1984; Gussenhoven & Rietveld, 1988; Connell & Ladd, 1990; Hirst & Di Cristo, 1998; Connell, 2002; Gussenhoven, 2004; Hirst, 2006).

Comme nous l'avons vu en section 1.2.2, les effets d'abaissements ne sont pas seulement dus à des contraintes physiologiques mais jouent également le rôle d' indice structurel ; ils permettent en effet à la fois la cohésion des unités linguistiques et, associés à une remise à niveau (ou *reset*), signalent un changement de topique ou une frontière syntaxique ou sémantique majeure. Conditionnés par la structure lexicale, tonale, morphosyntaxique d'un énoncé, ils revêtent en effet une fonction démarcative ou distinctive (Laniran & Clements, 2003). Ainsi liés à des contraintes linguistiques, ils sont abordés dans la théorie autosegmentale-métrique, par les phénomènes de **catathèse** ou **downstep** et de **downdrift** (Pierrehumbert, 1980; Liberman & Pierrehumbert, 1984; Beckman & Pierrehumbert, 1986; Clements, 1990; Inkelas & Leben, 1990; Connell & Ladd, 1990; Ladd, 1990, 1992; Van Den Berg et al., 1992; Ladd, 1996;

19. Nous reviendrons sur l'empan temporel de la déclinaison en 1.4.

Hirst, 1998; Truckenbrodt, 2002; Féry & Truckenbrodt, 2005). Si les auteurs s'entendent sur le caractère phonologique de ces deux effets, ce à quoi ils font référence, en revanche, n'est pas clairement établi. Ce qui est clair, c'est qu'ils se distinguent de l'effet de déclinaison [20] dans le sens où ils sont des phénomènes à la fois locaux et itératifs. En effet, contrairement à la déclinaison, leur empan temporel est plus localisé, agissant sur une ou plusieurs séquences de tons hauts. Stewart, Schachter, et Welmers (1964) (cité dans Stewart (1993)) proposent de distinguer le « downstep automatique » du « downstep non automatique ». Le premier renvoie à l'abaissement d'un ton H par rapport au ton H précédent dans une séquence HLH, où le ton bas (L) est rendu responsable de l'abaissement. Le deuxième renvoie également à l'abaissement du second ton H par rapport au ton H précédent mais dans une séquence HH et donc où le ton H n'est pas abaissé sous l'influence d'un ton L. Le terme de downdrift, quant à lui, outre le fait qu'il a pu être utilisé comme équivalent du terme de déclinaison (Pike, 1945; Pierrehumbert, 1980; Ohala et al., 2004), fait soit référence au downstep automatique que nous venons de décrire (Hombert, 1974; Inkelas & Leben, 1990; Snider & Hulst, 1992) soit à l'abaissement progressif d'une chaîne de tons de même nature (Laniran & Clements, 2003), i.e. abaissement des tons hauts uniquement ou abaissement des tons bas uniquement. Pour notre part, nous nous rangeons aux définitions apportées par, entre autres, Ladd (1984), Hirst (1987), Inkelas et Leben (1990), Connell et Ladd (1990), Snider et Hulst (1992), pour lesquelles le downstep est uniquement utilisé pour référer à l'abaissement itératif de tons hauts successifs sans qu'un ton bas intermédiaire soit responsable de l'abaissement alors que le downdrift correspond à l'abaissement itératif de tons hauts successifs sous l'influence d'un ton bas intermédiaire. Hirst (1987) propose aussi la distinction de ces deux phénomènes par l'utilisation des termes d'abaissement dictinctif (équivalent à *downstep*) et d'abaissement conditionné (équivalent à *downdrift*). Résulte, par exemple, la représentation schématisée donnée en 8 pour décrire l'effet de downdrift [21].

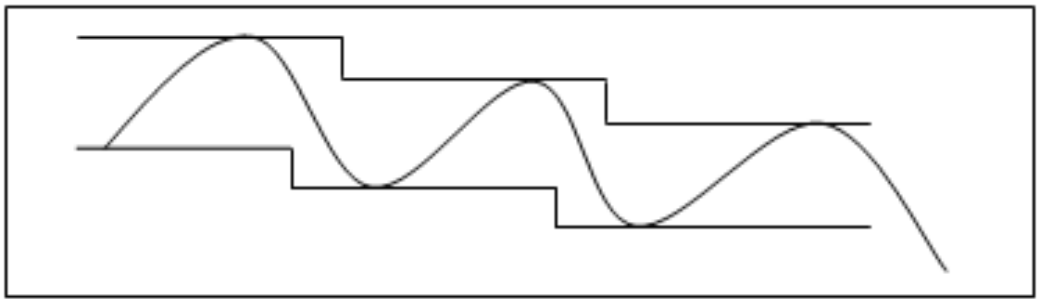


FIGURE 8 – Représentation schématique du downdrift.

20. Il est cependant à noter que le terme de downdrift a été utilisé comme synonyme de déclinaison, notamment chez Pike (1945), Pierrehumbert (1980) et Ohala, Dunn, et Sprouse (2004).

21. Pour plus de précisions sur les effets de déclinaison, downdrift, downstep, cf. les ouvrages éclairants de Ladd (1984), Connell et Ladd (1990) et Connell (2002).

Les effets d'abaissement concernent également le phénomène d'**abaissement final** ou **final lowering** qui renvoie communément à l'abaissement localisé de la hauteur du registre en fin d'énoncés déclaratifs ou encore avant un tour de parole (Liberman & Pierrehumbert, 1984; Vaissière, sous presse).

Les effets d'abaissement, nous l'avons dit, sont également associés au phénomène de **remise à niveau** ou **réhaussement** du registre (Vaissière, sous presse). Dans la littérature anglophone, les auteurs distinguent deux types de remise à niveau : l'effet d'**upstep** et celui de **reset**. Alors que l'upstep renvoie à un retour à la hauteur initiale du premier pic accentuel du début de l'énoncé, le reset renvoie plutôt à l'interruption de l'effet de downstep par un réhaussement tonal partiel. La remise à niveau totale (*upstep*) se distingue également de la remise à niveau partielle (*reset*) qu'elle opère en fin de domaine alors que la remise à niveau partielle se fait en début de domaine (Truckenbrodt, 2002). Ces effets de remise à niveau, comme nous l'avons mentionné plus haut et en 1.2.2, jouent donc aussi le rôle d'indice structurel ou de rupture et participent à la fonction démarcative, distinctive que revêtent les effets d'abaissement. Le degré de remise à niveau marque ainsi la force/ l'importance de la frontière entre deux constituants (Di Cristo, 2004; Vaissière, sous presse).

Les effets d'abaissements concernent donc les phénomènes de déclinaison/declination, de downdrift, de catathèse/ downstep et d'abaissement final/ final lowering. Nous reviendrons sur leur caractère phonétique vs. phonologique qu'on leur reconnaît en 1.4. Il est également important de noter dans les variations de registre, outre les effets d'abaissement « déclinants », les effets d'abaissement, d'élévation, d'expansion et de rétrécissement de l'espace tonal. Il est d'autant plus nécessaire de distinguer ces effets qu'ils affectent la réalisation phonétique des cibles tonales de façon différente. Alors que les effets d'abaissement que nous venons de décrire sont « déclinants », i.e. affectent la réalisation des cibles tonales « pas à pas », un ton H se faisant plus bas que son précédent et ce jusqu'à une remise à niveau, les effets d'abaissement/ d'élévation, d'expansion et de rétrécissement de l'espace tonal, quant à eux, affectent l'ensemble des cibles tonales de façon similaire, les cibles se trouvant alors abaissées/ élevées selon un même degré (Liberman & Pierrehumbert, 1984; Ladd & Terken, 1995; Shriberg et al., 1996). Nous reprenons en figure 9 la représentation schématique explicite proposée par (Ladd, 1996) [22]. Il apparaît clair dans cette représentation que, dans les langues à tons à trois niveaux H, M et L (haut, médian et bas), l'élévation du registre entraîne une élévation de l'ensemble des tons H, M et L alors qu'une expansion cause une élévation de l'ensemble des tons hauts et un abaissement de l'ensemble des tons bas.

22. cf. également les représentations schématiques dans Clements (1979).

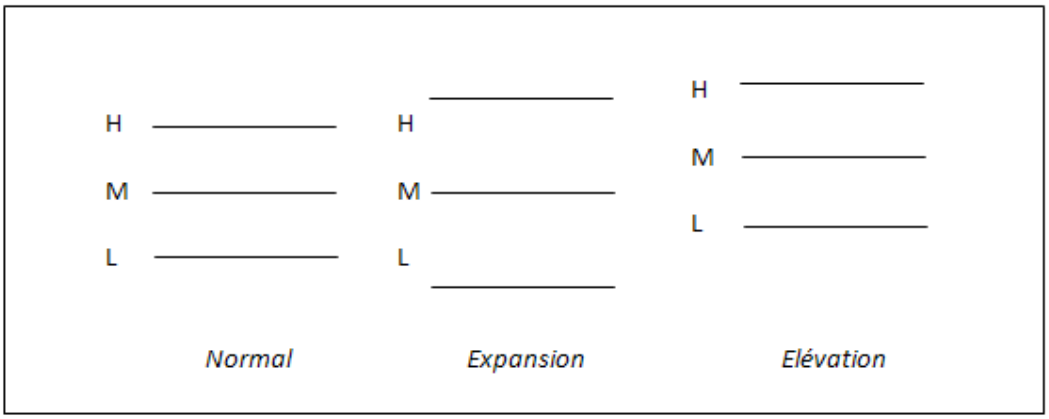


FIGURE 9 – Représentation schématique d'expansion et d'élévation du registre.

Nous proposons de nommer ces modifications de « verticales » afin de les distinguer des effets d'abaissement déclinant [23]. Il est d'ailleurs intéressant de se pencher sur les concepts de registre (*register*) et de plage tonale (*range*) ou d'espace tonal (*tonal space*) et de ligne de base (*base-line*) (Connell & Ladd, 1990; Ladd, 1990, 1992), ou encore sur les concepts de lignes focales (*focal lines*) et de lignes d'affirmation (*statement lines*) (Bruce, 1982), afin d'éclaircir cette distinction. Les premiers auteurs proposent en effet de distinguer l'espace tonal ou registre, dans lequel varient les cibles tonales, de la plage tonale (*range*) dont les limites, idéalement invariantes, viennent à s'élever où s'abaisser à des fins paralinguistiques (e.g. excitation, ennui). La schématisation est la suivante (figure 10).

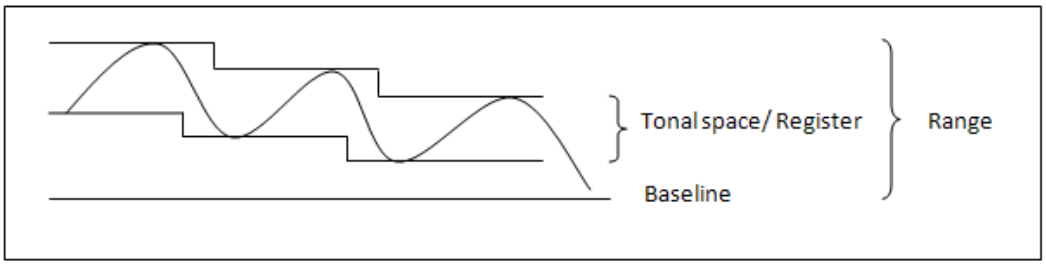


FIGURE 10 – Représentation schématique des modifications déclinantes et verticales du registre telles que proposée par Ladd (1992).

De la même façon [24], Bruce et Gårding (1978), Bruce (1982) distinguent les lignes d'affirmation,

23. On trouvera plutôt dans la littérature le terme d'abaissement graduel que nous réfutons ici afin de pouvoir englober sous ce terme des phénomènes à la fois graduels et catégoriels.

24. Nous rendons équivalentes les conceptions de Ladd (1992) et de Bruce (1982) en termes de représentation

qui délimitent les variations des cibles tonales, des lignes focales, invariantes, qui délimitent les valeurs minimales et maximales atteintes des mots focalisés (figure 11).

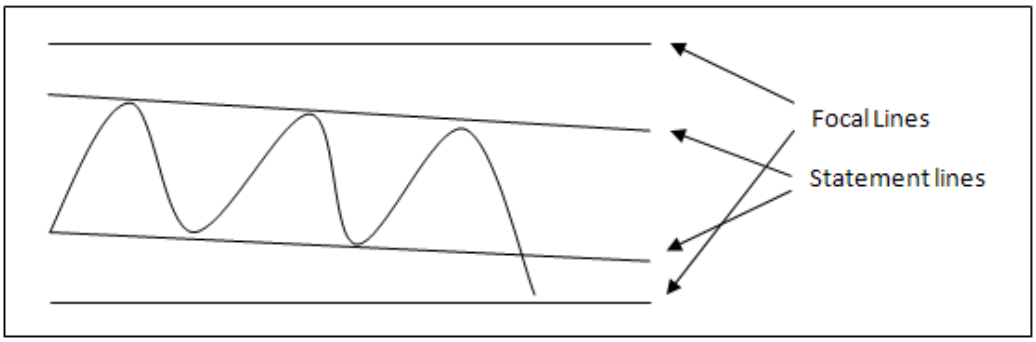


FIGURE 11 – Représentation schématique des modifications déclinantes et verticales du registre telles que proposée par Bruce (1982).

De telles conceptions permettent très clairement de comprendre la distinction que nous faisons des modifications déclinantes et verticales. Alors que l'espace tonal de Ladd (1992) ou les lignes d'affirmation de Bruce (1982) permettent la schématisation des effets d'abaissement déclinant, la plage tonale (*range*) ou les lignes focales permettent, elles, celle des effets verticaux.

Pour notre part, nous représentons les modifications de registre par la schématisation donnée en 12. Les lignes plancher et plafond délimitant la plage tonale restent constantes et servent de limites aux variations verticales [25] ; les lignes du registre délimitent l'espace tonal à partir duquel s'échelonne les cibles tonales. Une telle schématisation permet en effet selon nous de rendre compte simultanément de l'ensemble des possibles formes que peut revêtir l'espace tonal. Contrairement à Ladd (1992) donc, les lignes plancher et plafond délimitant la plage tonale (*range* chez l'auteur) ne s'élèvent ou ne s'abaissent pas à des fins para-linguistiques. Elles sont constantes et servent de limites aux modifications de l'espace tonal. De la même façon, nous considérons l'abaissement final comme un abaissement local du registre, par conséquent nous l'incluons dans notre représentation schématique de l'espace tonal (figure 13).

schématique de l'espace tonal. En aucun cas, nous ne considérons les deux modèles identiques. cf Ladd (1990, p39) et Ladd (1996, p73) pour plus de précisions.

25. Nous verrons au cours du chapitre 2 que cette schématisation, en termes de limites constantes, est validée par nos expériences.

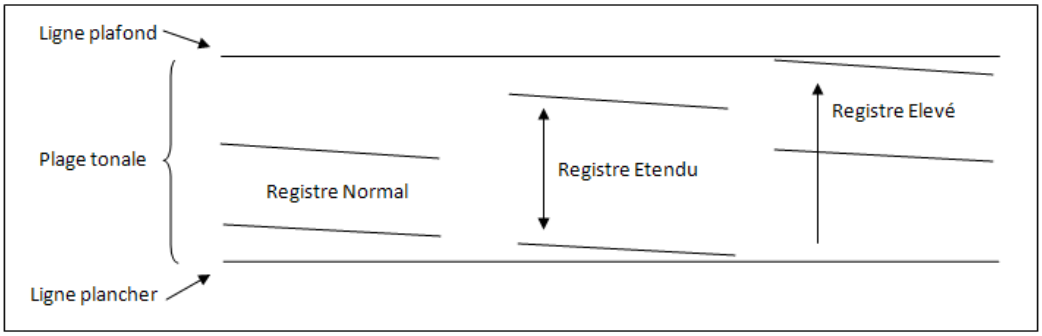


FIGURE 12 – Proposition d'une représentation schématique des modifications déclinantes et verticales du registre.

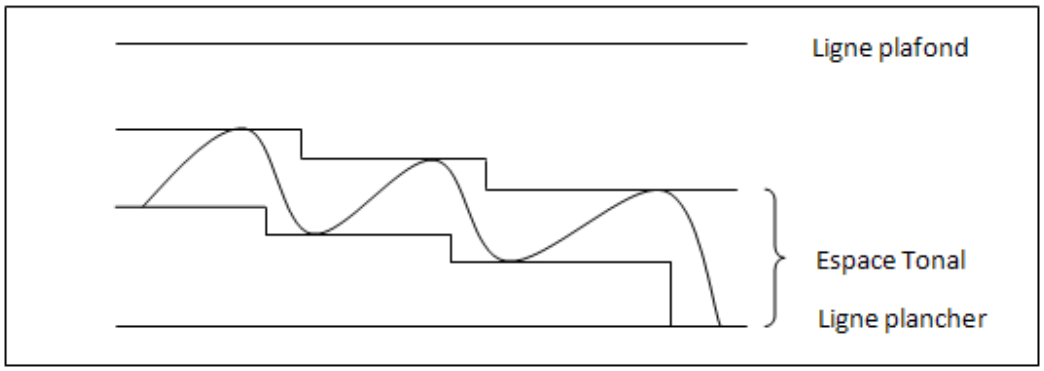


FIGURE 13 – Représentation schématique des modifications déclinantes et verticales du registre.

Nous souhaitons ici apporter une précision : La plage tonale est à différencier de la tessiture, plage plus large qui prend en compte les valeurs extrêmes qu'un locuteur est capable d'émettre (i.e. physiologique). La plage tonale, elle, représente les valeurs extrêmes qu'un locuteur utilise en parole, une bande limite, comme nous l'avons dit, aux variations de registre (figure 14).

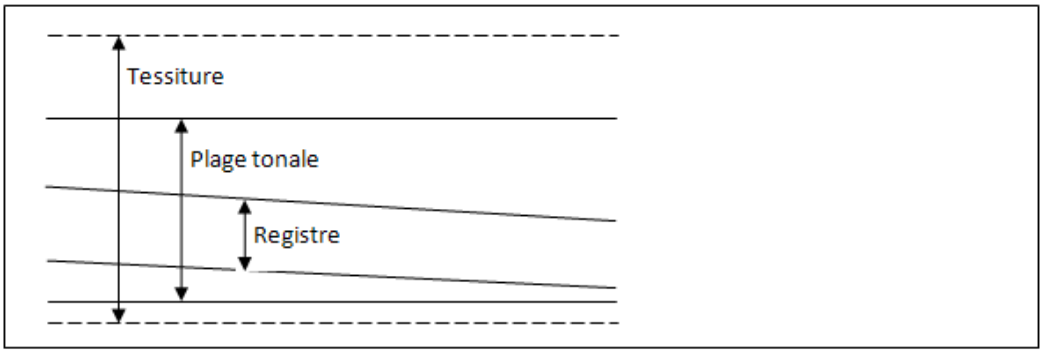


FIGURE 14 – Représentation schématique de la tessiture, de la plage tonale et du registre.

La schématisation que nous avions proposée précédemment peut être alors affinée par l'utilisation des termes décrits dans cette section :

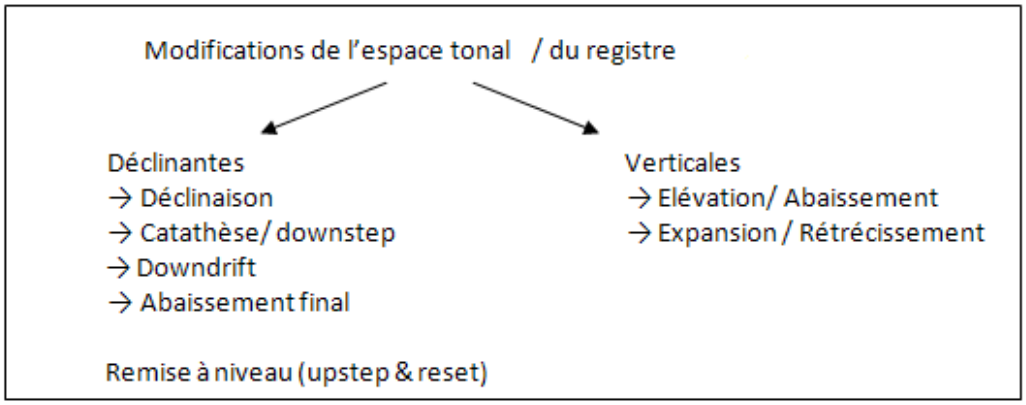


FIGURE 15 – Représentation schématique des variations de registre.

Suite à ces considérations, nous proposons de revisiter la définition que nous avions donnée au terme de registre en 1.1.

Nous rappelons tout d'abord cette définition...

> Le registre est l'espace tonal effectivement utilisé dans un énoncé, une bande de hauteur de sons émis sans difficulté qui, généralement ne prend pas en compte, sans les exclure, les valeurs extrêmes qu'un locuteur est capable d'émettre. On distingue deux dimensions au sein du registre : sa hauteur (hauteur globale de voix perçue par un auditeur) et son étendue (espace tonal délimité par les valeurs maximale et minimale émises par un locuteur).

... Et proposons désormais la suivante :

> Le registre, composante orthogonale à la chaîne des tons, est l'espace effectivement utilisé dans un énoncé, une bande de hauteur de sons émis sans difficulté, qui, généralement, ne prend pas en compte, sans les exclure, les valeurs extrêmes qu'un locuteur est capable d'émettre. Ses modifications, régies par des contraintes de bas niveau (e.g. pression sous-glottique, ajustements laryngés) ou revêtant de fonctions linguistiques, para-linguistiques et extra-linguistiques, affectent ainsi la réalisation des cibles tonales. Les modifications de registre peuvent être à la fois déclinantes, i.e. relever d'effets de déclinaison, de catathèse, d'abaissement final, associés aux phénomènes de remise à niveau (partiel ou non), et verticales, i.e. relever de modifications de hauteur (élévation/ abaissement) et d'étendue (expansion / rétrécissement).

Les variations de registre, à présent clairement établies, nous pouvons nous pencher sur la façon dont elles ont pu être interprétées et intégrées dans les théories phonologiques de l'intonation, et ce afin de déterminer l'empan temporel qui leur est accordé en anglais et en français.

1.4 De l'interprétation et de la représentation phonologique des variations de registre dans les théories de l'intonation : quel(s) empan(s) temporel(s) ?

[26]

Dans les théories phonologiques de l'intonation, les modifications déclinantes de l'espace tonal sont envisagées de deux façons distinctes. Soit elles sont considérées résultantes uniquement du phénomène de déclinaison (Lehiste, 1975; Maeda, 1976; Yule, 1980; Sorensen & Cooper, 1980; Menn & Boyce, 1982; Umeda, 1982; Vaissière, 1983; Collier, 1985; Levelt, 1989; Terken & Collier, 1989)[27], soit elles sont attribuées à la fois au phénomène de déclinaison et au phénomène de downstep (Pierrehumbert, 1980; Ladd, 1984; Liberman & Pierrehumbert, 1984; Beckman & Pierrehumbert, 1986; Ladd, 1988, 1990; Van Den Berg et al., 1992; Truckenbrodt, 2002). Les modifications déclinantes sont donc envisagées soit d'un point de vue global, soit d'un point de vue à la fois global et local, où les modifications globales sont le résultat d'une succession de changements locaux.

26. cf. également les travaux éclairants de Van Den Berg et al. (1992), de Ladd (1996), de Patterson (2000), de Truckenbrodt (2002) et de Féry et Truckenbrodt (2005).

27. Il est à noter que même si dans ces travaux les auteurs attribuent les effets d'abaissement uniquement à la déclinaison, ils ne considèrent pas pour autant cet effet uniquement relevant de contraintes physiologiques, comme le terme de déclinaison, tel que défini plus haut, pourrait le laisser croire. La conception ici révèle plutôt du caractère phonétique qu'ils admettent aux effets d'abaissement.

La déclinaison, nous l'avons dit, est généralement reconnue[28], et ce, dans de nombreuses langues, sur un empan global, de la taille d'un énoncé (*utterance*)[29] (Cohen & Hart, 1967; Sorensen & Cooper, 1980; J. de Pijper, 1983; Liberman & Pierrehumbert, 1984; Collier, 1987; Ladd, 1988; Terken & Collier, 1989; Ladd, 1990; Swerts, Strangert, & Heldner, 1996; Truckenbrodt, 2002). Parce qu'elle est liée à la structure hiérarchique et à l'organisation informationnelle du discours, la déclinaison est cependant difficilement réferable à un domaine phonologique/ syntaxique en particulier. Elle couvre aussi bien des unités de la taille du syntagme que de longs segments de discours. Nous retrouvons ainsi dans la littérature, notamment pour l'anglais et le français, le syntagme, la proposition, la phrase, le groupe de sens ou groupe de souffle, l'unité intonative, le paraton, etc. comme domaines de la déclinaison (Lehiste, 1975; Maeda, 1976; Yule, 1980; Sorensen & Cooper, 1980; Umeda, 1982; Vaissière, 1983; Collier, 1985; Levelt, 1989; Terken & Collier, 1989; Vaissière, 2002). Elle est également établie pour tout domaine dont les constituants sous-jacents partagent un même topique (G. Brown et al., 1980; Yule, 1980; Menn & Boyce, 1982; Umeda, 1982; Vaissière, 1983; Beckman & Pierrehumbert, 1986; Hirschberg & Pierrehumbert, 1986). Les effets de remise à niveau partielle, qui accompagnent la déclinaison, se trouvent donc en amont de chacun de ces constituants, jouant le rôle de marqueur de frontière (Lehiste, 1975; Maeda, 1976; Sorensen & Cooper, 1980; Menn & Boyce, 1982; Vaissière, 1983, 2002).

Depuis Pierrehumbert (1980), les tenants de l'approche auto-segmentale métrique (Liberman & Pierrehumbert, 1984; Beckman & Pierrehumbert, 1986; Ladd, 1988, 1990; Van Den Berg et al., 1992; Truckenbrodt, 2002; Féry & Truckenbrodt, 2005) s'accordent également à attribuer aux phénomènes d'abaissement l'effet de downstep, qui consiste, nous l'avons dit, en l'abaissement des accents mélodiques successifs et qui, lorsque appliqué itérativement, rend compte des phénomènes d'abaissements globaux. Chez Pierrehumbert (1980), l'effet de downstep, en anglais, opère à certains points de la séquence tonale, et ce, au sein d'un même syntagme. Beckman et Pierrehumbert (1986) définissent plus précisément ce domaine d'application et reconnaissent le syntagme intermédiaire (*intermediate phrase*) comme domaine du downstep. Tout récemment, D'Imperio et A. (2009) ont également montré que le downstep agissait au niveau du syntagme intermédiaire en français.

La répétition de l'effet de downstep pour rendre compte des effets d'abaissement plus globaux est cependant envisagée de façon différente au sein de cette théorie. Alors que chez Pierrehumbert (1980), les modifications globales relèvent de la répétition « linéarisée » du

28. Il est cependant à noter que certains auteurs ont pu montrer l'absence du phénomène de déclinaison dans certains contextes. cf. entre autres Umeda (1982) et Lieberman, Katz, Jongman, Zimmerman, et Miller (1985).

29. Shattuck-Hufnagel et Turk (1996) reconnaissent l'énoncé comme domaine maximal de la hiérarchie prosodique ; il y est défini par des pauses silencieuses et correspond plus ou moins à une phrase ou plusieurs phrases syntaxiques.

downstep, elles sont expliquées par l'effet de downstep emboité (*nested downstep* ou *wheels within wheels model of downstep*, Van Den Berg et al. (1992)) chez Ladd (1988) [30], Van Den Berg et al. (1992), Truckenbrodt (2002) et Féry et Truckenbrodt (2005). Le concept de downstep emboité, à l'origine proposé par Ladd (1988), répond à une problématique toujours actuelle, celle du caractère graduel vs. catégoriel que l'on reconnaît aux variations de registre [31]. Alors que Pierrehumbert (1980) réduit la représentation phonologique du registre aux effets de downstep, et ce, selon une représentation uniquement intrinsèque des patrons intonatifs, renvoyant ainsi toute autre modification de l'espace tonal à l'hypothèse de variabilité graduelle libre (*free gradient variability hypothesis* ; Ladd (1994)), Ladd (1988) propose plutôt d'intégrer les effets de registre locaux en tant que composante extrinsèque à la chaîne des tons. L'auteur reconnaît ainsi à la dimension orthogonale à la chaîne des tons, à la fois des distinctions graduelles, et des distinctions catégorielles, qu'il propose de représenter formellement par le downstep emboité (de la même façon que Inkelas et Leben (1990) et Clements (1990) introduisent une nouvelle rangée (*tier*), appelée registre). Sous forme de représentation arborescente (empruntée aux théories métriques), le downstep emboité permet de relier les constituants prosodiques entre eux, par des branchements binaires, où chaque paire de noeuds est étiquetée [h] (*high* ; haut) ou [l] (*low* ; bas). Toute branche droite étiquetée [l] abaisse ainsi le registre d'un niveau. Parce qu'il peut opérer à différents niveaux de la structure, le downstep emboité permet de rendre compte des modifications déclinantes de l'espace tonal sur différents empans, au sein de syntagmes et sur plusieurs syntagmes (e.g. *accentual downstep* et *phrasal downstep* chez Van Den Berg et al. (1992)) et ainsi traduire des différences de registre au cours de larges domaines, e.g. l'énoncé, représenté dans cette conception par l'élément terminal le plus haut (*Highest Terminal Element*). Les remises à niveau du registre, totales et partielles, sont donc également considérées sur des empans à la fois locaux (élévation de la hauteur de l'accent mélodique [32]) et globaux (élévation du registre sur un syntagme qui affecte l'ensemble des segments tonals). Chaque noeud terminal de la représentation métrique représente donc un point potentiel de changement de registre. Une telle représentation rend compte des niveaux d'emboitement de la structure et détermine ainsi, selon le principe *The Deeper, the Steeper* (Féry & Truckenbrodt, 2005), la force ou le degré d'abaissement entre deux constituants soeurs.

De façon similaire, Di Cristo et al. (2004) envisagent les variations de registre orthogonales à la chaîne des tons. Ils proposent que les effets d'abaissement déclinant soient représentés sur une ligne séparée des faits prosodiques, des effets, qui, comme dans la représentation arborescente de Ladd (1988), peuvent opérer à différents niveaux de la structure. Les effets d'abaissement sont ainsi annotés « Dd » (*Downtrends*) et les domaines au niveau desquels ils opèrent sont indiqués entre parenthèses. Un abaissement entre deux unités intonatives, empan maximal que

30. cf. également Ladd (1983), Ladd (1990), Ladd (1992), Ladd (1993), Ladd (1994), Ladd (1996).

31. cf. discussions approfondies dans Beckman et Pierrehumbert (1986), Ladd (1993) et Ladd (1996).

32. Nous rappelons ici que la remise à niveau totale opère en fin de domaine alors que la remise à niveau partielle se fait en début de domaine (Truckenbrodt, 2002).

semblent reconnaître les auteurs, est ainsi marqué : « Dd(UI) ».

Le phénomène d'abaissement final opère, quant à lui, très localement. Il est généralement admis qu'il est réalisé à la fin de constituants déclaratifs assez larges, e.g. de la taille d'un énoncé (Liberman & Pierrehumbert, 1984; Beckman & Pierrehumbert, 1986; Hirschberg & Pierrehumbert, 1986). Cependant, Beckman et Pierrehumbert (1986) expliqueront qu'il est difficile de définir le domaine que conclut l'abaissement final. En effet, parce qu'il est contrôlé par la structure du discours, il ne peut pas être rattaché à un syntagme phonologique en particulier.

Concernant les modifications verticales de l'espace tonal, parce qu'elles sont dites graduelles, et donc, en dehors de toute considération phonologique, les auteurs ne spécifient pas le ou les empans temporels qu'elles couvrent. Elles sont cependant reconnues comme des variations globales ou au moins de la taille d'un syntagme (Liberman & Pierrehumbert, 1984), dans lesquelles les variations déclinantes peuvent se nicher (Ladd, 1996).

Di Cristo et al. (2004), dans leur représentation phonologique des variations verticales, distinguent clairement les variations globales de hauteur et d'étendue du registre des réajustements locaux, traduits par le phénomène de remise à niveau :

> Cette ligne [ReLe] sert à la notation du niveau global du registre (*register level*) pour un domaine donné et à celle des réajustements locaux de ce paramètre. En ce qui concerne le premier cas, nous utilisons les symboles N (*Normal*), Rai (*Raised*) et Low (*Lowered*) pour noter, respectivement le registre moyen de référence du locuteur (ou sa dynamique de base), un rehaussement et un abaissement du registre par rapport à cette référence. Pour le second cas, nous employons le symbole Re (*Reset*) afin de signaler un phénomène abondamment décrit dans la littérature (Wichmann, 2000) qui concerne un retour au niveau de référence du locuteur indiquant l'ouverture d'une nouvelle unité prosodique majeure (Di Cristo et al., 2004, p55).

Il est à noter, à travers cette citation, que, contrairement à Ladd (1996) qui considère que les effets extrinsèques globaux d'élévation et d'expansion du registre doivent être exclus de la représentation phonologique par une procédure de normalisation, les auteurs incluent ces phénomènes dans leur représentation par une notation catégorielle disposée sur deux couches, l'une concernant les variations d'étendue du registre (*Register Span*), l'autre concernant les variations de sa hauteur (*Register Level*).

Bien qu'elles soient reconnues sur un empan global, il est difficile à la lecture de ces différents travaux, d'établir un domaine strict pour lequel les variations verticales opèrent. Cette dif-

ficulté repose tout simplement sur le fait que les variations verticales, comme les variations déclinantes, résultent des choix (conscients ou inconscients), de l'intention du locuteur dans le message qu'il veut faire passer. Le registre d'un locuteur qui exprime par exemple sa colère peut être élevé sur un élément en particulier comme il peut l'être sur plusieurs phrases successives. Quelle que soit l'interprétation que les auteurs font du registre, quelle que soit l'entité qu'ils lui reconnaissent, la définition de l'empan temporel des variations de registre ne peut être aisément établie. Il n'empêche que la prise en compte de l'empan temporel à long terme de ces variations permettrait très certainement une meilleure approche de la réalisation phonétique des cibles tonales. C'est ce que nous chercherons à montrer dans le troisième chapitre de cette thèse dans lequel nous présentons un algorithme de détection automatique des variations de registre.

Outre la problématique de l'empan temporel, une autre difficulté se pose à l'étude du registre, et de ses variations, c'est celle de sa mesure. Il nous a donc semblé nécessaire d'éclairer cette problématique afin de justifier en amont nos choix théoriques dans ce ce travail.

1.5 Mesures du registre et de ses variations

Si l'on synthétise les difficultés rencontrées, dans la littérature, quant à la mesure du registre et de ses variations, trois grandes problématiques sous-jacentes se distinguent : La première concerne les erreurs de détection et de calcul de la fréquence fondamentale, et les contraintes de production « interactives » (Di Cristo, 2004) auxquelles elle est soumise ; la deuxième est relative au choix d'une échelle ou unité de mesure afin de capturer au mieux le registre et ses variations ; enfin, la troisième relève de la mesure même du registre. Doit-elle se faire à partir de la distribution fréquentielle ou à partir de points d'abstraction extraits de cette distribution (e.g. cibles tonales H et L) ?

1.5.1 Erreurs de détection/ de calcul de la f_0 et contraintes de production interactives

La mesure du registre et de ses variations est difficile car elle repose sur une détection de la fréquence fondamentale (désormais annotée f_0) parfois erronée. Si les algorithmes de détection de la f_0 sont aujourd'hui robustes, ils présentent toujours des erreurs de détection et de calcul dans certaines situations : cas d'irrégularités des cycles vocaux (e.g. glottalisation, voix craquée, diplophonie, etc.), de changements abrupts de la f_0, de parole bruitée (i.e. selon les conditions d'enregistrement, e.g. combiné téléphonique), etc (Niemann et al., 1994; Kiessling, Kompe, Niemann, Nöth, & Batliner, 1995; Brøndsted, 1997). La plupart des algorithmes

révèlent en effet deux types d'erreurs [33] : les erreurs de détermination/ décision de voisement (ou erreurs de voisement) et les erreurs de calcul de la f_0 à partir des segments voisés, que nous présentons ci-après, succinctement.

Les erreurs de voisement sont des erreurs de détection de périodicité ou de non périodicité dans les segments de la parole. Il arrive en effet que la f_0 soit calculée à partir d'un signal non voisé, ou inversement ; le détecteur peut omettre le calcul de la f_0 sur des parties du signal voisées, résultant ainsi des valeurs aberrantes. Ghio (2007) explique :

> La décision voisé/non voisé est un compromis délicat : soit le détecteur élimine trop de parties voisées en les considérant comme non voisées et par conséquent, la f_0 n'est pas calculée sur ces parties intéressantes ; soit le détecteur favorise trop le voisement et des parties non voisées sont soumises à un calcul de f_0 qui se solde par des valeurs aberrantes car portant sur du bruit apériodique.

Les erreurs de calculs de la f_0 concernent les « ratés » des algorithmes de détection pour déterminer une valeur fréquentielle. Une des erreurs les plus communément rencontrées est le saut d'octave. Il correspond à des sauts de fréquence situés à l'octave supérieure ou inférieure de la valeur réelle, perçue. Les causes de ces erreurs sont diverses car elles dépendent de la technique utilisée dans le calcul de la f_0 (e.g. méthode temporelle, méthode d'autocorrélation, méthode spectrale d'analyse harmonique ou méthode du cepstre, méthode hybride, etc. [34]). Dans les techniques temporelles, par exemple, les erreurs d'octaves sont le résultat d'une mauvaise sélection des possibles candidats à la périodicité ; l'algorithme donne à la période fondamentale T0 la valeur 2T0 et estime ainsi la valeur f_0 à $f_0/2$ (*halving error*). Dans les techniques spectrales d'analyse harmonique, les erreurs d'octaves sont le résultat d'une mauvaise sélection des candidats à l'intervalle spectral, l'algorithme estimant inversement la valeur f_0 à $2f_0$ (*doubling error*) (Noll, 1968; Rabiner et al., 1976; Dziubinski & Kostek, 2004; Ghio, 2007).

Toute mesure faite à partir de la f_0, notamment lorsqu'elle se fait automatiquement, doit se faire avec la prise en compte des valeurs aberrantes possibles résultant des erreurs de détection et de calcul de la f_0. Mesurer automatiquement le registre revient donc à réfléchir en amont à la façon dont ces effets peuvent être réduits, voire éliminés, afin que la mesure obtenue ne soit pas le reflet d'erreurs de voisement ou de sauts d'octave.

Il est à noter également que la fréquence fondamentale, et ce, quelle que soit la langue étudiée, est « perturbée » par la nature intrinsèque des segments phonémiques et par leur coarticu-

33. cf. correspondant aux quatre types d'erreurs listées dans Rabiner, Cheng, Rosenberg, et McGonegal (1976).

34. c.f. Noll (1968), Rabiner et al. (1976), Hess et al. (1983), Dziubinski et Kostek (2004), Ghio (2007) pour une description de ces différentes méthodes.

lation (House & Fairbanks, 1953; Di Cristo, 1985; Di Cristo & Hirst, 1986; Silverman, 1986; Hanson, 2009). L'influence très localisée de certains segments phonémiques ou « effets intrinsèques » peuvent en effet modifier les valeurs de la f_0. Les voyelles hautes ou fermées ([i] / [u]), par exemple, auront tendance à accroître la f_0 associée à leur production quand les voyelles basses ([a]) auront un effet inverse, résultant en un abaissement important de la f_0 (House & Fairbanks, 1953; Di Cristo, 1985; Di Cristo & Hirst, 1986; Vaissiere, 1988). Di Cristo (1985), synthétisant les différentes études portant sur ces phénomènes, montrera, en français, que les écarts intrinsèques ou rapports entre la valeur moyenne des valeurs hautes et la valeur moyenne des valeurs basses se situent entre 6.2% et 16% (une différence moyenne de 1 à 2 demi-tons ; Di Cristo et Hirst (1986)). La f_0 associée à des obstruentes voisées, en français, serait aussi abaissée ; à contrario, quasiment aucune perturbation n'est rapportée à l'association de sonorantes.

Des effets « co-intrinsèques », ie. résultant de la coarticulation de deux segments phonémiques adjacents, participent également à perturber les valeurs fréquentielles en anglais et en français, et dans d'autres langues d'ailleurs (Di Cristo, 1985; Di Cristo & Hirst, 1986; Silverman, 1986; Hanson, 2009). Le non-voisement des consonnes, notamment, altèrent les caractéristiques prosodiques des voyelles adjacentes, i.e. une voyelle précédée d'une consonne non voisée est généralement marquée par un saut de hauteur mélodique (*pitch skip*) [35] au niveau de l'attaque. Les syllabes CVC ont en effet une valeur fréquentielle plus haute lorsqu'elles commencent par une consonne non-voisée que lorsqu'elles débutent par une consonne voisée. Il est à noter cependant que selon le patron intonatif d'un énoncé, tout effet de saut de hauteur mélodique peut être supprimé (Kohler, 1982). Outre le voisement et bien que secondaire, le mode articulatoire des consonnes perturbent également les valeurs fréquentielles des voyelles adjacentes. La f_0 initiale des voyelles est plus perturbée au contact de fricatives qu'au contact de plosives.

Plusieurs explications ont été proposées aux effets intrinsèques et co-intrinsèques, i.e. physiologiques, aérodynamiques ou encore acoustiques. L'étude de Di Cristo (1985) validera plutôt la théorie de l'attraction linguale (*tongue-pull theory*) aux effets intrinsèques, selon laquelle les variations intrinsèques de la f_0 dépendent du couplage mécanique des muscles de la langue et ceux du larynx. Les effets co-intrinsèques, eux, répondraient plutôt à des théories aérodynamiques qui expliquent ces variations par deux facteurs, i.e. les modifications de la vitesse d'écoulement du débit d'air transglottique et l'activité des effecteurs laryngiens. Si les différentes explications apportées ne font pas consensus, certains facteurs sont cependant unanimement reconnus dans les effets intrinsèques et co-intrinsèques de la f_0. La hauteur du registre global et des contextes tonals plus locaux, le caractère accentuel de la syllabe, la position de la syllabe dans l'énoncé et les caractéristiques propres au locuteur (Di Cristo & Hirst, 1986; Silverman, 1986; Hanson, 2009).

35. cf. (Hanson, 2009) pour une revue de la littérature sur le phénomène de *pitch skip*.

Il est donc important dans toute évaluation acoustique dérivée de la mesure de la f_0 d'être attentif aux perturbations ou valeurs aberrantes résultantes de cette dernière. Une détection automatique « à l'aveugle », i. e. sans se soucier des paramètres à ajuster en amont et sans considérer les limites d'une telle détection, est dangereuse, car les résultats obtenus par une telle méthode peuvent s'avérer erronés.

1.5.2 Echelles de mesure

Mesurer le registre et ses variations revient également à se poser la question d'une unité de mesure. Faut-il opter pour une échelle linéaire ou logarithmique ? Le choix d'une unité repose clairement sur ce que l'on cherche à mesurer (e.g. mouvements mélodiques, proéminence, hauteur de voix, etc.). Dans le cas du registre, nous rappelons que deux dimensions sont à prendre en compte : sa hauteur et son étendue. Il faut donc opter pour une ou plusieurs unités qui permettront de mesurer au mieux ces deux dimensions.

Dans l'étude de la parole, la fréquence fondamentale est généralement exprimée en Hertz (ie. cycles de vibration des cordes vocales). L'échelle de fréquence linéaire est principalement utilisée pour distinguer les voix d'hommes des voix de femmes. Elle apparaît en effet satisfaisante pour mesurer la hauteur du registre et permet ainsi l'étude des différences inter-locuteurs et des variations intra-locuteurs (Patterson & Ladd, 1999; Patterson, 2000). A contrario, mesurer l'étendue du registre à partir d'une échelle linéaire ne permet pas de rendre compte correctement de ces différences et de ces variations. Admettons que le registre d'un locuteur varie entre 100 et 150 Hz, et 150 à 250 Hz pour un autre. Si l'on peut conclure que le premier a un registre plus bas que le deuxième, en aucun cas, il serait juste de dire, et ce, de par la nature du système auditif (Nolan, 2003), que l'étendue de son registre est de moitié celle du deuxième.

Si certains auteurs ont mesuré l'étendue du registre à partir d'une échelle linéaire (Jassem, 1971; C. Williams & Stevens, 1972; Liberman & Pierrehumbert, 1984; Clark, 1999; Möhler & Mayer, 1999; Mayer et al., 2006), d'autres auront préféré une échelle logarithmique (Apple et al., 1979; Graddol, 1986; Collier & Cohen, 1990; Traunmüller & Eriksson, 1995; Henton, 1995; Nicolas & Hirst, 1995; Patterson & Ladd, 1999; Paeschke & Sendlmeier, 2000; Patterson, 2000; Lennes & Anttila, 2002; Nolan, 2003; T. Rietveld & Vermillion, 2003; Xu & Xu, 2005; Hirst, 2007; I. Mennen, Schaeffler, & Docherty, 2008). Ce type d'échelle rend en effet compte de la loi de Weber-Fechner ou de Bouguer-Weber selon laquelle la grandeur d'une sensation perçue est directement proportionnelle au logarithme de la grandeur physique ou intensité d'un stimulus (Bonnet, 1986, p131). Les échelles musicales des demi-tons et des octaves sont par exemple des transformations logarithmiques de l'échelle linéaire. Elles rendent compte de la distance entre deux valeurs tonales en termes d'intervalles musicaux, ie. en termes de demi-tons ou en termes

d'octave. Dans ce type d'échelle donc, chaque intervalle représente une proportion fréquentielle identique (Graddol, 1986; Hermes & Van Gestel, 1991). Les études comparatives de Graddol (1986), Traunmüller et Eriksson (1995), Patterson et Ladd (1999), Patterson (2000) et Nolan (2003) montreront ainsi que lorsqu'on cherche à mesurer l'étendue du registre entre locuteurs (e.g. entre hommes et femmes) mais également les variations de cette étendue (e.g. parole neutre vs. parole « vive »), il est préférable d'utiliser une échelle logarithmique plutôt qu'une échelle linéaire. Les auteurs relèvent d'ailleurs l'importance du choix de l'échelle lorsqu'on mesure l'étendue du registre inter-locuteurs et intra-locuteurs car selon l'échelle choisie, les résultats obtenus peuvent être divergents. Dans l'étude de Graddol (1986), par exemple, alors que le registre des femmes est constaté plus étendu que celui des hommes lorsqu'il est mesuré en Hertz, il est dit plus étroit dès lors qu'il est mesuré en demi-tons. De même, Traunmüller et Eriksson (1995) expliquent que, mesuré en Hertz, le registre d'une femme doit s'étendre plus que celui d'un homme pour atteindre un même degré de « vivacité » alors qu'il lui suffit de s'étendre de la même façon que celui de l'homme lorsque le registre est mesuré en demi-tons.

D'autres échelles « psycho-acoustiques » (i.e. dont le but est de fournir des « pas » qui correspondent à des intervalles perceptifs égaux (Nolan, 2003)) sont également utilisées afin de mesurer les fréquences des sons de la parole : l'échelle Bark, l'échelle ERB et l'échelle de Mel, que nous proposons de présenter ci-après, succinctement.

- L'échelle de fréquence des bandes critiques, ou échelle Bark, établit au sein d'une bande de bruit une largeur de bande « critique » pour laquelle la perception d'un ton est altérée. Elle est dite semi-logarithmique, ie. linéaire en dessous des 500 Hz, logarithmique au-delà (Zwicker, Flottorp, & Stevens, 1957; Zwicker, 1961; Moore, 1989; Traunmüller, 1990; Hermes & Van Gestel, 1991; Nolan, 2003).

- Similaire à l'échelle Bark, l'échelle ERB (*ERB-rate scale* ; *Equivalent Rectangular Bandwith*) est une mesure de la largeur de bande critique. La méthode utilisée est cependant différente. La largeur de bande est en effet mesurée à l'aide de la méthode du bruit à bande réjectée (*notched-noise*) et non par un effet de masquage. Elle est également logarithmique à de hautes fréquences ; en deçà, elle est entre linéaire et logarithmique (Moore, 1989; Traunmüller, 1990; Hermes & Van Gestel, 1991; Nolan, 2003).

- Enfin, l'échelle de Mel (*Mel scale*), proposée par Stevens, Volkman et Newman en 1937, est une échelle basée sur une mesure subjective de la grandeur (*magnitude*) mélodique. Définie arbitrairement à partir de la méthode de dédoublement (i.e. les sons sont déterminés par l'auditeur comme ayant une hauteur de moitié ou double des sons standards), elle a été conçue pour que 1000 mels correspondent à 1000 hz. Elle est linéaire en dessous des 500 Hz, logarithmique au-dessus (Hermes & Van Gestel, 1991; Nolan, 2003).

Dans l'étude du registre, bien que moins utilisée dans la littérature, l'échelle ERB est également reconnue, au même titre que l'échelle des demi-tons, comme une échelle appropriée de mesure de la f_0 (Ladd & Terken, 1995; Hermes & Van Gestel, 1991; Patterson & Ladd, 1999; Nolan, 2003) . Patterson et Ladd (1999) montreront, par exemple, que, bien que les différences d'étendue du registre inter-locuteurs soient correctement révélées lorsque mesurées en Hertz, en demi-tons ou en ERB, les échelles des demi-tons et ERB sont plus fortement corrélées à la perception des auditeurs. Nolan (2003) montrera également que les échelles des demi-tons et ERB sont préférables à l'étude de l'échelonnage des contours intonatifs, les échelles Bark et Mel générant de plus grandes erreurs de réplication[36]. Ces deux dernières semblent donc moins appropriées à la mesure du registre. A notre connaissance, elles n'ont d'ailleurs jamais été utilisées dans l'étude du registre et de ses variations.

Les études menées en perception sur le choix d'une échelle de mesure des variations de registre montrent donc que l'échelle fréquentielle linéaire, exprimée en Hertz, est la plus adéquate pour la mesure de la hauteur du registre alors que les échelles logarithmiques musicales des demi-tons et des octaves et l'échelle semi-logarithmique des ERB sont plus adaptées à la mesure de son étendue.

Nous avons vu que la difficulté de la mesure du registre repose à la fois sur les perturbations de la fréquence fondamentale et sur le choix d'une échelle adaptée. Si la fréquence fondamentale est manifestement le paramètre acoustique utilisé dans l'étude du registre, les mesures qui en sont dérivées, elles, ne relèvent pas d'un consensus. Deux types de mesures, à partir de la fréquence fondamentale, peuvent être cependant distingués : les mesures acoustiques et les mesures linguistiques, auxquelles nous allons nous attacher.

1.5.3 Mesures acoustiques vs. linguistiques

Les mesures acoustiques sont des mesures statistiques faites à partir de la f_0. Dans la littérature, les auteurs utilisent communément la moyenne (Jassem, 1971; Cosmides, 1983; Johns-Lewis, 1986; Graddol, 1986; Bhatt & Léon, 1991; Daly & Zue, 1992; Clark, 1999; Breitenstein et al., 2001; Braun & Kunzel, 2003; T. Rietveld & Vermillion, 2003; I. Mennen et al., 2007) ou la médiane (C. Williams & Stevens, 1972; Graddol, 1986; Carlson, Elenius, & Swerts, 2004) pour mesurer la hauteur du registre. La médiane est généralement considérée plus « pertinente » que la moyenne car elle atténue l'influence perturbatrice des valeurs aberrantes et est une mesure non-paramétrique, indépendente du type d'échelle utilisée. L'étendue du registre a été mesurée de diverses façons, i.e. comme (1) la différence entre la valeur maximale et la valeur minimale de la f_0 (Cosmides, 1983; Clark, 1999), (2) comme la différence entre les 97.5^{me}

36. L'erreur de réplication est une mesure utilisée par l'auteur afin de comparer différentes échelles de mesure. cf. Nolan (2003) pour plus de précisions.

et 2.5^{me} quantiles (couvrant donc 95% de la distribution) (Daly & Zue, 1992), (3) comme la différence entre les 95^{me} et 5^{me} quantiles (90%) (Graddol, 1986), (4) comme la différence entre les 90^{me} et 10^{me} quantiles (80%) (C. Williams & Stevens, 1972; I. Mennen et al., 2007) et (5) en termes d'écart(s) type(s) autour de la moyenne ou de la médiane (Jassem, 1971; Graddol, 1986; Johns-Lewis, 1986; Bhatt & Léon, 1991; Daly & Zue, 1992; Traunmüller & Eriksson, 1994; Clark, 1999; Breitenstein et al., 2001; Braun & Kunzel, 2003; I. Mennen et al., 2007). Nombreux sont les auteurs à avoir eu recours aux mesures acoustiques dans l'étude du registre du fait qu'elles présentent deux avantages non négligeables, celui de l'obtention d'une mesure objective du registre et de ses variations et celui d'une expérimentation « peu coûteuse » en termes de temps, puisqu'elles peuvent être obtenues automatiquement.

Ces mesures distributionnelles à long terme (*long term distributional measures* ; Patterson (2000)) ont cependant été critiquées pour diverses raisons :

(1) Les mesures acoustiques de la f_0, notamment lorsque obtenues automatiquement, prendraient en compte des valeurs aberrantes, causées par une détection et un calcul erronés de la f_0, tels que soulevés en 1.5.1, et ne seraient donc pas des mesures fiables du registre.

(2) Mesurer l'étendue du registre à partir de la distribution de la f_0 autour de la moyenne est tout particulièrement remis en cause car cela suppose une distribution normale de la f_0. Or, Patterson (2000) rapporte de son étude que la distribution peut être normale, i.e. symétrique autour de la moyenne, comme elle peut être positivement ou négativement asymétrique (*skewed*). Les patrons de la f_0 autour de la moyenne sont en effet propres aux locuteurs. Une telle mesure est donc remise en cause car elle est théoriquement incorrecte.

(3) Les mesures acoustiques seraient très faiblement corrélées aux mesures perceptives du registre. C'est ce que révèle l'étude de Patterson (2000) que nous décrivons ci-après. A partir d'un échantillon d'enregistrements, l'auteur évalue 2 mesures acoustiques de hauteur du registre (i.e. la moyenne et la médiane) et 3 mesures acoustiques d'étendue du registre (i.e. 4 écarts types autour de la moyenne, la distance entre les quantiles 95 et 5 et entre les quantiles 90 et 10). Les enregistrements sélectionnés sont des lectures oralisées de deux passages neutres, i.e. qui n'engagent aucune implication émotionnelle de la part du sujet, lus par 32 locuteurs (16 sujets anglais, 16 sujets écossais ; également répartis hommes - femmes). Dans une première expérience de perception, 48 sujets ont pour tâche de décrire les caractéristiques personnelles de ces 32 locuteurs, à partir d'un ensemble d'adjectifs : sûr de soi, tendu, sévère, expressif, profond, faible, irrité, joyeux, effrayé, détendu, énergique et ennuyeux. Les mesures acoustiques sont ensuite corrélées aux mesures perceptives du registre obtenues, i.e. en termes de jugement perceptif des caractéristiques personnelles des locuteurs. L'auteur rapporte alors que les mesures acoustiques, notamment les mesures d'étendue, sont très peu corrélées aux mesures perceptives du registre, en tous cas bien moins que peuvent l'être les mesures lin-

guistiques (détaillées ci-après). Dans une deuxième expérience de perception (ou expérience de réplication), l'auteur cherche à vérifier de tels résultats, notamment quant à la validité des mesures de hauteur qui sont difficilement interprétables en termes d'efficacité dans la première expérience. Il montre ainsi que la moyenne et la médiane sont, dans cette deuxième expérience, corrélées aux mesures perceptives du registre ainsi que la distance entre les 90^{me} et 10^{me} quantiles. Cependant, parce que les mesures acoustiques s'avèrent être corrélées ou non aux jugements perceptifs des auditeurs en fonction de l'expérience, l'auteur en conclut que, contrairement aux mesures linguistiques, elles ne sont pas très fiables pour mesurer le registre.

Certains auteurs préfèreront donc se détourner des mesures acoustiques et opteront pour des mesures linguistiques, i.e. des mesures extraites de points d'inflexion de la f_0, ou points cibles, codés au moyen de symboles discrets et donc considérés linguistiques (Liberman & Pierrehumbert, 1984; Ladd, 1990; Grosz & Hirschberg, 1992; Ladd & Terken, 1995; Nicolas & Hirst, 1995; Shriberg et al., 1996; Patterson & Ladd, 1999; Patterson, 2000; Fagyal, 2002; Portes & Di Cristo, 2003; T. Rietveld & Vermillion, 2003; Biersack & Kempe, 2005; Mayer et al., 2006; Hirst, 2007). Les points cibles situés à des valeurs maximales locales correspondent à des tons hauts, ceux situés à des valeurs minimales locales à des tons bas[37] . Cependant, à partir des cibles tonales, on ne sait quelles sont celles qui permettent au mieux de mesurer la hauteur et l'étendue du registre.

Une des proposition pour la mesure de la hauteur consisterait en la moyenne des tons bas finaux d'un énoncé du fait de leur stabilité (Menn & Boyce, 1982; Liberman & Pierrehumbert, 1984). Cependant, parce que leur dite stabilité a été controversée (Ladd & Terken, 1995; Shriberg et al., 1996) et parce que les tons bas finaux sont « en marge » par rapport aux autres tons bas, Patterson (2000) se pose la question de la pertinence d'une telle mesure. Dans son étude, il propose donc de tester deux mesures de hauteur, i.e. la moyenne des tons bas finaux de phrase (F) et la moyenne des creux post-accentuels (L), qu'il corrèle ensuite aux mesures perceptives du registre. L'étude de perception que nous avons décrite plus haut affirme que la hauteur du registre peut être aussi bien exprimée en termes de F qu'en termes de L, qu'elles soient mesurées en ERB ou en Hertz, bien que F, exprimée en Hertz, apparaisse plus corrélée aux mesures perceptives. Au contraire, T. Rietveld et Vermillion (2003) mettent en évidence, dans leur étude de perception, que les auditeurs n'utilisent pas le ton bas final pour déterminer la hauteur du registre mais plutôt la hauteur relative des tons bas. A partir d'enregistrements de la lecture oralisée d'un locuteur (Anglais Britannique), les auteurs modifient la hauteur du registre par la méthode d'échelonnage de Ladd et Morton (1997) (*the Total Rescaling method*) . Les stimuli sont ensuite présentés à 30 sujets anglais (8 hommes et 22 femmes)

37. Notons qu'une telle correspondance entre points cibles et catégories phonologiques n'est pas clairement établie et qu'elle ne résulte pas d'un consensus. cf. les discussions de Ladd (1996) et Petrone (2008).

qui ont pour tâche de les comparer en fonction de leur hauteur tonale. Le deuxième stimulus doit être validé « plus haut » ou « plus bas » que le premier. Après avoir mesuré les tons bas et les tons bas finaux d'énoncés, les auteurs se sont aperçus que les tons bas sont utilisés par l'auditeur juge pour déterminer la hauteur du registre ; à contrario, les indices finaux ne seraient pas suffisants, les auditeurs ayant besoin de plus d'informations pour évaluer cette dernière. Dans l'algorithme MOMEL-INTSINT (Hirst, 2007), la hauteur du registre est, elle, obtenue en prenant compte de l'ensemble des cibles tonales observées. La hauteur est calculée par optimisation, comme les deux valeurs situées entre 20 Hz en-dessous de la moyenne et 20Hz au-dessus de la moyenne. Le script teste toutes les combinaisons de ce paramètre et choisit le codage optimal des cibles à partir de la valeur pour laquelle la somme des carrés des écarts entre cibles tonales observées et les cibles tonales modélisées est minimale.

L'étendue du registre, i.e. délimitée par une ligne plafond et une ligne plancher, est généralement exprimée comme la différence entre la cible tonale la plus haute et la cible tonale la plus basse du contour ou comme la différence entre la valeur moyenne des tons hauts et la valeur moyenne des tons bas (Hirschberg & Pierrehumbert, 1986; Connell & Ladd, 1990; Ladd, 1990; Mayer et al., 2006) . Afin de valider la pertinence perceptive d'une telle mesure, Patterson (2000) propose de corréler 4 mesures linguistiques aux mesures perceptives du registre, toujours en termes de jugements des caractéristiques personnelles d'un locuteur. Selon le protocole décrit plus haut, l'auteur évalue la combinaison de deux lignes plafond, i.e. la valeur moyenne des tons hauts initiaux de phrase (H) ou la valeur moyenne des pics accentuels non-initiaux (M), et de deux lignes planchers, i.e. la valeur moyenne des tons bas finaux (F) ou la valeur moyenne des creux post-accentuels (L). L'auteur testera donc quelle distance parmi les distances H-F, H-L, M-F, M-L est la plus corrélée aux mesures perceptives du registre. Dans la première expérience de perception, la distance M-L s'avère être la plus corrélée aux mesures perceptives, notamment lorsque exprimée en demi-tons. Dans l'expérience de réplication, la distance M-F est également rapportée comme mesure fiable de l'étendue. Dans l'algorithme MOMEL-INTSINT (Hirst, 2007), l'étendue du registre, comme la hauteur, est obtenue en prenant compte de l'ensemble des cibles tonales observées. L'étendue est calculée par optimisation, comme les deux valeurs situées entre 0.5 octaves et 2.5 octaves. Comme pour la hauteur, le script teste toutes les combinaisons de ce paramètre et choisit le codage optimal des cibles à partir de la valeur pour laquelle la somme des carrés des écarts entre cibles tonales observées et les cibles tonales modélisées est minimale.

Les mesures linguistiques semblent donc présenter un avantage certain, celui d'une forte corrélation avec des mesures perceptives du registre. Cependant, comme les mesures acoustiques, les mesures linguistiques susciteront des critiques :

(1) Mesurer le registre à partir d'un inventaire de catégories tonales peut être inadapté car la distribution de ces catégories peut être différente selon les langues. En effet, I. Mennen et

al. (2008) rapporteront de leur étude que les locuteurs allemands (*Northern Standard German*) utilisent plus communément les creux accentuels alors que les pics accentuels sont plus communs chez les locuteurs anglais (*Southern Standard British English*). D'autres différences de distribution sont également rapportées, elles concernent les pics et les creux accentuels, généralement hauts en allemand, bas en anglais (cf. figure 16 empruntée à I. Mennen et al. (2008)).

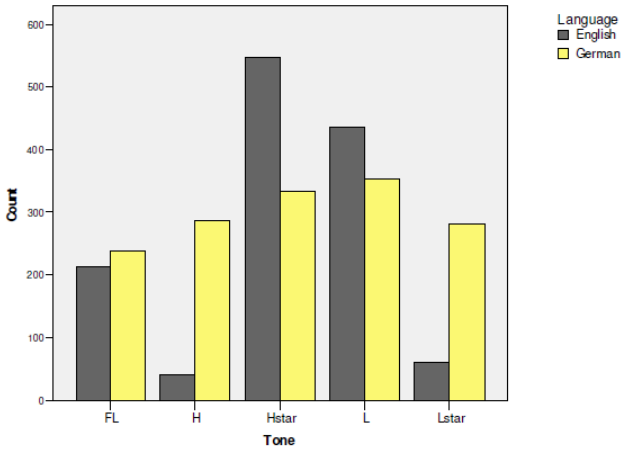


FIGURE 16 – Distribution des cibles mesurées en Allemand Standard et en Anglais Standard (FL=Final Low, H=High tones, L=Low tones, star=stressed syllable).

Sur les 4 mesures linguistiques évaluées, seule la différence entre la moyenne des pics post-accentuels et la moyenne des creux post-accentuels permet de caractériser des différences de registre entre les locuteurs anglais et les locuteurs allemands. Les auteurs concluent donc qu'utiliser des mesures linguistiques pré-déterminées dans l'étude du registre inter-langues n'est pas si évident. Elle requiert au contraire, en amont, l'étude de la distribution tonale de chaque langue, afin de valider une possible comparaison inter-langues à partir de catégories tonales communes. Cette contrainte révèle une deuxième critique :

(2) La mesure du registre en termes de mesures linguistiques dépendrait des présupposés théoriques. En effet, parce que les théories peuvent ne pas admettre le même statut phonologique à certaines cibles tonales, la mesure du registre donnerait lieu à différentes possibilités : soit il serait mesuré en termes de tons bas post-accentuels, soit il serait mesuré en termes de tons bas alignés aux syllabes accentuées et de tons bas alignés aux syllabes inaccentuées, dans le cas où la théorie accorde un statut phonologique différent aux cibles tonales selon le caractère accentuel de la syllabe auxquelles elles sont assignées.

(3) Les mesures linguistiques ont un coût : à l'heure du traitement automatique des langues,

une annotation manuelle par la spécification de cibles tonales peut en effet apparaître trop coûteuse en termes de temps. Patterson (2000) reconnaît d'ailleurs :

> Nevertheless future research should attempt to find properties of speaker's long term distributions of f_0 that approximate the parameters used in the model. It will be great benefit for future research if such properties can be found, as the current method of data collection is very time consuming and labour intensive.
>
> (Cependant, des recherches futures devraient s'attacher à trouver les propriétés des distributions à long terme de la f0 d'un locuteur qui s'approchent des paramètres utilisés dans le modèle. Il serait d'un grand intérêt, pour de futures recherches, si de telles propriétés pouvaient être trouvées, puisque la méthode actuelle de récolte des données est très longue et fastidieuse.)

(4) Par ailleurs, la localisation temporelle des cibles tonales, notamment des cibles tonales hautes, pose toujours difficulté (D'Imperio et al., 2007; Petrone, 2008).

Le choix d'une mesure acoustique ou linguistique ne semble donc pas évident. Il est donc important de considérer les limites de chacune et d'évaluer, par l'apport de solutions, celle qui sera la plus « efficace » dans la mesure du registre.

Si nous rappelons les critiques avancées contre l'utilisation de mesures acoustiques, nous retenons : (1) elles prendraient en considération des erreurs de détection et de calcul de la f_0 ; (2) elles seraient fondées sur le postulat théorique selon lequel la distribution est normale autour de la moyenne, ce qui s'avère être parfois erroné ; (3) elles sont faiblement corrélées aux mesures perceptives du registre. La première problématique pourrait être résolue par l'utilisation de mesures quantiles (e.g. la distance entre les 90^{me} et 10^{me} quantiles), qui sont moins sensibles aux distributions allongées et aux valeurs aberrantes [38]. Les mesures quantiles pourraient être complétées par l'utilisation des coefficients de dissymétrie et d'aplatissement afin de pallier à la difficulté d'une mesure fiable de la f_0 (Patterson, 2000). La problématique de la distribution des données n'est pas, à notre sens, un « argument de poids » qui justifierait le choix de mesures linguistiques. Tout d'abord, dans le cas de données asymétriques, une transformation en log permettrait l'obtention d'une distribution proche d'une loi normale ; de plus, à défaut de la moyenne, la médiane peut être utilisée comme mesure de la hauteur du registre puisqu'elle partage la distribution en deux parties égales, ie. deux parties constituées d'un même nombre d'éléments. Enfin, lorsque les données sont normalisées, la différence entre la valeur maximale et minimale peut être utilisée comme mesure fiable de l'étendue du registre. La différence entre des valeurs quantiles, une fois un intervalle de confiance estimé, ou encore le coefficient de kurtosis qui informe de la disposition des masses de probabilités autour de leur centre, peuvent

38. Nous verrons ce point dans le corps de nos chapitres expérimentaux.

être aussi de possibles mesures. Quant à la problématique de la validité perceptive des mesures acoustiques, si certaines études prônent des mesures linguistiques, d'autres montrent que les mesures acoustiques sont également corrélées aux mesures perceptives du registre. La médiane et la moyenne sont considérées intéressantes dans les études de Bezooijen (1984) [39], de T. Rietveld et Vermillion (2003) et de Grawunder, Bose, Hertha, Trauselt, et Anders (2006) par exemple. D'ailleurs, l'expérience de réplication de Patterson (2000) montre aussi que ces deux mesures sont corrélées aux mesures perceptives du registre. Bezooijen (1984) affirme que l'écart type est un indice acoustique fiable de l'étendue du registre quand Patterson (2000) valide plutôt la distance des 90^{me} et 10^{me} quantiles. Si les mesures acoustiques apparaissent dans ces études moins corrélées aux mesures perceptives du registre et moins stables que les mesures linguistiques, elles ne sont pas pour autant réfutables. D'ailleurs, après avoir confronté le jugement perceptif des auditeurs, et ce, en termes d'origine géographique du locuteur (Anglais vs. Allemand) à diverses mesures linguistiques et à diverses mesures acoustiques, S. F. Mennen I. et Docherty (2008) concluent que les mesures linguistiques ne sont pas plus corrélées aux mesures perceptives du registre que ne le sont les mesures acoustiques. Les différences que l'on trouve dans les études révèlent plutôt, selon nous, de la tâche de perception des auditeurs juges. Dans l'étude de Patterson (2000), la validité perceptive des mesures du registre est évaluée en fonction des caractéristiques personnelles attribuées aux locuteurs, alors qu'elle l'est en fonction de l'origine géographique dans l'étude de S. F. Mennen I. et Docherty (2008) ou encore en fonction du jugement perceptif de hauteur et d'étendue du registre dans l'étude de T. Rietveld et Vermillion (2003), où les auditeurs ont pour tâche de comparer explicitement le registre d'un stimulus par rapport à un autre en termes de « plus haut » ou « plus bas ».

Si nous rappelons à présent les critiques avancées contre l'utilisation de mesures linguistiques, nous retenons qu'il est difficile d'évaluer les différences de registre inter-langues en termes de mesures linguistiques car l'inventaire tonal et la distribution tonale peuvent différer selon les langues, que les mesures linguistiques répondent à des présupposés théoriques et qu'elles sont « coûteuses » en termes de temps. Le coût que demande les mesures linguistiques pourrait être par exemple résolu par des systèmes de modélisation de la f_0 qui permettent la détection de cibles tonales et des systèmes de codage automatiques des catégories tonales à partir de la détermination des points cibles (e.g. MOMEL-INTSINT).

Il est difficile en l'état d'établir lesquelles de ces mesures linguistiques vs. acoustiques sont les plus appropriées à la mesure du registre et de ses variations mais nous serions tentée de dire que le choix devrait être porté par l'objectif de l'étude et que ces mesures devraient être envisagées complémentaires. Nous verrons d'ailleurs, au cours du chapitre 2, que ces mesures sont en fait très similaires et, qu'en termes de registre, elles s'équivalent. Un algorithme comme MOMEL-INTSINT apparaît dans cette problématique comme la réponse aux critiques

39. citée dans T. Rietveld et Vermillion (2003).

principales apportées aux mesures acoustiques et linguistiques, à savoir celles de la fiabilité de la mesure et du coût d'annotation.

Nous avons pu apprécier, dans la section précédente, la difficulté récurrente à l'étude des patrons intonatifs d'une langue, celle de l'interaction qu'ils connaissent avec les variations de registre. Nous comprenons qu'il est nécessaire de prendre en compte les effets à long terme pour une meilleure analyse et représentation de la chaîne des tons. Pourtant, il émane de la littérature une certaine nébuleuse autour de son empan temporel et comprenons que cette problématique difficilement résolvable ait été laissée pour compte dans de nombreux systèmes de l'intonation.

Cette difficulté s'ancre également dans l'étude des variations temporelles. Parce qu'elles sont régies par différents paramètres, situés à différents niveaux (intrinsèque, co-intrinsèque, linguistique, para-linguistique), leur chevauchement et leur interaction rendent leur étude difficile. Notamment, dès lors que l'on s'intéresse à la structure rythmique, temporelle d'un énoncé, se pose le problème de la distinction entre la durée d'un phonème et l'effet des facteurs orthogonaux temporels sur cette dernière, tels que les variations de tempo. Comment en effet distinguer un phonème court dans un tempo ralenti d'un phonème long dans un tempo accéléré ? Cette problématique se doit d'être traitée car, non établie, elle rendrait l'analyse et la modélisation des durées segmentales fragiles. En effet, en synthèse de parole, un modèle de durée se doit d'être capable de prédire des durées relativement proches de celles que l'on peut trouver en parole naturelle, afin d'améliorer le naturel de la parole synthétisée. Pour cela, prendre en compte les différents facteurs qui peuvent influencer la durée segmentale, i.e. un ensemble de variables prédictives, reste primordial. Il est aujourd'hui reconnu que, outre l'entité phonémique et le contexte segmental, la structure et les faits prosodiques influencent la durée observée des segments phonétiques. Pour notre part, nous nous intéressons à la façon dont le tempo influence et interagit avec la durée segmentale. Dans une nouvelle section donc, nous proposons de comprendre les enjeux qui gravitent autour des variations de tempo. Quel définition lui donne t-on ? Comment le mesure t-on ? Quel empan temporel lui reconnaît-on ? Quels sont ces effets sur la chaîne segmentale et supra-segmentale ?

2 Tempo

2.1 Une définition

En phonétique ou en phonologie, de nombreux termes, « tempo », « débit de parole », « vitesse d'élocution », « vitesse d'articulation », « speech tempo », « speaking tempo », « speech rate », « speaking rate », ont été utilisés pour faire référence, en substance, à la vitesse de parole d'un

locuteur.

Très tôt, il est pourtant proposé de définir plus précisément le concept de « vitesse de parole » en distinguant deux de ses composants : la vitesse d'articulation et les pauses (Goldman-Eisler, 1956, 1961). De nombreux auteurs adopteront cette distinction, ce qui résultera en la création prolifique de termes, ceux d'un côté, référents aux deux phénomènes simultanément, ceux, d'un autre côté, excluant de leur conceptualisation les pauses. Ainsi, les auteurs distinguent les termes « tempo » (Eefting & Rietveld, 1989), « speaking rate » (Vaane, 1982; Miller et al., 1984; Gros, Miheli, & Pave, 1999), « speech rate » (Goldman-Eisler, 1961; Butcher, 1981; Eefting, 1988; T. Crystal & House, 1982; Vaane, 1982; Zellner, 1994), « vitesse de tempo » (Grosjean & Deschamps, 1972a), « débit de parole » (Saint-Bonnet & Boe, 1977), comme incluant les pauses, des termes « articulation rate » (Goldman-Eisler, 1961; Butcher, 1981; Miller et al., 1984; Eefting, 1988; Eefting & Rietveld, 1989; Zellner, 1994; Gros et al., 1999; T. Crystal & House, 1982), « phonation time » (Vaane, 1982), « vitesse d'articulation » (Grosjean & Deschamps, 1972a), « durée de phonation » ou « vitesse d'élocution » (Saint-Bonnet & Boe, 1977), comme excluant les pauses.

La vitesse d'articulation est ainsi définie en terme de durée, ie. le temps de phonation utilisé par un locuteur (Trouvain & Grice, 1999), ou en termes d'unités, ie. le nombre d'unités de parole produites par unité de temps (Zellner, 1998). Nous verrons que de nombreuses unités de parole ont été choisies dans la littérature pour mesurer la vitesse d'articulation mais qu'il n'existe, aujourd'hui, toujours pas de franc consensus quant au choix de cette unité.

Les pauses, elles, sont définies en termes de contenu et de durée.

Grosjean et Deschamps (1972a) proposent de distinguer les pauses sonores (décomposables en syllabes allongées et en pauses remplies référant à tout procédé d'hésitation) des pauses non sonores (pauses de respiration, pause de non-respiration, ie. les pauses d'hésitation, les pauses stylistiques ou une combinaison de ces différentes pauses). Plus globalement, même si les pauses ne font pas strictement référence aux mêmes phénomènes dans la littérature, les auteurs s'accordent à distinguer les pauses remplies des pauses non remplies (ou silencieuses). Les pauses remplies correspondent à la perception de sections voisées dans le signal de la parole, ie. des hésitations de types « heu », « hum » causées par des problèmes de planification ; les pauses non remplies correspondent à une portion silencieuse dans le signal de la parole, généralement à une prise de souffle ou jouant le rôle de marqueurs de discours dans le dialogue (Butcher, 1981; Zellner, 1994; Beinum & Donzel, 1996; Dankovicova, 1997; Fon, 1999; Trouvain, 2004).

La pause est également définie en terme de durée. La durée, d'une pause silencieuse notamment est supérieure à un seuil fixé à x ms bien que certains auteurs proposent qu'une pause corresponde plutôt au jugement perceptif d'un auditeur qu'à une durée fixe qui lui serait don-

née. D'un côté, donc, les auteurs se basent seulement sur le signal acoustique et considèrent les pauses dites réelles, ie. acoustiquement visibles. Les durées proposées dans la littérature varient entre 50ms (Lee & Oh, 1999) et 1400ms (Butcher, 1981). Pour certains auteurs, une pause est considérée entre 100 et 150ms (J. Pijper & Sanderman, 1994; Dankovicova, 1997; Tsao & Weismer, 1997; Trouvain & Grice, 1999), pour d'autres aux alentours de 200-250ms (Grosjean & Collins, 1979; Miller et al., 1984), pour d'autres encore aux alentours de 300-350ms (Lass & Deem, 1972; Lass & Clegg, 1974). D'un autre côté, les auteurs se basent sur le jugement perceptif d'un auditeur et distinguent les pauses perceptibles, supérieures à 200ms, des pauses non perceptibles ou acoustiques, inférieures à 200ms (Butcher, 1981). Les pauses perceptibles sont classées en sous catégories. Butcher (1981) distingue les pauses longues dont la durée dépasse 1000ms des pauses courtes de 500ms, Bartkova (1991) les pauses courtes (200ms), des moyennes (entre 200 et 500ms) et des longues (plus de 500ms).

Le choix d'une telle distinction (vitesse d'articulation et pauses) n'est pas le résultat d'un besoin grandissant de définir avec rigueur le concept de « vitesse de parole », mais plutôt celui de la nécessité de comprendre, de ces deux phénomènes, lequel joue un rôle dans l'accélération et le ralentissement de la vitesse de parole, et de déceler, plus précisément, de quelles stratégies usent les locuteurs pour varier leur vitesse de parole.

Les premières études menées sur l'anglais et le français (Goldman-Eisler, 1954, 1956; Goldman-eisler, 1958; Goldman-Eisler, 1961) résumées dans Goldman-Eisler (1968), Lane et Grosjean (1973), Grosjean et Deschamps (1975) et Grosjean (1977)) ont montré que la variation de la vitesse de parole est le résultat de la variation de la durée et du nombre de pauses plutôt que de la vitesse d'articulation. Goldman-Eisler (1956) rapportent en effet pour l'anglais, à partir d'un corpus de parole spontanée (interviews), une moyenne de vitesse d'articulation de 4.4 syllabes par seconde (désormais syll/sec) pour les locuteurs les plus lents et de 5.9 syll/sec pour les locuteurs les plus rapides, alors que cette variation inter-sujets est 5 fois supérieure pour la durée des pauses. Dans différentes études (Goldman-eisler, 1958; Goldman-Eisler, 1961, 1968), Goldman-Eisler démontre que, alors que la durée et le nombre de pauses varient en fonction du type de production, reflétant, selon l'auteur, le processus de sélection et de planification de la parole, la vitesse d'articulation, elle, reste constante. Il semblerait donc que les processus d'abstraction engagés dans la formulation du sens et dans l'encodage de l'information n'ont aucun effet sur les mouvements de la parole. L'auteur conclura, de par la forte variabilité inter-sujets rapportée de son étude, que la vitesse d'articulation est une « constante personnelle d'invariance remarquable » (traduit de l'anglais : « a personality constant of remarkable invariance » ; Goldman-Eisler (1961)). De même, Lane et Grosjean (1973) pour l'anglais et Malecot et al. (1972) et Grosjean et Deschamps (1975) pour le français, rapporteront à partir de corpus de parole lue, que la vitesse d'articulation inter-sujets et intra-sujets, pour passer d'un débit accéléré à un débit normal ou ralenti, ne varie que très peu.

Trouvain (2004), pour l'allemand, conclura que la modélisation du tempo passe d'abord par la modélisation des pauses.

Pourtant, Miller et al. (1984) critiqueront les conclusions hâtives de trois de ces études (Goldman-Eisler, 1956; Malecot et al., 1972; Grosjean & Deschamps, 1975) qui considèrent que seule la variation de la durée et du nombre de pauses joue un rôle essentiel dans la variation de la vitesse globale de la parole. Après avoir réévalué les résultats obtenus dans ces études, Miller et al. (1984) concluent qu'il existe aussi une certaine variabilité de la vitesse d'articulation, avec un coefficient de variation de 11.5% à 25.1% dans l'étude de Goldman-Eisler (1956), de 14.5% (productions lente et rapide) dans l'étude de Malecot et al. (1972) et de 7.5% à 10% dans l'étude de Grosjean et Deschamps (1975). Dans leur expérience notamment, conduite à partir d'un corpus d'interviews en anglais tiré de l'étude de Grosjean et Deschamps (1975), ils notent une forte variabilité de la vitesse d'articulation intra et inter-locuteurs, avec un coefficient de variation de 31%. Ainsi, ils concluent que, bien que la variation de la vitesse d'articulation soit moindre que celle de la variation de la durée et du nombre de pauses, elle mérite également d'être prise en compte puisqu'elle participe considérablement à l'accélération et au ralentissement de la vitesse de parole. Des résultats similaires sont rapportés des études de T. Crystal et House (1982), T. Crystal et House (1990) et Lass et Deem (1972), à partir de corpus de textes anglais lus, où le ralentissement de la vitesse globale de la parole est la conséquence, par ordre décroissant d'importance, de l'insertion de nouvelles pauses, de l'augmentation de la durée des pauses et de l'allongement des segments de la parole. Pour le français, et pour d'autres langues d'ailleurs [40], les auteurs confirment ces mêmes stratégies. La variation de la vitesse de parole inter- et intra-locuteurs, en lecture de texte est le résultat de la combinaison de trois facteurs indépendants : le nombre, la durée des pauses et la vitesse d'articulation par l'allongement et l'ajout de nouvelles syllabes (Saint-Bonnet & Boe, 1977; Bartkova, 1991; Pasdeloup, 2004). Par ailleurs, les pauses varient en termes de durée et de fréquence, en fonction de leur position syntaxique. En effet, on constate, pour les deux langues, à partir de corpus de parole lue, que, à débit accéléré, seules les pauses placées sur une frontière prosodique majeure ou seules les pauses de respiration situées en fin de phrase ou de proposition, ou encore marquées à l'écrit par un point, sont conservées ; à débit ralenti, les pauses sur une frontière prosodique mineure ou les pauses de non-respiration, situées en fin de syntagme ou de mot lexical, qui peuvent être marquées par une virgule à l'écrit ou ne correspondant à aucune forme de ponctuation, sont ajoutées. La suppression et l'ajout de pauses sont donc également fonction de la structure syntaxique et hiérarchique d'un énoncé (Goldman-Eisler, 1968; Lass & Deem, 1972; Saint-Bonnet & Boe, 1977; Grosjean & Collins, 1979; Butcher, 1981; K. A. Fant G. & Nord, 1990; Bartkova, 1991; Zellner, 1998; Trouvain, 2004).

40. cf. pour l'allemand, Butcher (1981), Trouvain et Grice (1999) et Trouvain (2004) ; pour le néerlandais, Eefting (1988)

Les pauses semblent donc jouer un rôle bien plus important que celui de la vitesse d'articulation dans la vitesse globale de la parole, bien que le rôle joué par la vitesse d'articulation ait été démontré. Il est à noter toutefois, que dans les études de Berkovits (1991) pour l'hébreu et Zellner (1998) pour le français, le ralentissement et l'accélération de la vitesse globale de la parole sont plutôt dus à une variation de la vitesse d'articulation qu'à une variation de la fréquence et de la durée des pauses. A partir d'un corpus de phrases lues, Zellner (1998) rapporte, en effet, que le ralentissement de la vitesse de la parole est caractérisé par l'ajout et l'allongement de syllabes, plutôt que par l'allongement des pauses déjà existantes et l'insertion de nouvelles pauses, chacun participant respectivement à 8.8% et 69.96% contre 1.07% et 6,83% à la variation globale de la vitesse de parole. Zellner conclura donc que « l'importance des pauses pour ralentir la parole s'avère être un mécanisme beaucoup moins considérable chez ce locuteur, que ne l'aurait suggéré la littérature ». Notons de cette citation que Zellner a mené son étude sur un seul locuteur et que les auteurs, dans la littérature, ont rapporté une forte variabilité inter-sujets (Goldman-Eisler, 1961; Saint-Bonnet & Boe, 1977; Beinum & Donzel, 1996; Trouvain, 2004). Les variations de la vitesse de parole dans cette étude sont donc le reflet des stratégies utilisées par un locuteur en particulier et ne peuvent être, en aucun cas, considérées comme représentatives des stratégies utilisées par les locuteurs français.

Au regard de ces diverses études, nous nous attacherons dans notre travail à distinguer vitesse d'articulation et pauses dès que nécessaire. Nous utiliserons le terme « tempo » de façon générique (ou encore de « débit de parole »), pour faire référence à la vitesse globale à laquelle parle un locuteur, comme on le ferait en musicologie pour distinguer un tempo andante d'un tempo allegro. Nous pensons en effet que le terme de tempo est approprié de par son origine étymologique, qui renvoie à la notion de temps, et à son emploi, notamment en musicologie, pour décrire le mouvement d'une oeuvre, l'allure d'exécution d'un morceau de musique, indiqué en termes d'unités de temps. Nous adopterons le terme « vitesse ou débit d'élocution » ou encore « vitesse d'articulation » pour la vitesse à laquelle les mouvements de la parole sont exécutés et pour laquelle les pauses remplies ou non sont exclues.

2.2 De la mesure du tempo et de ses variations

L'étude du tempo se heurte à un certain nombre de difficultés. La première relève de la nécessité d'une mesure objective de ses variations mais aussi de la nécessité d'asseoir des critères objectifs analogues aux impressions perceptives. En effet, mesurer le tempo et ses variations de façon objective, c'est aussi considérer une réalité perceptive qu'ont les auditeurs de cet effet. Ainsi, trois « catégories » de tempo se dégagent de la littérature : le tempo objectif et les tempos subjectifs intentionnel et perçu. Le tempo objectif « quantifie un extrait de parole de façon quantitative sur une échelle continue, e.g. comme le nombre de syllabes par seconde »

(Trouvain, 2004). Les tempos subjectifs, intentionnel et perçu, se rapportement respectivement au tempo que le locuteur a cherché à produire et au tempo perçu par l'auditeur. Les tempos subjectifs intentionnel et perçu sont donc décrits en termes qualitatifs, e.g. très rapide, rapide, normal, lent, très lent. La mesure du tempo objectif passe d'abord par le choix d'une unité de mesure, que nous proposons de décrire ci-après.

2.2.1 Choix d'une unité de mesure

Mesurer le tempo de façon objective nécessite de mesurer séparément les éléments qui le composent, à savoir la vitesse d'élocution et les pauses. Dans la littérature, la vitesse d'élocution est basée sur la durée des unités linguistiques (e.g. la durée moyenne des syllabes) ou sur le nombre d'unités produites par unité de temps (e.g. le nombre de syllabes par seconde).

Les études basées sur la durée moyenne des unités linguistiques ou sur le nombre d'unités produites par unité de temps mesurent le débit d'élocution en termes de :

- durée totale du texte (Son & Pols, 1989),

- durée moyenne des phrases (Magen & Blumstein, 1991; Gandour, Tumtavitikul, & Satthamnuwong, 2000),

- durée moyenne des mots (Adams, Weismer, & Kent, 1993; Arvaniti, 1999),

- nombre de mots par minute (Saint-Bonnet & Boe, 1977; Grosjean & Collins, 1979),

- nombre de mots par seconde (Goldman-Eisler, 1961; Lass & Deem, 1972; Fowler & Housum, 1987; Berkovits, 1991; Kessinger & Blumstein, 1997),

- durée moyenne des syllabes (Miller & Baer, 1983; Eefting, 1988; T. Crystal & House, 1990; Beinum & Donzel, 1996; Ladd, Faulkner, Faulkner, & Schepman, 1999; Kirkham, 2002; Igarashi, 2004; Beller et al., 2006),

- nombre de syllabes par minute (Goldman-Eisler, 1956; Tsao & Weismer, 1997),

- nombre de syllabes par seconde (Grosjean & Deschamps, 1972a; Malecot et al., 1972; Lass & Clegg, 1974; Saint-Bonnet & Boe, 1977; Bartkova, 1985; O'Shaughnessy, 1981; Walker et al., 1992; Pfitzinger, Burger, & Heid, 1996; Duez, 1997; Fougeron & Jun, 1998; Pfitzinger, 1998; Trouvain & Grice, 1999; Zellner, 1998; Cutugno & Savy, 1999; Dankovicova, 1999; Fon, 1999; Gros et al., 1999; Janse, Sennema, & Slis, 2000; Dellwo & Wagner, 2003; Hansson, 2003; Verhoeven et al., 2004; Pasdeloup, Espesser, & Faraj, 2006),

- durée moyenne des phonèmes (Fónagy & Magdics, 1960; Osser & Peng, 1964; G. Fant, Kruckenberg, & Nord, 1991; Campbell & Sagisaka, 1992; Verhasselt & Martens, 1996; Batliner et al., 1997; Brøndsted, 1997; Heerden & Barnard, 2006),

- durée moyenne des voyelles (Bartkova, 1991; Carlson, 1991),

- durée moyenne des mores [41] (Kuwabara, 1996; Ohno, Fukumiya, & Fujisaki, 1996; Koiso et al., 1998),

- nombre de phonèmes par seconde (Lobacz, 1976; Walker et al., 1992; Verhasselt & Martens, 1996; Pfitzinger, 1998; Pasdeloup et al., 2006),

- nombre de voyelles par seconde (Pellegrino et al., 2004),

- nombre de mores par seconde (Takamaru, Hiroshige, Araki, & Tochinai, 2000; Hirose & Kawanami, 2002).

Les unités linguistiques à partir desquelles le débit d'élocution est calculé sont donc le texte, la phrase, le mot, la syllabe, le phonème, la voyelle, ou encore la more. Parmi ces unités, les plus utilisées sont le mot, la syllabe, le phonème et la voyelle. La more sert à mesurer le débit d'élocution dans les langues isomoraïques [42]. La question qui se pose à présent est celle du choix d'une unité optimale pour mesurer le débit d'élocution et ses variations.

Pour y répondre, Trouvain (2004) s'appuie sur une liste de cinq critères : le degré de popularité de l'unité, la possibilité d'une étude inter-langues à partir de cette unité, la facilité de sa mesure, la facilité de sa définition et le reflet de sa variance temporelle. Pour notre part, nous pensons que trois critères sont essentiels : la possibilité d'une étude inter-langues, la variance temporelle et la façon dont l'unité rend compte du tempo perçu. Si l'on se base sur ces trois critères, il semble que le mot et la syllabe ne sont pas des mesures adéquates du débit d'élocution. En effet, la structure des mots et des syllabes peuvent affecter la mesure du débit, où les unités les plus complexes sont intrinsèquement marquées par un débit plus lent. Oller (1973) explique, par exemple, que si l'on cherche à comparer les débit d'élocution de locuteurs japonais et états-uniens, la syllabe ne peut être utilisée comme unité de mesure, du fait que les structurations syllabiques de ces deux langues sont très différentes. Les auteurs rappellent les patrons syllabiques les plus fréquents de l'anglais et du japonais : d'un côté, l'anglais révèle des syllabes complexes de type CVS, CCVS, CCCVS, CVC, CVCC, CVCCC et CCVCCCC ; d'un autre côté, les structures syllabiques en japonais sont plus simples, de type V, CV, CVS et C (où C représente une consonne, V une voyelle et S une semi-voyelle). De la même façon,

41. La more est l'unité minimale

42. e.g. le japonais. Dans les langues isomoraïques, les mores successives (unités plus petites que la syllabe) sont dites de durée égale. Une more correspond, par exemple, à une voyelle brève ou à une voyelle brève et la consonne qui la précède.

une mesure du débit en termes de mots peut être affectée par la structure même du mot, structure qui, selon les langues, est plus ou moins complexe (e.g. langues agglutinantes). Oller (1973) suggère donc de mesurer le débit d'élocution en termes de phonèmes ; il rendrait aussi compte, de façon assez fiable, de la variance temporelle (Trouvain, 2004).

En revanche, dans son étude, Kohler (1986) démontre que le phonème, comme la syllabe, n'est pas une mesure adéquate du débit, et ce, pour deux raisons : d'un côté, ces mesures ne peuvent pas exprimer d'importants effets de tempo ; de l'autre, il est très peu probable que l'auditeur se base sur ces unités pour estimer le débit d'élocution d'un locuteur, du fait de la charge cognitive que demanderait une telle tâche.

Une mesure en termes de syllabes ou de phonèmes soulève aussi une autre question, celle de la distinction de la représentation sous-jacente et de la forme actuelle de ces unités. Koreman (2003, 2006) montre en effet que, pour mesurer au mieux le débit d'élocution, il est nécessaire de considérer à la fois la forme canonique (i.e. la représentation sous-jacente des unités) et leur forme actuelle (i.e. la réalisation de ces unités). En d'autres mots, mesurer le débit en termes de taux de phones réalisés uniquement (*realised phones*), i.e. en termes de phones réellement prononcés, n'est pas suffisant. Il faut aussi prendre en compte le taux de phones intentionnels (*intended phones*), i.e. le nombre de segments réalisables à partir de la forme canonique ou encore présents dans l'abstraction phonologique ou linguistique de l'extrait (Pfitzinger et al., 1996).

Au vu de ces observations, il est difficile de déterminer quelle unité permet au mieux de mesurer le débit d'élocution. Trouvain (2004) explique qu'il n'existe pas d'unité optimale et que la sélection d'une unité dépend plutôt des objectifs que se fixe l'étude. Pfitzinger (1998) suggère, pour sa part, que la combinaison du taux de phones et du taux de syllabes permet une meilleure estimation du débit d'élocution. Alors que la corrélation entre le taux de phones et le débit perçu et la corrélation entre le taux de syllabes et le débit perçu sont respectivement de 0.73 et 0.81, la corrélation entre la combinaison linéaire des taux des syllabes et des phones et le débit perçu est, elle, de 0.88. Dans l'étude de Roach (1998), les auditeurs perçoivent l'italien et le français comme ayant un tempo plus rapide que le néerlandais et l'allemand. Ces intuitions sont confirmées par l'étude lorsque le débit d'élocution est mesuré en syllabes mais non lorsqu'il est mesuré en phonèmes. Hirst (2006) propose que cette impression reflète le fait qu'en italien et en français, les syllabes contiennent généralement moins de phonèmes que les syllabes en néerlandais et en allemand. L'auteur suggère aussi que l'asymétrie entre production et perception peut être aussi une explication de l'impression qu'ont les auditeurs du tempo d'une langue.

Campbell (1992) propose, quant à lui, une autre façon de mesurer le débit d'élocution. Selon l'auteur, quantifier le débit en terme du nombre de segments par unités de temps n'est pas

suffisant pour rendre compte de ses variations. Parce que les durées moyennes des phonèmes peuvent être très différentes, le nombre de phonèmes par unité de temps peut être biaisé par la prédominance de phones longs ou courts. De même, parce que la durée moyenne d'une unité dépend du nombre de segments dans cette unité et du caractère accentuel de l'unité, elle ne estimer correctement le débit d'élocution. Ou alors, cela nécessite, en amont, de gommer l'effet de ces facteurs. L'auteur propose donc deux alternatives. La première consiste en une normalisation des durées segmentales et révèle ainsi des différences de longueur [43] ; la seconde compare les durées prédites des phonèmes (i.e. les durées moyennes des phonèmes) et celles des durées observées [44]. Le résidu obtenu permet ainsi de décrire les effets du débit d'élocution sur les changements de durée. Ces deux mesures semblent d'ailleurs refléter l'impression auditive. En effet, les auditeurs appliqueraient des techniques de normalisation pour juger le débit d'élocution d'un locuteur (Campbell & Sagisaka, 1992; Monaghan, 2001). Campbell (1988) observe, par ailleurs, une assez bonne corrélation entre la différence des valeurs prédites et observées, et les changements perçus du débit d'élocution, en lecture oralisée. En revanche, Koiso et al. (1998) n'optent pas pour une normalisation des mores, dans leur étude du débit en japonais, du fait qu'elle ne reflète pas forcément le débit d'élocution perçu.

Pour notre part, parce que la différence entre les valeurs prédites et les valeurs observées répond aux trois critères de sélection que nous nous étions fixée, elle sera utilisée comme mesure du débit d'élocution dans nos chapitres expérimentaux (chapitre 3).

Outre le choix d'une unité optimale de mesure du débit d'élocution, il faut aussi faire celui, dès lors que l'on cherche à mesurer le tempo, de la durée de la pause silencieuse. Nous avons pu voir, dans la définition du tempo que nous donnons plus haut, que pour déterminer une pause, soit les auteurs se basent uniquement sur des critères acoustiques, i.e. toute pause visible, acoustique, au-delà d'un seuil fixé est annotée, soit ils s'appuient plutôt sur le jugement perceptif du locuteur, i.e. toute pause est annotée si elle est perçue. Nous ne reviendrons pas sur ce point que nous avons développé dans la définition du tempo mais il est important de le noter, du fait que la durée des pauses mais aussi leur nombre jouent un rôle important dans la perception du tempo global d'un locuteur (Goldman-Eisler, 1968; Lass & Clegg, 1974; Eefting & Rietveld, 1989; Dellwo, 2006; Koreman, 2006).

Mesurer le tempo en fonction du débit d'élocution et des pauses, de façon objective, n'assure pas cependant la parfaite correspondance entre tempo objectif et tempo perçu. En effet, cela requiert aussi de déterminer le seuil à partir duquel les auditeurs perçoivent des différences de tempo, ou encore à partir duquel ils détectent des changements de tempo intra-locuteurs. Cette variation est par exemple estimée aux alentours des 20% dans les études de Goldman-

43. cf. aussi Campbell et Isard (1991), Pfau, Faltlhauser, et Ruske (2000), Pfitzinger (2002) et Heerden et Barnard (2006).

44. cf. aussi Cedergren et Perreault (1994).

Eisler (1968), Grosjean et Deschamps (1975) et Miller et al. (1984), aux alentours des 10% dans les études de Lehiste (1970). Mesurer le tempo nécessite encore de considérer l'ensemble des facteurs qui participent à la perception du tempo.

2.2.2 Mesure objective vs. subjective

Certaines études ont pu montrer que, outre la vitesse d'élocution et le nombre et la durée des pauses silencieuses, d'autres facteurs participent à la perception des variations de tempo, tels que la structure prosodique, le contour de la f0, le registre, la longueur de l'énoncé, l'hypo- vs. l'hyper-articulation, les disfluences, les élisions, les assimilations, etc. (Goldman-Eisler, 1968; Lehiste, Olive, & Streeter, 1976; Os, 1985; Kohler, 1983, 1986; A. Rietveld & Gussenhoven, 1987; Eefting & Rietveld, 1989; Fon, 1999; Trouvain, 2004; Dellwo, 2006).

En effet, A. Rietveld et Gussenhoven (1987) ont pu observer, en néerlandais, que, bien que marquées par un même débit, les phrases synthétisées dont le contour intonatif est lié (*linked intonation contours*) sont perçues comme plus rapides que les phrases synthétisées dont les contours intonatifs ne sont pas liés (*linked intonation contours*). Cette impression auditive, selon les auteurs, résulte de la complexité phonétique (i.e. le nombre de mouvements descendants et montants) ou phonologique (i.e. le nombre de segments tonals) des phrases. Le contour lié, impliquerait une absence de frontière tonale (*tone domain boundary*) et serait donc perçu comme plus rapide, du fait qu'un contour lié est généralement associé à un tempo rapide. La structure prosodique, i.e. la présence vs. l'absence de frontières d'unités intonatives dans cette étude, affecterait donc la perception du débit.

Outre la structure prosodique, les variations de la f0 et la hauteur du registre influencent aussi la perception du débit (Lehiste et al., 1976; Os, 1985; Kohler, 1986; A. Rietveld & Gussenhoven, 1987; Eefting & Rietveld, 1989; Dellwo, 2006). Les phrases monotones sont en effet perçues plus rapides que les phrases dont la courbe intonative est normale. Lorsque le registre est haut, le débit est également perçu plus rapide. Lehiste et al. (1976) et Kohler (1983) ont d'ailleurs observé une forte corrélation entre la durée segmentale et le contour de la f0. Plus les mouvements des tons sont complexes, plus la durée segmentale augmente.

La relation entre structure sous-jacente et structure de surface participe aussi à l'impression globale du tempo et de ses variations. Par ailleurs, cet effet est étroitement lié à celui de la qualité d'élocution qui intervient dans la perception du tempo, à savoir l'effet d'une parole hypoarticulée, relâchée vs. celui d'une parole hyperarticulée. Selon Koreman (2006), ces stratégies de parole (i.e. une parole hypo- vs. hyper-articulée) se reflètent en effet directement dans le taux de phones réalisés et le taux de phones intentionnels. L'auteur observe, en allemand, une forte corrélation entre les taux de phones intentionnels et réalisés. En effet, lorsque le

taux de phones intentionnels augmente, le nombre de phones supprimés dans la prononciation canonique augmente aussi. Il formule donc l'hypothèse que les auditeurs se basent sur la différence entre phonèmes intentionnels et phonèmes réalisés pour percevoir le débit. L'hypothèse est validée : l'auteur montre en effet que la suppression des segments affecte la perception du tempo. Par ailleurs, un taux bas d'événements de surface (i.e. de phones réalisés) fait que l'auditeur perçoit la parole relâchée comme plus lente que la parole claire, quand les taux de phones intentionnels sont identiques. Ou encore, plus le taux de phones réalisés est élevé dans une parole claire, plus l'auditeur a l'impression d'un débit rapide. La perception du débit dépend donc aussi du taux sous-jacent et du taux de surface.

Les disfluences, telles que les bégaiements, les lapsus, les interruptions, les hésitations, les allongements, les pauses remplies et les rires, jouent aussi un rôle dans le débit global perçu (Goldman-Eisler, 1968; Fon, 1999; Trouvain, 2004; Koreman, 2006). Une parole lente est en effet marquée par plus de bégaiements et d'hésitations ou de pauses remplies, qu'une parole rapide, plutôt marquée, elle, par des lapsus, des allongements et des rires (Koreman, 2006). Ou encore, la longueur de l'unité joue un rôle sur la perception des variations du débit (Eefting & Rietveld, 1989). Kohler (1986) observe notamment en allemand que les phrases longues sont perçues plus lentes que les phrases courtes. La complexité de l'unité joue aussi dans la perception du débit. Hoequist et Kohler (1986) ont par exemple montré qu'une syllabe complexe est perçue plus rapide qu'une syllabe moins complexe.

Au vu de ces observations, on constate que l'effet cumulé de ces divers paramètres participent à l'impression globale du tempo et de ses variations. Mais parce qu'il a été montré que les auditeurs sont capables de percevoir de petites différences de tempo seulement à partir de la vitesse d'élocution, nous pouvons conclure que la vitesse d'élocution, mais aussi les pauses, sont les facteurs les plus importants dans la perception du tempo. Pour notre part, nous nous en tiendrons à ces deux composantes dans notre mesure objective du tempo.

Un autre facteur qui interfère dans la mesure du tempo est l'effet de la fenêtre à partir de laquelle il est mesuré. En effet, Cedergren et Perreault (1994) rapportent qu'une fenêtre temporelle incorrectement choisie, dans la modélisation de la parole, peut mener à une impression de moindre variabilité que la variabilité effective. Il est donc important de déterminer le domaine qui reflète au mieux les variations de tempo.

2.3 De l'empan temporel des variations de tempo

Une première distinction relative à l'empan temporel du tempo est celle d'un effet local vs. global. Soit ces deux termes permettent de distinguer le débit global (i.e. intrinsèque) d'un locuteur et ses variations (Koiso et al., 1998; Dioubina, 2004), soit ils sont employés pour

décrire l'empan temporel des variations de tempo, opérant localement vs. globalement. Ce que les auteurs entendent par local et global, dans cette seconde conception, en revanche ne renvoie pas à un domaine spécifique. Par exemple, alors qu'Ohno et al. (1996) définissent le tempo global comme un ensemble de phrases et le tempo local comme le domaine de la phrase, G. Fant et al. (1991), Campbell et Sagisaka (1992), Mozziconacci (2000) ou encore Trouvain (2004) considèrent la phrase comme le domaine du tempo global, le tempo local étant, lui, défini au sein d'unités plus petites. Du fait que ces termes ne correspondent pas à un domaine en particulier, on voit que divers domaines ont été utilisés dans la littérature pour mesurer le débit d'élocution. Les domaines choisis excluent en effet les pauses, les auteurs s'intéressant notamment dans ces études à la façon dont les variations du débit d'élocution affectent la chaîne segmentale.

Déterminer l'empan temporel des variations du débit nécessite donc de définir un domaine pour lequel le débit reste constant. Les domaines à partir desquels les auteurs ont pu mesurer le débit d'élocution sont généralement basés sur des critères syntaxiques : la phrase, la proposition ou encore le syntagme (groupe verbal, nominal, prépositionnel) (Lobacz, 1976; Eefting, 1991; G. Fant et al., 1991; Campbell & Sagisaka, 1992; Ohno et al., 1996; Verhasselt & Martens, 1996). Batliner et al. (1997) expliquent en effet que les changements du débit d'élocution se situent à des frontières syntaxiques majeures. Ils seraient 70 fois plus fréquents à ces frontières qu'entre les mots. Cependant, pour Brøndsted (1997), le débit d'élocution ne devrait pas être mesuré au niveau de la phrase du fait des nombreuses variations qui peuvent opérer au sein de ce domaine (e.g. au niveau des groupes accentuels). Par ailleurs, Cedergren et Perreault (1994) suggèrent que la phrase n'est pas appropriée pour des corpus de parole spontanée, pour laquelle les faux départs et les répétitions sont nombreux.

Pour mesurer le débit d'élocution, certains auteurs choisissent aussi l'extrait de parole inter-pausal (*inter-pausal run, temporal phrase*) (Miller et al., 1984; Eefting & Rietveld, 1989; T. Crystal & House, 1990; Walker et al., 1992; Beinum & Donzel, 1996; Tsao & Weismer, 1997). Dans l'étude de Miller et al. (1984), le domaine inter-pausal est compris entre deux pauses de 250 ms au moins, dans celle de Trouvain (2004) entre deux pauses de 100 ms. Il est généralement constitué d'un ou deux syntagmes norminaux, verbaux ou encore prépositionnels. Cependant, parce que le débit n'est pas constant au sein de ce domaine, Dankovicova (1999) conclut que l'extrait inter-pausal ne permet pas de mesurer correctement le débit. Elle montre en revanche que l'unité intonative, qui correspond, dans cette étude, à l'*intonation phrase* de Beckman et Pierrehumbert (1986) permet de décrire les variations de débit en tchèque. Cette unité est aussi utilisée dans les études de A. Rietveld et Gussenhoven (1987), Cedergren et Perreault (1994) ou encore Koreman (2003). On notera toutefois qu'elle n'est pas toujours clairement définie et qu'il est par conséquent difficile d'évaluer une analyse comparative de ces diverses études. Cette unité correspond d'ailleurs parfois au domaine inter-pausal

(e.g. Fougeron et Jun (1998)). En fait, excepté quelques travaux, les auteurs ne se posent pas vraiment la question du domaine des variations du débit d'élocution. Le domaine est défini en amont, puis, le débit d'élocution est calculé.

Au vu de ces observations, il est difficile de déterminer quel domaine permet une meilleure mesure du débit d'élocution. Eefting (1991) suggère que le débit varie à différents niveaux et que par conséquent, plusieurs domaines doivent être étudiés pour décrire les variations de débit.

La difficulté du choix du domaine vient aussi du fait, comme pour les variations de registre, que les variations de tempo dépendent de nombreux facteurs : nous avions en effet soulevé dans le premier chapitre, le fait que, outre les caractéristiques physiologiques (i.e. la vitesse des mouvements des organes de la parole, les dimensions et la forme du tractus vocal, le processus respiratoire, les contraintes neuromusculaires, l'horloge biologique interne, etc. ; (Fónagy & Magdics, 1960; Lobacz, 1976; Tsao & Weismer, 1997)), l'âge, le sexe, l'origine géographique et sociale du locuteur, son état émotionnel, le style de parole qu'il adopte, ou encore la structure hiérarchique et organisationnelle du discours influencent la vitesse d'élocution. Notamment, les intentions du locuteur, i.e. sa volonté de se faire comprendre, son adaptation à la situation de communication, aux besoins environnementaux et au type d'audience, sont marquées par des variations de débit (Nooteboom, 1985; Fowler & Housum, 1987; Eefting, 1991; Campbell & Sagisaka, 1992; Zellner, 1996; Duez, 1997; Takamaru et al., 2000; Dioubina, 2004). Goldman-Eisler (1968) ou encore Fon (1999) suggèrent aussi que les fluctuations de tempo révèlent la charge cognitive impliquée pendant la planification des énoncés. Plus la charge cognitive est importante, plus le débit est lent ; et inversement, moins elle est importante, plus le débit est rapide. Ou encore, la longueur des énoncés influencerait le débit d'élocution, où la vitesse de parole est plus lente pour des phrases courtes que pour des phrases longues (Goldman-Eisler, 1968; Fónagy & Magdics, 1960; Malecot et al., 1972; Beinum & Donzel, 1996; Dankovicova, 1999; Hirose & Kawanami, 2002).

Mesurer le débit est donc complexe du fait que la qualité de sa mesure dépend de l'unité et du domaine que l'on décide de choisir, et de la considération que l'on aura porté aux facteurs qui infuencent à la fois sa réalisation et sa perception.

Un autre phénomène relatif au tempo, que nous n'avons pas encore mentionné, est celui de l'allongement de frontière (*boundary-related lengthening*) (Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992). Parce que les deux phénomènes sont étroitement liés, leur séparation est difficile. Dans certains travaux d'ailleurs, l'allongement de frontière est considéré comme la décélération *localisée* du débit de parole. (Cummins, 1999) le définit comme « un processus graduel qui reflète la décélération globale et la réduction de l'effort articulatoire sur un

certain nombre de syllabes » [45]. Ou encore, Yoon, Cole, et Hasegawa-Johnson (2007) traitent ce phénomène comme « une réduction du débit de parole à la fin du syntagme » [46]. Pourtant, l'allongement de frontière est bien distinct des variations globales de tempo. Comme elles, il est un facteur important de la modélisation segmentale. Ainsi, si l'on veut mesurer correctement le débit d'élocution, outre le choix d'une unité et d'un domaine de mesure, il est important de prendre en compte les effets de frontière sur sa mesure ; et inversement, étudier les effets de frontière nécessite en amont de gommer les effets des variations de tempo. Selon nous, l'étude séparée de ces phénomènes n'est possible que si leur empan temporel est déterminé. Or, l'allongement de frontière, comme les variations du débit d'élocution, pose aussi la difficulté de la détermination de son empan temporel. Si de nombreuses études portent sur le phénomène d'allongement de frontière, les travaux consacrés à la détermination de son empan temporel sont rares. A ce jour, l'empan temporel de l'allongement de frontière n'est pas clairement établi. Or, dans le cadre d'une modélisation « fine » de la durée segmentale, il est nécessaire de comprendre la répartition de ce phénomène, son comportement, son influence sur la durée segmentale, ce qui nécessite, en amont, de définir l'unité ou le domaine à partir duquel il opère. Nous proposons donc de traiter ces points.

2.4 De l'empan temporel de l'allongement de frontière

L'allongement de frontière, i.e. un allongement des segments aux bornes des constituants morphosyntaxiques et/ ou prosodiques, a été abondamment décrit dans la littérature et ce, pour de nombreuses langues [47]. Les langues à quantité vocalique et/ou consonantique, que l'on disait « non touchées » par des effets de frontière, du fait qu'elles reposent déjà sur distinction de longueur phonémique (comme le finnois par exemple ; Lehiste (1965) cité dans Hockey et Fagyal (1998)), montrent aussi l'influence de la frontière de constituants sur les durées des segments de la parole (Hockey & Fagyal, 1998). Vaissière (1983) suggère par ailleurs que ce phénomène est universel.

On rencontre, dans la littérature anglophone et francophone, un certain nombre de termes liés à l'effet de frontière. Si l'on parlait, à l'origine de ces recherches, d'allongement prépausal (*pre-pausal lengthening*), les termes d'allongement de domaine (*domain lengthening*), et

45. notre traduction.

46. notre traduction.

47. Anglais : Oller (1973), Klatt (1975), Lehiste (1976), Beckman et Edwards (1990), Campbell (1992), Wightman et al. (1992), Fougeron et Keating (1997), Cummins (1999), White (2002), Byrd et Saltzman (2003), Turk et Shattuck-Hufnagel (2007) ; Français : Hirst et Di Cristo (1984), Pasdeloup (1990), Vaissière (1991), Jun et Fougeron (2000), Post (2000), Michelas et D'Imperio (soumis) ; Allemand : Kohler (1983) ; Espagnol : Aguilar et Escudero (2009) ; Suédois : Lindblom (1968), Horne, Strangert, et Heldner (1995) ; Néerlandais : (Cambier-Langeveld, 1997) ; Hébreu : Berkovits (1993) ; Hongrois : Hockey et Fagyal (1998) ; Mandarin : Cao (2004), Lin et Fon (2009) ; Koréen et taïwanais : Keating, Cho, Fougeron, et Hsu (1999)

plus exactement d'allongement final (*domain-final lengthening*) ou pré-frontalier (*pre-boundary lengthening*) et de renforcement initial (*initial strengthening*) ou post-frontalier (*post-boundary lengthening*), sont aujourd'hui davantage usités. L'allongement final est en effet décrit comme un allongement des segments en position finale de constituants ou précédant une frontière, et non plus uniquement comme l'allongement des segments précédant une pause ; le renforcement initial est celui des segments en position initiale de constituants ou à droite d'une frontière. Dans la description de ces deux phénomènes, deux approches majeures de l'effet de frontière se dégagent : l'une repose sur une investigation acoustique, l'autre se pose d'un point de vue articulatoire, où les effets d'allongement sont décrits en termes de renforcement articulatoire, ou de réduction de la coarticulation.

L'étude de l'allongement de frontière nécessite de déterminer à la fois l'empan temporel ou le domaine à partir duquel il opère et le locus qu'il effecte au sein de ce domaine. Nous rappelons la distinction de ces deux termes que nous avions proposée en début d'introduction. Le terme de domaine ou d'empan temporel renvoie à l'unité à partir de laquelle un phénomène opère, ici l'allongement de frontière, le locus renvoie à la localisation exacte des effets de ce phénomène au sein du domaine défini. Pour reprendre l'exemple de White (2002), si un allongement de frontière est observé à la fin de syntagmes intonatifs, alors le domaine de l'allongement est le syntagme intonatif ; au sein de ce syntagme, le « morceau de parole » qui est affecté par l'effet d'allongement, e.g. la rime de la dernière syllabe du syntagme, est le locus, la localisation précise de l'allongement. Le domaine et le locus nécessitent donc d'être identifiés séparément.

Les premières études menées sur l'allongement de frontière ont tout d'abord montré qu'une syllabe en fin d'énoncé était généralement plus longue qu'une syllabe positionnée au sein de l'énoncé (entre autres, Oller (1973), Klatt (1975), Klatt (1976), Lehiste et al. (1976), Cooper et Danly (1981)). Se pose alors la possibilité d'autres domaines d'allongement final. Plusieurs constituants définis sur la base de critères syntaxiques et/ ou prosodiques sont ainsi considérés. Outre l'énoncé, le paragraphe, la phrase, la proposition, le syntagme et le mot sont aussi reconnus comme les domaines de l'allongement final (Klatt, 1975; Lehiste, n.d.; Nakatani, O'Connor, & Aston, 1981; Rakerd, Sennett, & Fowler, 1987; T. Crystal & House, 1988; Beckman & Edwards, 1990; Campbell & Isard, 1991; Campbell & Sagisaka, 1992; Campbell, 1992; Erickson, 2000; Turk & Shattuck-Hufnagel, 2000). Plus précisément, en anglais, Klatt (1975) montre qu'un allongement final s'insère entre le syntagme nominal et le syntagme verbal, entre le syntagme nominal et le syntagme prépositionnel, en amont de conjonctions et de proposition enchâssée (*embedded clause*). L'allongement en fin de mot est en revanche controversée. Klatt (1976, p1213) explique :

> Early investigators reported large word-final lengthening effects (Barnwell, 1971 ; Lehiste, 1972 ; Klatt, 1973) but they didn't always control for phrase-final lengthening effects. Word-final lengthening has not always been observed by all inves-

> tigators (Harris & Umeda, 1974) and is probably too small an effect to contribute significantly to the decoding of word boundaries location.
>
> (Très tôt, les chercheurs ont montré des effets importants d'allongement final au niveau du mot (Barnwell, 1971 ; Lehiste, 1972 ; Klatt, 1973) mais ils n'ont pas toujours contrôler les effets d'allongements finaux au niveau du syntagme. L'allongement final au niveau du mot n'est pas unanimement observé (Harris & Umeda, 1974) et peut être un effet trop faible pour contribuer significativement au décodage des localisations de frontières de mots.)

L'allongement rapporté en fin de mot ne serait donc qu'un artefact et révèlerait plutôt un allongement au niveau de constituants supérieurs. Cutler et Butterfield (1990, p208) ont un avis différent et proposent plutôt que, contrairement à l'allongement de syntagme, l'allongement final de mot ne serait pas systématique. Il dépendrait du débit du locuteur, et se ferait plus évident en débit ralenti. C'est d'ailleurs ce que Michelas et D'Imperio (soumis) ont pu remarquer dans leur étude menée sur le français. Alors qu'un effet d'allongement est rapporté pour trois domaines phonologiques (le syntagme accentuel, le syntagme intermédiaire et le syntagme intonatif) à débit normal, seulement deux (le syntagme intermédiaire et intonatif) sont marqués par un allongement à débit rapide. L'effet du débit influence ainsi la restructuration phonologique des constituants prosodiques et le fait que l'allongement de mot ne soit pas révélé dans certaines études (e.g. Turk et Shattuck-Hufnagel (2000)) pourrait être expliqué par cet effet. Or, Beckman et Edwards (1990) montrent un allongement final au niveau du mot, selon trois débits distincts, et ce, même lorsque le mot n'est pas aligné au syntagme intonatif. Le domaine du mot reste donc encore à explorer.

Dans le cadre de la constituance prosodique, d'autres domaines sont également définis. En français, l'allongement final est l'une des marques phonétiques suprasegmentales qui identifie 3 niveaux de constituance : (1) le syntagme accentuel (*accentual phrase*[48]), aussi appelé unité rythmique (*Rhythm unit*[49]), syntagme phonologique (*phonological phrase*[50]), ou encore mot prosodique (*prosodic word*[51]) ; (2) le syntagme intermédiaire (*intermediate phrase*[52]) ou segment de syntagme intonatif (*Intonation phrase segment*[53]) ; et (3) le syntagme intonatif (*intonation phrase*[54]) ou unité intonative (*intonation unit*[55])[56]. Plus précisément, Pasdeloup (1990) constate un allongement de moins de 50% sur la syllabe finale du syntagme accentuel

48. Jun et Fougeron (2000).
49. Di Cristo (1998).
50. Post (2000).
51. Martin (1980) et Vaissière (1991).
52. Jun et Fougeron (2000) et Michelas et D'Imperio (soumis).
53. Di Cristo et Hirst (1996).
54. Jun et Fougeron (2000) et Post (2000).
55. Di Cristo et Hirst (1996) et Di Cristo (1998).
56. Il est à noter que l'équivalence faite ici des constituants prosodiques n'est pas aussi stricte qu'elle n'y

et un allongement de plus de 50% sur la syllabe finale du syntagme intonatif. Ces résultats sont confirmés par l'étude de Michelas et D'Imperio (soumis). Les auteurs montrent en effet un allongement significatif de la voyelle et de la syllabe en fin de syntagme accentuel et de syntagme intonatif, en débit normal. L'allongement de la voyelle est aussi estimé en-deçà des 50% au niveau du syntagme accentuel et au-delà des 50% au niveau du syntagme intonatif. Elles mettent également en évidence un niveau intermédiaire, le syntagme intermédiaire, pour lequel la voyelle finale est allongée aux environs des 50%. Zellner (1996) établit, quant à elle, un allongement de la syllabe entre 0.5 et 1 écart type par rapport à la moyenne, en fin de groupe mineur (Ostendorf & Veilleux, 1994) et un allongement supérieur à 1 écart type par rapport à la moyenne, en fin de groupe majeur (Ostendorf & Veilleux, 1994).

En anglais, les auteurs observent aussi l'effet d'allongement final à la frontière droite du mot prosodique, du syntagme intermédiaire et du syntagme intonatif[57] (Edwards & Beckman, 1987; Beckman & Edwards, 1990; Wightman et al., 1992; Fougeron & Keating, 1997; Cummins, 1999; Byrd, Krivokapić, & Lee, 2006; Hirst et al., 2007; Yoon et al., 2007; Turk & Shattuck-Hufnagel, 2007). Il est à noter que les auteurs n'utilisent pas forcément les termes de cette hiérarchie prosodique (i.e. syntagme accentuel, intermédiaire et intonatif) mais nous pouvons penser, au vu des contextes prosodiques proposés dans ces études, que les résultats peuvent être commentés sous l'étiquette de ces trois constituants. Hirst et al. (2007) rapportent aussi un allongement du segment final au niveau de l'unité rythmique étroite (*narrow rhythm unit* ; Jassem (1952)), en anglais britannique.

Il est intéressant de noter que le degré d'allongement, en anglais et en français, et pour d'autres langues d'ailleurs, reflète le niveau du constituant dans la hiérarchie prosodique. Plus le constituant est élevé dans la hiérarchie, plus le degré d'allongement de frontière est important (Ladd & Campbell, 1991; Vaissière, 1991; Wightman et al., 1992; Di Cristo & Hirst, 1996; Cambier-Langeveld, 1997; Fougeron & Keating, 1997; Swerts, 1997; Cambier-Langeveld, 2000; Jun & Fougeron, 2000; Gee & Grosjean, 2002; Byrd & Saltzman, 2003; White, 2002; Yoon et al., 2007). L'allongement de frontière permet ainsi d'identifier des regroupements syntagmatiques importants et d'estimer le degré des démarcations phonologiques, pragmatiques ou morpho-syntaxiques entre les constituants de la hiérarchie prosodique. En français, nous l'avons vu, le syntagme accentuel est marqué par un allongement final de moins de 50%, le syntagme intermédiaire par un allongement final d'environ 50% et le syntagme intonatif par un allongement final de plus de 50% (Pasdeloup, 1990; Michelas & D'Imperio, soumis). En anglais, Yoon et al. (2007) montrent que l'allongement du syntagme accentuel est moins important que celui du syntagme intermédiaire, inférieur aussi à son tour au syntagme intonatif. Si les auteurs

paraît. Le syntagme phonologique est parfois comparé au syntagme intermédiaire (Wightman et al., 1992; Fougeron & Keating, 1997), ou encore le syntagme accentuel de Beckman et Pierrehumbert (1986) est comparé au syntagme mineur de Selkirk (1980). cf. Wightman et al. (1992) pour plus de précisions.

57. tels que définis par Beckman et Pierrehumbert (1986).

s'entendent sur le fait que le degré d'allongement marque le degré, la force de la frontière, le nombre de niveaux qu'ils définissent, lui, n'est pas toujours le même. En anglais, par exemple, Yoon et al. (2007) constatent que l'allongement final permet la distinction de 3 niveaux de constituance prosodique. Wightman et al. (1992) mentionnent que 4 types distincts de frontière peuvent être identifiés sur la base de l'allongement final. En effet, après avoir étudié les effets d'allongement à partir de 7 niveaux de constituance, i.e. (1) le mot orthographique, (2) le mot prosodique, (3) le syntagme accentuel, (4) le syntagme intermédiaire, (5) le syntagme intonatif, (6) le groupement d'unités intonatives et (7) la phrase, les auteurs rapportent un effet significatif d'allongement de frontière final au niveau du mot prosodique, du syntagme accentuel, du syntagme intermédiaire et du syntagme intonatif. Ces résultats vont de pair avec ceux de Ladd et Campbell (1991) qui proposent aussi une hiérarchie à 4 niveaux, définie en fonction de la durée de la syllabe finale de chaque constituant.

Outre les marques phonétiques supra-segmentales, les marques articulatoires de frontière semblent établir parallèlement la hiérarchie des constituants prosodiques. L'allongement de frontière est en effet articulatoirement réalisé par des mouvements mandibulaires très longs et moins rapides, la tension ou raideur (*stiffness*) de ces mouvements étant donc réduite. Pareillement à une réduction du tempo en débit ralenti, l'allongement final consiste donc en une diminution de la tension des gestes articulatoires, un ralentissement de l'articulation en fin de constituants prosodiques (Beckman & Edwards, 1990). Les effets d'allongement de frontière sont donc également révélés au travers d'approches articulatoires (Fougeron & Keating, 1997; Byrd, 2000; Byrd & Saltzman, 2003; Meynadier, 2003; Byrd, Lee, Riggs, & Adams, 2005; Byrd et al., 2006). Fougeron et Keating (1997) montrent par exemple, en français, qu'une diminution progressive du contact linguopalatal maximal de la voyelle /o/ en position finale de constituant et une réduction graduelle de la coarticulation VC, parallèlement à une durée acoustique croissante, permet la distinction de 2 niveaux de constituance prosodique. La durée acoustique à elle-seule permet la distinction de 2 à 4 niveaux de constituance, selon les locuteurs. Byrd et ses collègues (Byrd, 2000; Byrd & Saltzman, 2003; Byrd et al., 2005, 2006) ont également analysé l'articulation en frontière de constituants prosodiques. Ils déterminent trois niveaux prosodiques (absence de frontière, frontière mineure et frontière majeure) selon l'amplitude spatio-temporelle des gestes articulatoires. Byrd (2000), par exemple, rapporte un allongement graduel du geste vocalique final, dont la tension articulatoire est réduite en fin de constituants. L'allongement de frontière de constituants serait marqué à la fois par une durée croissante des segments finaux et par un ralentissement articulatoire localisé sur la voyelle finale des constituants. Les changements spatio-temporels articulatoires augmentent donc graduellement en position finale de constituant, reflétant ainsi la force de la frontière prosodique.

Outre un allongement pré-frontalier, les études révèlent un effet d'allongement post-frontalier,

i.e. à l'initiale de constituants prosodiques. Ces études s'ancrent principalement dans une approche articulatoire, d'où l'utilisation du terme de renforcement initial. Le renforcement initial renvoie ainsi à une augmentation de la durée vocalique ou consonantique, au renforcement des articulations laryngées, relatives à l'aspiration consonantique et à la glottalisation vocalique, à la durée croissante du VOT (*voice onset time*), à un contact linguopalatal plus important, ou encore à une réduction de la coarticulation, en position initiale de constituant. Plus le poids de la frontière prosodique est important, plus le renforcement de ces effets et la durée acoustique des segments initiaux sont augmentés (Oller, 1973; Pierrehumbert & Talkin, 1992; Fougeron & Keating, 1997; Keating et al., 1999; Byrd, 2000; Turk & Shattuck-Hufnagel, 2000; Fougeron, 2001; White, 2002; Byrd & Saltzman, 2003; Byrd et al., 2005, 2006). Oller (1973) montre, par exemple, que la consonne initiale de mot est incrémentée de 30ms en anglais. Ou encore, Fougeron (2001) rapporte, en français, une augmentation progressive de la durée du segment initial dans la syllabe, le mot prosodique et le syntagme accentuel. Dans le cadre de la phonologie articulatoire (Browman & Goldstein, 1992), il est à noter que les auteurs (Byrd et al., 2006) ne considèrent pas les effets d'allongement de frontière final et initial comme des événement séparés. Ils parlent plutôt d'intervalle temporel (*temporal scope*) où les événements prosodiques, tels que les frontières de constituants, ont un intervalle temporel d'activation similaire aux actions de constriction des différents organes dans le conduit vocal. Les effets d'allongement de frontière sont ainsi implémentés en tant que gestes π (π-*gestures*) ou gestes prosodiques (*prosodic gestures*)[58]. Le geste π pourrait être aussi attiré, dans certaines langues, par la syllabe accentuée. Cela suggèrerait la possibilité d'une interaction entre constituance prosodique et proéminence accentuelle (Turk & Sawusch, 1997; Turk & White, 1999).

Si l'on synthétise à présent les études menées en acoustique et en articulatoire, il est difficle d'établir, en anglais et en français, les domaines de l'allongement initial et final de frontière, du fait de la complexité de la constituance prosodique. Outre cette difficulté, se pose celle de la question du locus de l'allongement de frontière au sein de ces domaines.

Turk et Shattuck-Hufnagel (2007) suggèrent que trois approches majeures ont cherché à déterminer le locus de l'allongement de frontière : l'approche structurelle (*structure-based*), l'approche de contenu (*content-based*) et l'approche hybride (*hybrid view*). A travers une approche structurelle, les auteurs définissent le locus de l'allongement selon la structure linguistique du constituant, e.g. la rime d'une syllabe en position finale de constituant. La théorie structurelle suggère donc un domaine fixe, dans le sens où la région affectée par l'allongement de frontière est invariante, quel que soit le constituant et quel que soit son contenu phonologique. L'approche de contenu envisage au contraire le locus en fonction du contenu phonologique du constituant. Le locus dépend donc de la nature intrinsèque du segment, de la syllabe et du nombre de segments en bordure de constituant. Enfin, l'approche hybride, comme son nom

58. cf. Byrd et Saltzman (2003) pour plus de précisions.

l'indique, se trouve à mi-chemin des approches structurelle et de contenu. Elle suggère que le domaine de l'allongement de frontière est fixe, structurellement déterminé, et que les propriétés phonologiques et phonétiques de la syllabe finale ou initiale déterminent l'ajout éventuel d'un allongement additionnel en dehors du locus défini structurellement.

Les travaux situés dans le cadre d'une approche structurelle montrent ainsi que, en anglais et en français, les locus d'allongement de frontière finale sont les dernières syllabes du constituant (Cummins, 1999), la syllabe finale (Oller, 1973; Klatt, 1975; Pasdeloup, 1990; Vaissière, 1991; Michelas & D'Imperio, soumis), la voyelle finale (Klatt, 1975; Beckman & Edwards, 1990; Erickson, 2000; Post, 2000; Michelas & D'Imperio, soumis; Yoon et al., 2007) et la rime de la syllabe finale (Lindblom, 1968; T. Crystal & House, 1990; Wightman et al., 1992; White, 2002).

Les travaux menés sous le couvert d'une approche hybride révèlent aussi que la rime de la syllabe finale de constituant est le locus de l'allongement de frontière final. Mais en fonction du caractère accentuel (*stressed* vs. *unstressed*) ou du contenu phonétique de la syllabe finale, le locus se trouve « dispersé » sur d'autres segments du constituant. White (2002) spécifie que, par exemple, en anglais britannique, l'allongement de frontière affecte la durée de la rime (noyau et coda) de la syllabe positionnée en fin de mot, en fin de syntagme phonologique et en fin de syntagme intonatif. Il ajoute que, si l'avant-dernière syllabe est accentuée (*stressed*), alors le coda de cette syllabe est également allongé, bien que la syllabe finale soit toujours la plus affectée ; il précise encore que, si la syllabe antépénultième est accentuée, alors le coda de cette syllabe est également allongé, et les syllabes inaccentuées (*unstressed*) qui suivent subissent progressivement l'effet d'allongement jusqu'à la frontière du constituant. De la même façon, Oller (1973) montre que lorsque la syllabe finale est accentuée, seuls le noyau et le coda de cette syllabe sont affectés alors que lorsqu'elle est inaccentuée, ce sont l'attaque, le noyau et le coda qui sont allongés. Ou encore, dans l'étude de Cambier-Langeveld (2000), l'allongement final affecte la dernière syllabe du mot et se propage au niveau de la syllabe accentuée du mot (*main-stress syllable*) si la voyelle de la dernière syllabe est courte ou réduite. Turk et Shattuck-Hufnagel (2007), quant à eux, suggèrent que le locus ne peut être défini à partir d'une approche structurelle, de contenu ou hybride. Ils formulent plutôt l'hypothèse d'un allongement final multiple au vu de la littérature et de leurs résultats. Les auteurs mentionnent, en effet, que, bien que l'allongement de frontière affecte principalement la rime de la syllabe finale, 7 à 18% de l'allongement affecte aussi la rime de la syllabe accentuée (*main-stress syllables*) lorsqu'elle n'est pas en position finale. Les auteurs suggèrent donc que l'allongement final ne peut être expliqué à travers un seul mécanisme, puisque il n'a pas toujours le même locus. C'est pourquoi ils proposent un modèle selon deux mécanismes.

Quant au locus de l'allongement de frontière initial, quelle que soit l'approche, les auteurs s'entendent sur le fait que seul l'attaque de la syllabe initiale de constituant est affecté par

l'effet d'allongement (Oller, 1973; Fougeron & Keating, 1997; Byrd, 2000; Turk & Shattuck-Hufnagel, 2000; Fougeron, 2001; White, 2002; Hirst, 2009). L'allongement initial induit donc une approche structurelle.

Ces différentes études soulèvent d'ailleurs un point intéressant, celui de la répartition de l'effet d'allongement final et de sa progression au sein du locus. Au vu des divers travaux conduits, l'allongement de frontière final semble progressif (Kohler, 1983; Berkovits, 1993; Cambier-Langeveld, 1997, 2000; White, 2002; Turk & Shattuck-Hufnagel, 2007). Plus le segment est proche de la borne droite du constituant, plus il est allongé ; et inversement, plus il s'en éloigne, moins il est touché. Wightman et al. (1992) montrent, par exemple, que l'allongement de frontière est plus fort dans la rime que dans l'attaque de la syllabe finale ; ou encore, Campbell et Isard (1991) conviennent d'un allongement plus important au niveau du coda qu'au niveau de la rime, au sein de la syllabe finale. Pourtant, les travaux de Cambier-Langeveld (1997) et Turk et Shattuck-Hufnagel (2007) révèlent que certains segments, situés entre les locus de l'allongement de frontière, ne sont pas touchés par cet effet. On peut alors se demander si l'allongement de frontière est réellement progressif. Nous répondrons qu'un allongement progressif et une absence d'effet d'allongement sur les segments situés au sein de cette progression, sont tout à fait possibles. La progression a lieu jusqu'à la frontière droite du constituant, bien que certains segments ne soient pas affectés par l'allongement. Dans l'étude de Turk et Shattuck-Hufnagel (2007), même si les segments entre la syllabe accentuée et la rime de la syllabe finale ne sont pas affectés, il n'empêche que la syllabe accentuée est moins touchée par l'effet d'allongement de frontière que la syllabe finale du constituant. Nous pouvons donc conclure que, dès lors que le locus se défait de la rime finale et se propage sur d'autres segments, généralement situés sur la syllabe accentuée, une progression de l'allongement de frontière final et une absence de l'effet d'allongement sur les segments « sandwich », i.e. situés entre la syllabe accentuée et la rime syllabe, peuvent être observés simultanément.

A lecture de ces différents travaux, nous comprenons que les allongements de frontière final et initial opèrent sur différents niveaux de la hiérarchie prosodique. Ils indiquent, par leur quantité, le degré ou la force de la frontière du constituant. L'allongement de frontière est donc à étudier sur différents niveaux de constituance. Au vu de la littérature, le locus qui capture au mieux l'allongement final est la rime de la syllabe finale du constituant, bien que, selon le caractère accentuel et le contenu phonétique de la syllabe finale, ce locus puisse être déplacé sur les syllabes précédentes. Plusieurs interprétations sont alors envisageables : soit le locus est plus large que la rime (de la syllabe accentuée à la syllabe finale du constituant), soit le locus est multiple et, dans ce cas, il ne peut être exprimé que par un modèle basé sur deux mécanimes. Dans tous les cas, cela suggère que certains segments ne sont pas touchés par l'effet d'allongement final.

Si nous avons jusque là traité uniquement la problématique des variations de tempo et de l'al-

longement de frontière, et la façon dont ils peuvent interférer dans l'analyse et la modélisation de la durée segmentale, nous ne réfutons pas pour autant, ni n'amoindrissons, l'effet d'autres facteurs. Nous reconnaissons en effet que d'autres variables prédictives doivent être prises en compte dans la visée d'une amélioration du naturel de la parole de synthèse. Les variables prédictives de l'organisation temporelle de la parole sont d'ailleurs aujourd'hui bien établies. Klatt (1987, p761) [59] proposait déjà un modèle fondé sur onze règles de prédiction de la durée segmentale. Le modèle s'appuyait tout d'abord sur le postulat selon lequel chaque segment phonétique a une durée inhérente, i.e. qui lui est propre (Peterson & Lehiste, 1960; House, 1961; Klatt, 1976; Lehiste, 1976). Etaient ensuite affectés des pourcentages d'allongement et de compression aux durées inhérentes des segments phonétiques en fonction de ces onze règles.

De nombreux facteurs influencent donc la durée segmentale. Nous proposons de les synthétiser ci-après :

(1) la nature intrinsèque du phonème (Peterson & Lehiste, 1960; House, 1961; Klatt, 1976),

(2) le contexte phonétique (Klatt, 1976),

(3) la nature grammaticale du mot (Keller & Zellner, 1996; Turk & Shattuck-Hufnagel, 2000),

(4) le nombre ou l'organisation des segments dans l'(es) unité(s) supérieure(s) (Lindblom, 1968; Liberman, 1975; A. Prince, 1983; Selkirk, 1984; Nespor & Vogel, 1983; White, 2002),

(5) le caractère accentuel ou effet de proéminence - accent lexical, nucléaire, contrastif (Hirst, 1977; T. Crystal & House, 1988, 1990; Keller & Zellner, 1996; Turk & Sawusch, 1997; Cummins, 1999; Turk & White, 1999; Cambier-Langeveld, 2000; Erickson, 2000; White, 2002; Turk & Shattuck-Hufnagel, 2007),

(6) la position du phonème, i.e. sa proximité avec la frontière du constituant (morpho)syntaxique et/ou prosodique (Oller, 1973; Lehiste, 1976; Hirst, 1977; Keller & Zellner, 1996; Cambier-Langeveld, 2000; Erickson, 2000; Turk & Shattuck-Hufnagel, 2000; White, 2002; Turk & Shattuck-Hufnagel, 2007),

(7) le degré de la frontière (Hirst, 1977; Ladd & Campbell, 1991; Vaissière, 1991; Wightman et al., 1992; Di Cristo & Hirst, 1996; Cambier-Langeveld, 1997; Fougeron & Keating, 1997; Swerts, 1997; Cummins, 1999; Cambier-Langeveld, 2000; Jun & Fougeron, 2000; Gee & Grosjean, 2002; Byrd & Saltzman, 2003; White, 2002; Yoon et al., 2007), et

(8) le débit de parole (Goldman-Eisler, 1968; Klatt, 1976; Keller & Zellner, 1996; Campbell, 1992; Cummins, 1999).

59. Klatt propose dans cet article une description complète de son modèle, développé succinctement, entre autres, dans Klatt (1973), Klatt (1975) et Klatt (1976).

Nous comprenons dès lors la difficulté d'une analyse et d'une modélisation de l'organisation temporelle de la parole, puisqu'elles nécessitent la prise en compte d'un nombre important de facteurs. Et la difficulté se fait d'autant plus grande que l'étude d'un facteur en particulier peut être interférée par l'influence d'autres facteurs, ou par l'interaction que ces facteurs entretiennent ; d'où la nécessité d'un « contrôle » ou d'une prise en compte de l'ensemble des facteurs dans l'étude d'une variable prédictive en particulier. Cela n'empêche pas pour autant le choix de certains facteurs « au détriment » d'autres, si leur(s) effet(s) sont suffisants dans a description de l'organisation temporelle de la parole. Campbell (1992) montre, par exemple, que la durée, une fois définie par des règles d'élasticité (*elasticity*) et d'ajustement (*accommodation*) au niveau de la syllabe, peut être prédite à partir de 3 facteurs d'allongement : la proéminence, la frontière et le débit de parole.

3 Conclusion : une synthèse

Nous avons proposé une définition du registre établie à travers plusieurs problématiques sous jacentes à sa définition. Nous avons abordé le caractère relatif des variations du registre, et ainsi de la relativité des faits tonals, en abordant les approches initialisante et normalisante. Dans l'approche initialisante, l'espace tonal est défini par la différence entre une cible tonale initialisante et une cible tonale finale. Cette représentation a cependant été critiquée, car elle néglige tout repère à un espace tonal de référence et ne peut donc décrire avec précision les mouvements mélodiques d'une langue. L'approche normalisante viendra donc répondre à cette problématique en proposant la notion d'espace tonal ou d'échelonnage des cibles tonales.

C'est à partir de cette première conception que nous avons envisagé une définition du registre. Or, dès lors qu'on s'attache à la notion d'espace tonal, on se trouve confronté à de nombreux termes et se pose alors la question de leur emboitement et de leur synonymie. Cette densité de mots satellites nous a amené à définir le registre et ses variations sous les termes de variations déclinantes (qui répondent ainsi aux termes de déclinaison, downdrift ou downstep) et les termes de variations verticales (qui se réfèrent aux termes de modifications de hauteur et d'étendue du registre).

Cette définition nous a amené directement à la problématique de l'empan temporel des variations de registre. La définition de cet empan temporel s'ancre dans la façon dont les théories phonologiques de l'intonation considèrent les modifications de l'espace tonal. En abordant l'interprétation et la représentation phonologique des variations de registre dans les théories de l'intonation, nous avons pu ainsi présenter les divers domaines établis à partir desquels elles ont été décrites. Mais, notre regard sur la littérature révèle que la problématique de l'empan temporel des variations de registre est complexe. C'est donc à partir des conceptions théo-

riques que nous avons pu définir et décrire ce qu'on entend par « registre » et par « empan temporel de ses variations ».

Par ailleurs, nous avons soulevé la difficulté de sa mesure. Nous avons dégagé 3 grandes problématiques sous jacentes, à savoir (1) les erreurs de détection et de calcul de la f_0 et les contraintes de production interactives auxquelles elle est soumise, (2) le choix d'une échelle ou d'une unité de mesure et enfin (3) la mesure même du registre.

Les erreurs relatives à la f_0 sont les erreurs de voisement ou erreurs de détection de périodicité qui font que le calcul de la f_0, sur certaines parties du signal, résulte en des valeurs aberrantes. Nous avons aussi expliqué que la f_0 est perturbée par la nature intrinsèque des segments phonémiques et par leur coarticulation. Ces erreurs de calcul de la f_0 et les contraintes de production auxquelles elle est soumise montrent qu'il est important, lorsqu'on cherche à mesurer le registre à partir de la f_0 de façon automatique, d'avoir connaissance de ces problématiques qui peuvent altérer la bonne mesure du registre.

Nous avons aussi présenté les différentes unités de mesure qui peuvent être utilisées dans l'étude du registre. Nous avons ainsi présenté l'échelle linéaire des Hertz, l'échelle logarithmique (octaves, demi-tons) et les échelles psycho-acoustiques que sont l'échelle Bark, l'échelle ERB et l'échelle de MEL. Nous avons conclu que l'échelle fréquentielle linéaire exprimée en Hertz est préférable aux autres pour la mesure de hauteur de registre alors que les échelles logarithmiques et musicales des demi-tons et des octaves sont plus adaptées pour la mesure de l'étendue du registre.

Quant à la mesure même du registre, nous avons présenté la dichotomie entre mesures acoustiques (i.e. effectuées à partir de la distribution fréquentielle) et mesures linguistiques (i.e, effectuées à partir de points d'abstraction extraits de cette distribution) et nous nous sommes posé la question de savoir quel type de mesure serait le plus pertinent pour le registre. Nous avons donc abordé les avantages et inconvénients de chacune de ces mesures et avons conclu que le choix doit être porté par l'objectif de l'étude et, qu'au lieu de parler de dichotomie, il est préférable de parler de complémentarité de ces mesures.

Dans une seconde partie, nous avons défini le tempo à travers diverses problématiques : nous avons d'abord soulevé l'importance de distinguer au sein du tempo ses composantes, à savoir la vitesse d'articulation et les pauses. Les composantes du tempo une fois définies, et l'importance de leur séparation soulignée, nous avons abordé la problématique d'une mesure à la fois objective et subjective des variations de tempo. Nous avons ainsi défini trois catégories : le tempo subjectif, le tempo intentionnel et le tempo perçu et avons montré la difficulté de proposer une mesure objective qui estime le tempo perçu.

Par ailleurs, nous avons présenté les différentes unités de mesure qui ont été proposées dans

la littérature pour mesurer la vitesse d'élocution et avons cherché à déterminer, en abordant les avantages et les inconvénients de chacune, quelle unité de mesure est la plus adaptée pour les variations du débit. Nous avons conclu, après avoir élaboré des critères de sélection, que le phonème semble être une unité adéquate.

Enfin, nous avons présenté les différents domaines proposés dans la littérature pour mesurer l'empan temporel des variations de tempo et avons montré combien il est difficile, face à la multitude des domaines, de déterminer cet empan temporel. Nous avons par ailleurs souligné que la difficulté d'asseoir le domaine des variations de registre et de tempo relève du fait qu'elles revêtent de très nombreuses fonctions.

Nous avons terminé ce chapitre en abordant la problématique de l'empan temporel de l'allongement de frontière du fait qu'il interagit avec les variations de tempo. Nous avons en effet soumis l'idée que la distinction de ces deux phénomènes passe par l'étude et la détermination de leur domaine et locus respectifs. Nous avons ainsi montré que la détermination du domaine et du locus de l'allongement de frontière est complexe et qu'elle nécessiterait, à elle seule, une étude séparée.

Nous proposons donc de traiter au cours de trois chapitres expérimentaux les différentes problématiques soulevées dans le corps de ce chapitre. Les chapitres 2 et 3 chercheront à répondre aux problématiques liées à la notion de registre, à savoir sa mesure et la détermination de l'empan temporel de ses variations. Le quatrième chapitre sera consacré à la détermination de l'empan temporel des variations de tempo et de l'allongement de frontière.

Chapitre 2

Estimation des paramètres de registre

Dans ce chapitre, nous proposons de rappeler les problématiques relatives à la difficulté de mesure du registre et de ses variations et chercherons à valider expérimentalement les solutions avancées à ces problématiques.

1 Problématique

Nous avions déjà soulevé la difficulté de mesurer le registre et ses variations. La première difficulté relève des erreurs de détection et de calcul de la f_0, résultant en des valeurs aberrantes qui ont pour conséquence de fausser tout calcul automatique des différences de registre inter-locuteurs. Une solution proposée était celle d'une mesure à partir des quantiles de la distribution[60], afin d'éviter les valeurs faussées situées aux extrêmes de cette distribution. C'est pourquoi, dans un premier temps, nous testerons différents quantiles pour évaluer ceux qui décrivent au mieux des différences de registre inter-locuteurs.

La deuxième difficulté à laquelle on se heurte est celle de l'unité de mesure du registre et de ses variations. Dans ce travail, nous ne proposons pas d'évaluer les différentes échelles proposées dans la littérature, puisqu'il se dégage un certain consensus des divers travaux portant sur le

60. Les quantiles permettent de diviser une série de données en classes de taille égale. Il est à noter que le terme centile ou percentile serait plus adéquat mais nous nous rangeons dans cette thèse au terme de quantile, plus utilisé dans la littérature.

sujet, à savoir que l'échelle fréquentielle linéaire, exprimée en Hertz, est préférable à la mesure de la hauteur du registre et que les échelles logarithmiques musicales des demi-tons et des octaves ainsi que l'échelle semi-logarithmique des ERB sont plus adaptées à la mesure de son étendue. Nous avons donc choisi, pour notre travail, de mesurer la hauteur du registre en Hz (ce qui n'empêche pas une normalisation pour une étude inter-locuteurs), l'étendue en octaves, puisque ces deux unités sont unanimement reconnues corrélées à la perception du registre et de ses variations.

La troisième problématique est celle de la mesure même du registre : acoustique vs. linguistique, i.e. en termes de mesures statistiques effectuées à partir de la f_0 vs. en termes de mesures extraites de points d'inflexion de la f_0 . Contrairement au choix d'une unité de mesure, il n'apparaît pas clairement à partir de la littérature quels types de mesure décrit au mieux des variations de registre intra-locuteurs. Nous avions d'ailleurs suggéré dans le premier chapitre que ces deux types de mesure devaient plutôt être envisagés complémentaires. Nous proposons donc, dans ce chapitre, de comparer les résultats obtenus par mesures acoustiques et mesures linguistiques, afin d'évaluer les différences qui s'en dégagent et ainsi asseoir une mesure fiable du registre et de ses variations.

Nous commençons la première partie de ce chapitre par l'abord de la difficulté d'une mesure fiable de la f_0 .

2 De la difficulté d'une mesure fiable de la f_0

2.1 Corpus et base de données

61

Deux corpus ont été utilisés pour cette étude : le corpus PFC et le corpus AIX-MARSEC que nous proposons de présenter ci-après.

2.1.1 PFC

Le corpus Phonologie du Français Contemporain (PFC) (Delais-Roussarie & Durand, 2003) est le fruit d'un projet mené par Jacques Durand (Université de Toulouse-Le Mirail), Bernard Laks (Paris X, Nanterre) et Chantal Lyche (Olso et Troms), dont le but est de créer une base de données, permettant l'analyse comparative des variétés du français contemporain. Le PFC est

61. Les fichiers son utilisés pour ce chapitre se trouvent sur le CR ROM - dossier DATA_CHAP2.

en effet représentatif d'un nombre important de locuteurs (hommes et femmes, âgés entre 20 et 70 ans environ, originaires de régions diverses du monde francophone) et de différents types de production (lecture à voix haute d'une liste de mots, lecture d'un passage, conversations guidées et conversations libres).

Pour notre part, nous n'avons utilisé qu'un échantillon du corpus PFC. Les 10 locuteurs sélectionnés (6 femmes et 4 hommes) sont originaires de Marseille et ont entre 17 et 73 ans. Nous avons gardé de leur production les lectures de texte, les conversations guidées et spontanées, un total de 50 minutes d'enregistrement. La lecture de texte consiste en la lecture à voix haute d'un passage, de type article d'un journal régional, ne posant aucune difficulté de compréhension. La conversation guidée est un entretien semi-directif entre l'enquêté et un enquêteur non connu de lui. Le locuteur est interrogé sur ses activités, son travail, ses projets, son enfance, ou sur des questions d'actualité. La conversation libre est un entretien non-directif, une conversation à bâtons rompus entre l'enquêteur familier des enquêtés et les enquêtés.

2.1.2 AIX-MARSEC

Le corpus AIX-MARSEC, à l'origine SEC (Spoken English Corpus), est un corpus d'anglais britannique standard, élaboré dans les années 80, par G. Knowles, G. Leech et L. Taylor de l'université de Lancaster (UCREL : Unit for Computer Research on the English Language) et par P. Alderson, N. campbell, B. Pickering, A. Wichmann et B. Williams du centre scientifique d'IBM (IBM UKSC : United Kingdom Scientific Center). Ce corpus, en l'état, offrait un certain nombre d'informations, notamment une transcription orthographique, un étiquetage morpho-syntaxique (à l'aide de CLAWS) et une annotation prosodique (pour environ 1/5 du corpus total). Le corpus SEC est ensuite devenu le corpus MARSEC (Machine Readable Spoken English Corpus) lorsque rendu informatiquement exploitable. L'optimisation de ce corpus s'est donc faite par l'homogénéisation, la modification, sous format ASCII, des symboles prosodiques utilisés dans le SEC et par l'alignement temporel du signal sonore au niveau du mot. Le corpus MARSEC, enrichi par l'équipe EPGA du Laboratoire Parole et langage à Aix-En-Provence, est aujourd'hui le corpus AIX-MARSEC, annoté sous Praat, orthographiquement, prosodiquement et phonétiquement (Knowles et al., 1996; Auran & Bouzon, 2003; Auran, Bouzon, & Hirst, 2004).

Le corpus AIX-MARSEC présente de nombreux avantages. Il représente en effet 5h de parole authentique, 11 types de productions différents, principalement des monologues préparés, issus d'enregistrements de la BBC (des années 80), ie. des commentaires, des bulletins d'informations, des paroles publiques, des émissions religieuses, des reportages, des fictions, des poésies, des dialogues, des propagandes et autres types de production. Il compte 68 locuteurs (18 femmes et 50 hommes) dont l' accent est assez proche du « BBC English », un échantillon

donc utile pour l'étude de l'anglais britannique standard. De plus, ce corpus est riche d'annotations et répond aux besoins de notre étude. Il est en effet déjà découpé en différentes unités, ie. unités intonatives, mots, unités rythmiques (étroites et anacrouses telles que définies par Jassem (1952)), pieds (Abercrombie, 1964), syllabes, phonèmes, toutes alignées avec le signal et pour lesquelles la durée, le nombre et la position dans l'unité supérieure sont rendues tabulaires. Ces données ont donc un format approprié pour des logiciels de statistiques tels que R, un avantage précieux, notamment pour notre étude sur l'allongement et l'empan temporel des variations de tempo en anglais (cf. chapitre 4).

2.2 Protocole expérimental

Notre étude empirique s'est tout d'abord portée sur l'évaluation du registre des 68 locuteurs anglais du corpus AIX-MARSEC, du fait du caractère riche et varié d'une telle base de données. Nous avons tout d'abord annoté manuellement les valeurs extrêmes de la f_0 pour chaque locuteur (sur 1 minute de parole environ), que nous avons ensuite comparées aux mesures automatiques, obtenues à partir des quantiles de la distribution.

2.2.1 Annotation manuelle des valeurs extrêmes de la f_0

L'annotation manuelle des minima et des maxima de la f_0 s'est faite à partir d'une inspection à la fois visuelle et perceptive. L'inspection visuelle repose sur la courbe de la fréquence fondamentale, obtenue par l'utilisation du logiciel PRAAT (Boersma & Weenink, 2009) et observée dans la fenêtre d'édition d'un Objet Sound. La courbe de la f_0 correspond à une séquence de points calculés en Hertz, comme le nombre de périodes à la seconde, affichés et calculés après l'ajustement de paramètres, tels que les paramètres de registre, d'unités et autres paramètres avancés, que nous décrivons ci-après.

Les paramètres de registre assoient un seuil plancher et un seuil plafond qui délimitent les valeurs obtenues de la méthode d'analyse périodique. Les paramètres par défaut de ces seuils plancher et plafond sont ajustés à 75 Hz et 600 Hz, ce qui veut dire que les valeurs obtenues ne peuvent être inférieures à la valeur plancher ni supérieures à la valeur plafond. Les auteurs du logiciel préconisent cependant un ajustement des seuils en fonction du registre du locuteur, e.g. entre 75 Hz et 300 Hz pour une voix d'homme, entre 100 Hz et 500 Hz pour une voix de femme, en-deçà des 75 Hz pour des voix craquées. L'ajustement est en effet primordial car les valeurs obtenues en dépendent. Si la valeur plancher est trop basse, les changements rapides de la f_0 ne pourront être estimés ; à contrario, si la valeur plancher est trop haute, certaines valeurs très basses ne pourront être détectées. Cela vient du fait que déterminer une valeur plancher revient également à déterminer la fenêtre d'analyse pour laquelle les valeurs

fréquentielles sont calculées. La fenêtre d'analyse (f), qui correspond à un maximum de 3 périodes, est calculée comme suit : $f = \frac{3}{p}$, où p est la valeur donnée au seuil plancher. Si la valeur plancher est trop basse (e.g. 25 Hz), la fenêtre d'analyse est alors « trop longue », i.e. 120 ms ($\frac{3}{25} = 0.120$) pour estimer les changements rapides de la f_0 . Dans cette étude, nous avons donc ajusté les paramètres de registre en fonction du sexe du locuteur et selon les valeurs préconisées par les auteurs, bien que certains réajustements aient été opérés pour des voix de femmes très graves (i.e. abaissement de la valeur plancher) ou pour des voix de femmes assez monotones qui varient sur un espace tonal plutôt réduit(i.e. abaissement de la valeur plafond).

Le paramètre *unités* indique l'unité d'échelonnage des valeurs fréquentielles. Dans ce travail, nous avons choisi l'échelle fréquentielle linéaire des Hertz.

Les paramètres *avancés* concernent l'ajustement d'un ensemble de paramètres à partir desquels les valeurs fréquentielles candidates sont sélectionnées, tels que l'échelonnage des valeurs dans l'éditeur, le type de méthode utilisée dans la détection de périodicité (e.g. méthode d'autocorrélation), le pas d'analyse, le seuil de voisement, le seuil de non voisement, le coût d'octave, le coût de saut d'octave et le coût de détection de voisement. Dans ce travail, nous nous sommes rangée aux valeurs par défaut proposées par Boersma et Weenink (2009).

Une fois ces paramètres ajustés, nous avons effectué une annotation manuelle des valeurs minimale et maximale des registres des 68 locuteurs. Dans cette annotation, nous avons écarté toute erreur de détection et de calcul de la f_0 (e.g. erreurs de voisement et sauts d'octave) et toutes perturbations de la f_0 résultant des effets intrinsèques et co-intrinsèques tels que décrits en 1.5.1, et ce, par le biais d'une analyse perceptive. Tout changement abrupt de la f_0, à hauteur d'une octave par exemple, non perçu a été écarté de l'analyse (cf. figure 17). Ou encore, toute détection de f_0 dans un « cadre » de non voisement a également été exclu (cf. figure 18). Les valeurs maximale et minimale ont été annotées à partir d'une fenêtre grossissante ou zoom dont la partie visible du signal correspond à 200 ms et pour laquelle une valeur minimale et une valeur maximale fréquentielles sont données. Fonction du temps, les valeurs minimale et maximale sont remplacées si la valeur minimale de la fenêtre « en cours » est inférieure à la fenêtre précédente et, inversement, si la valeur maximale de la fenêtre « en cours » est supérieure à la valeur maximale de la fenêtre précédente. Ainsi, nous obtenons, en fin d'analyse, les valeurs extrêmes du registre du locuteur.

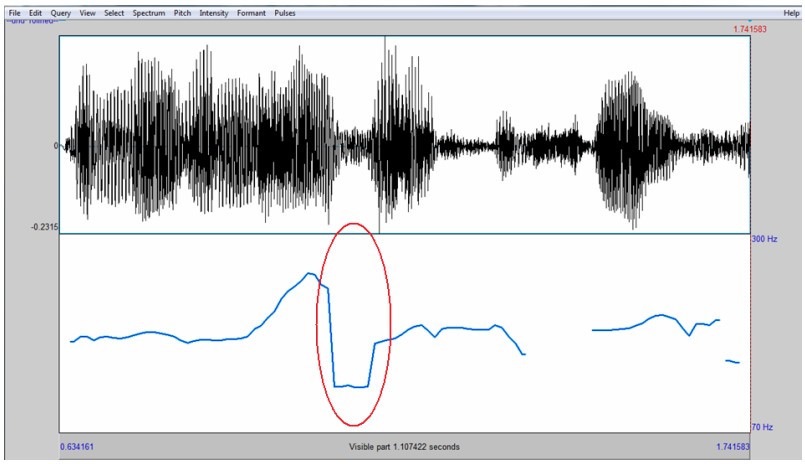


FIGURE 17 – Exemple de saut d'octave extrait du fichier son A1101G.

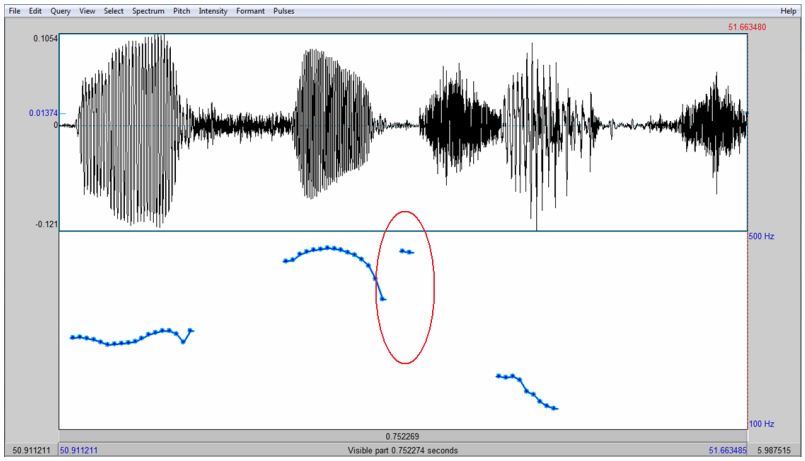


FIGURE 18 – Exemple d'erreur de détection de voisement extrait du fichier son A0101B.

2.2.2 Détection automatique des valeurs extrêmes de la f_0 : en l'état

La difficulté d'obtenir automatiquement des valeurs fiables de la f_0, nous l'avons dit, relève du fait que des valeurs aberrantes peuvent se glisser dans ces mesures automatiques. Après avoir effectué une détection automatique des registres de 53 locuteurs (39 hommes et 14 femmes) du corpus Aix-Marsec [62] [63], dont les paramètres plancher et plafond sont ajustés aux valeurs données par défaut (soit 75 Hz et 600 Hz), il apparaît qu'une mesure automatique (à l'aide des

62. Nous avons éliminé de notre analyse les fichiers audio marqués par trop de bruit pour lesquels une analyse de la f_0 est impossible.
63. cf. Tableau des valeurs - CD ROM, ANNEXES_CHAP2 : Table1.

commandes Get minimum... et Get maximum...) ne permet pas de rendre compte des valeurs extrêmes du registre d'un locuteur. La figure 19montre clairement que les valeurs maximales obtenues automatiquement (PRAATMAX)[64] sont généralement plus élevées que celles notées manuellement (MMAX), graphiquement situées entre 450 Hz et 650 Hz contre, respectivement, 200 et 450 Hz. Les valeurs minimales obtenues automatiquement (PRAATMIN) sont relativement proches de celles notées manuellement (MMIN) pour les locuteurs hommes, mais l'écart se creuse entre valeurs automatiques et valeurs manuelles pour les locutrices femmes. Cela s'exlique par le fait que un seuil estimé à 75 Hz convient à l'étude de voix d'hommes. Utiliser les paramètres par défaut revient donc à prendre en compte les valeurs aberrantes et donc ne permet pas de décrire les différences de registre inter-locuteurs. Les analyses de régression effectuées à partir de l'environnement de travail R montrent d'ailleurs un effet du sexe du locuteur sur les valeurs minimales et maximales annotées manuellement (respectivement, F(1,52)=249.1, p-value<2.2e-16 ; F(1,52)=87.13, p-value=1.062e-12) alors qu'aucun effet n'est établi avec les valeurs minimales et maximales obtenues automatiquement (respectivement, F(1,52)=2.25, p-value=0.1396 ; F(1,52)=1.471, p-value=0.2307).

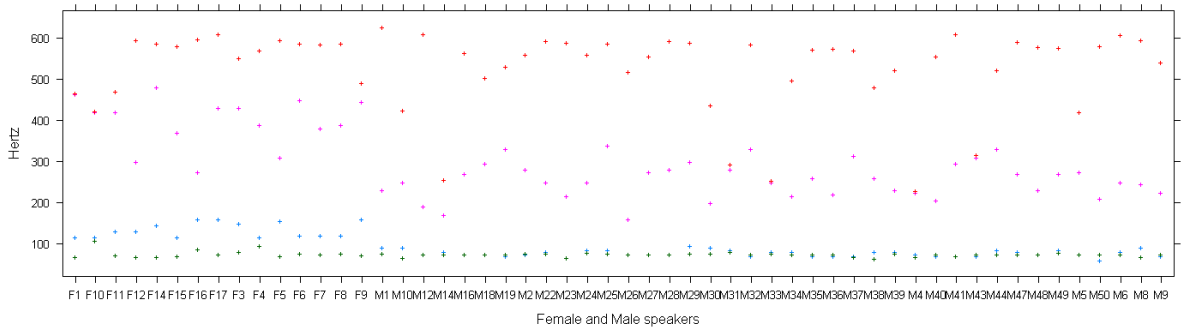


FIGURE 19 – Nuage de points des valeurs minimales annotées manuellement (MMIN) en bleu, obtenues par un ajustement des seuils par défaut (PRAATMIN) en vert, des valeurs maximales annotées manuellement (MMAX) en rose et obtenues par ajustement des seuils par défaut (PRAATMAX) en rouge.

Cependant de tels résultats ne sont pas surprenants, les auteurs préconisant, nous l'avons dit, un ajustement des paramètres plancher et plafond en fonction du sexe du locuteur afin d'obtenir une détection fiable de la f_0 . Un plancher donné à 75 Hz pour les voix d'hommes, 100 Hz pour les voix de femmes et un plafond donné respectivement à 300 Hz et 500 Hz pour

64. L'obtention des valeurs extrêmes se fait à la suite de la création d'un objet Pitch dont les paramètres *pas d'analyse*, *plancher* et *plafond* sont ajustés aux valeurs 0.01, 75 et 600 respectivement. Une fois l'objet Pitch créé, les valeurs extrêmes sont obtenues automatiquement par les commandes Get minimum... et Get maximum... .

les voix d'hommes et de femmes, permettent en effet de décrire des différences de registre interlocuteurs, notamment des différences hommes / femmes (p-value : $<$ 2.2e-16). La réduction des valeurs aberrantes aux extrêmes de la courbe est nette puisque la distance obtenue [65] entre valeurs obtenues par ajustement en fonction du sexe du locuteur (SEXMIN et SEXMAX) et valeurs de référence (MMIN et MMAX) est inférieure à celle obtenue entre les valeurs obtenues par ajustement des seuils par défaut (PRAATMIN/PRAATMAX) et MMIN/MMAX : 0.12 contre 0.16 pour la distance des minima et 0.13 contre 0.90 pour la distance des maxima. Cependant, cette procédure requiert en amont d'ajuster les paramètres plancher et plafond en fonction du sexe du locuteur et l'espace tonal résultant de l'ajustement de ces paramètres peut paraître également trop large, notamment pour les voix monotones. Il apparaît à la fois le besoin d'adapter les paramètres plancher et plafond au registre du locuteur et d'ajuster ces paramètres « sans coût ». Nous proposons donc une méthode d'ajustement de ces paramètres par l'utilisation de valeurs quantiles qui, associées à un coefficient, permettraient l'obtention des valeurs extrêmes de tout type de registre (e.g. voix d'hommes vs. voix de femmes, voix monotones vs. voix vives).

2.2.3 « Filtrage » des valeurs aberrantes par la méthode des quantiles

Expérience 1. —

La première partie de l'étude a consisté en l'évaluation systématique de la corrélation des quantiles q5, q10, q15, q20, q25, q30, q35 et q40 avec les valeurs MMIN et des quantiles q60, q65, q70, q75, q80, q85, q90 et q95 avec les valeurs MMAX. Nous avions vu en effet que lorsque les paramètres plancher et plafond étaient ajustés aux valeurs par défaut (i.e. 75 et 600 Hz), les valeurs maximales de la f_0 avaient tendance à être surestimées, les valeurs minimales sous-estimées. Une étude des quantiles apparaît donc intéressante afin d'élaguer les valeurs aberrantes. Aussi, à partir de chaque fichier son ou objet Sound, un objet Pitch est créé automatiquement [66] avec comme paramètres de pas d'analyse, de seuils plancher et plafond les valeurs respectives 0.01, 60 [67] et 600. A l'aide de la commande Get quantile..., les valeurs quantiles pour chaque objet Pitch sont obtenues et répertoriées dans un tableau [68]. Afin de comparer les valeurs quantiles obtenues aux valeurs de référence (MMIN et MMAX), on définit, par la formule suivante (1), une distance entre les valeurs quantiles et les valeurs de référence :

65. Le calul de la distance est décrit en 2.2.3.
66. cf. script Calculate Quantiles - CD ROM, dossier SCRIPTS : Calculate_Quantiles.
67. Nous descendons le seuil plancher à 60 Hz au lieu de 75 Hz afin de couvrir une large gamme de voix, notamment des voix d'hommes très graves.
68. cf. Tableau des valeurs - CR ROM, ANNEXES_CHAP2 : Table2.

$$\chi = \sum \frac{|q_i - r_i|}{r_i} \tag{1}$$

où q représente les valeurs quantiles et r les valeurs de référence.

Résultats 1. —

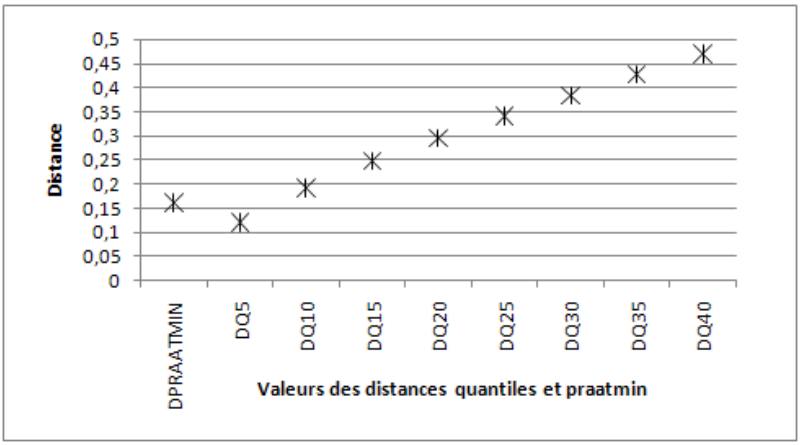


FIGURE 20 – Nuage de points des distances calculées entre valeurs de référence minimales (MMIN) et valeurs quantiles (Qn), comparées à la distance calculée entre MMIN et valeurs minimales obtenues par ajustement des seuils par défaut (PRAATMIN) - corpus AIX-MARSEC.

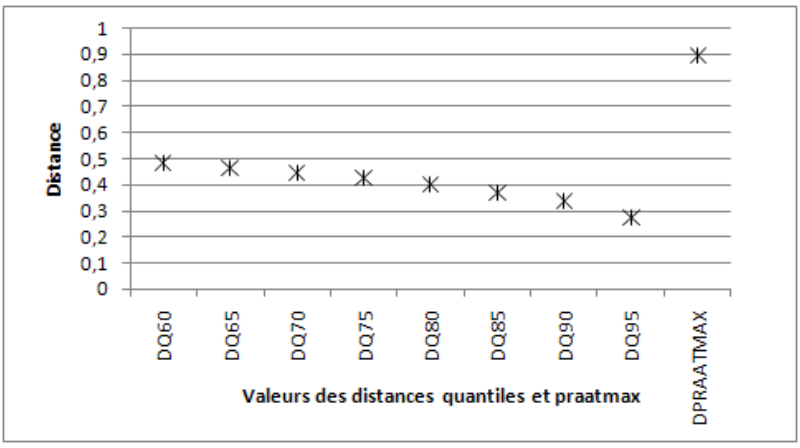


FIGURE 21 – Nuage de points des distances calculées entre valeurs de référence maximales (MMAX) et valeurs quantiles (Qn), comparées à la distance calculée entre MMAX et valeurs maximales obtenues par ajustement des seuils par défaut (PRAATMAX) - corpus AIX-MARSEC.

Les figures 20 et 21 [69] montrent que les valeurs quantiles q5 s'approchent plus des valeurs minimales de référence (MMIN) que PRAATMIN. Quant aux valeurs maximales, les valeurs quantiles q60, q65, q70, q75, q80, q85, q90 et q95 s'approchent plus des valeurs de référence MMAX que PRAATMAX. Si l'on regarde en détails les distances obtenues en fonction du sexe du locuteur, on observe (cf. figure 22), pour les voix de femmes, que les valeurs quantiles q5, q10, q15, q20, q25, q30, q35 et q40 sont plus corrélées aux valeurs MMIN que ne peuvent l'être les valeurs PRAATMIN ; au contraire, pour les voix d'hommes (cf. figure 24), les valeurs PRAATMIN sont les plus corrélées aux valeurs MMIN. Cela montre bien qu'ajuster la valeur plancher à 75 Hz convient aux voix d'hommes et non aux voix de femmes. Si l'on se penche sur les valeurs maximales, il apparaît que les valeurs quantiles q75, q80, q85, q90, q95 sont plus corrélées aux valeurs MMAX que les valeurs PRAATMAX pour les voix de femmes (cf. figure 23) ; pour les voix d'hommes (cf. figure 25), toutes les valeurs quantiles se montrent plus corrélées aux valeurs MMAX. Il est donc clair qu'un seuil plafond ajusté à 600 Hz est trop élevé pour les voix de femmes et ne convient pas aux voix d'hommes.

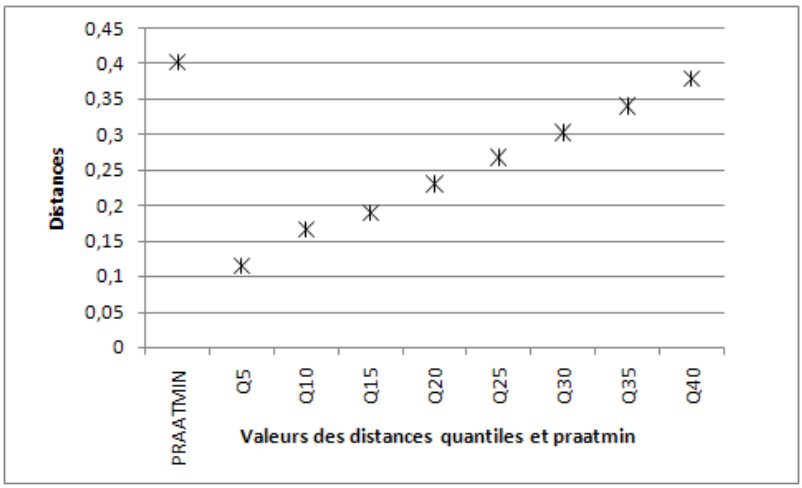


FIGURE 22 – Nuage de points des distances calculées entre valeurs de référence minimales (MMIN) et valeurs quantiles (Qn), comparées à la distance calculée entre MMIN et valeurs minimales obtenues par ajustement des seuils par défaut (PRAATMIN), pour les locutrices femmes - corpus AIX-MARSEC.

69. cf. Tableau des valeurs - CR ROM, ANNEXES_CHAP2 : Table3.

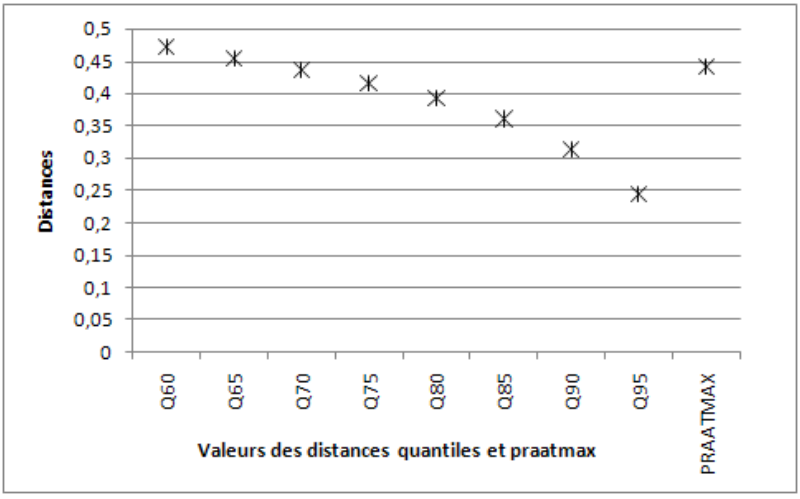


FIGURE 23 – Nuage de points des distances calculées entre valeurs de référence maximales (MMAX) et valeurs quantiles (Qn), comparées à la distance calculée entre MMAX et valeurs maximales obtenues par ajustement des seuils par défaut (PRAATMAX), pour les locutrices femmes - corpus AIX-MARSEC.

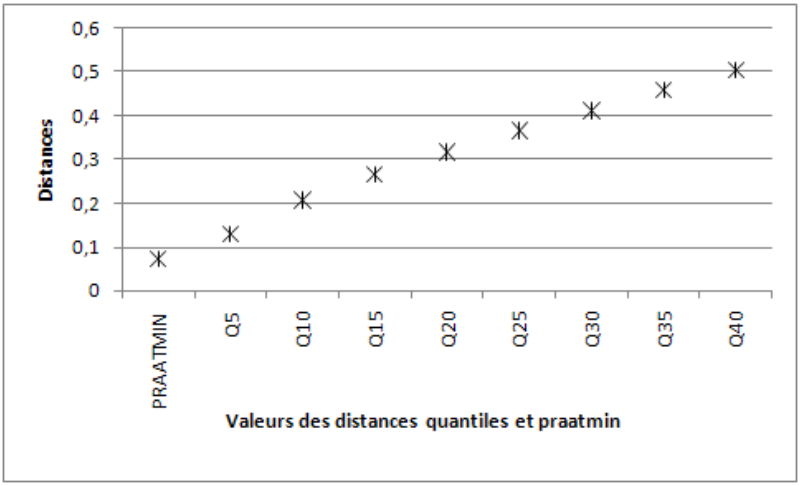


FIGURE 24 – Nuage de points des distances calculées entre valeurs de référence minimales (MMIN) et valeurs quantiles (Qn), comparées à la distance calculée entre MMIN et valeurs minimales obtenues par ajustement des seuils par défaut (PRAATMIN), pour les locuteurs hommes - corpus AIX-MARSEC.

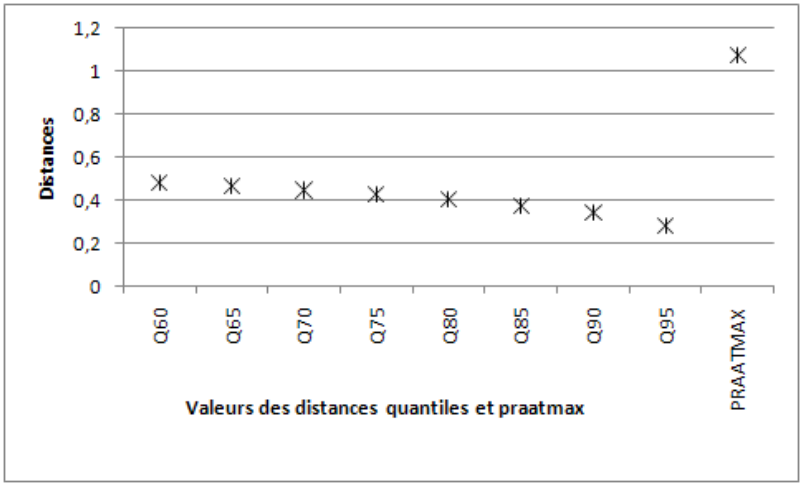


FIGURE 25 – Nuage de points des distances calculées entre valeurs de référence maximales (MMAX) et valeurs quantiles (Qn), comparées à la distance calculée entre MMAX et valeurs maximales obtenues par ajustement des seuils par défaut (PRAATMAX), pour les locuteurs hommes - corpus AIX-MARSEC.

De tels résultats justifient les paramètres proposés par les auteurs en fonction du sexe du locuteur (i.e. voix de femmes 100-500 ; voix d'hommes 75-300) et attestent également du possible ajustement des paramètres plancher et plafond en fonction du registre du locuteur et, par là, d'une meilleure détection des valeurs extrêmes de la f_0. Utiliser les valeurs quantiles comme paramètres plancher et plafond permettrait un ajustement plus précis au registre du locuteur. Les valeurs paramétriques ne dépendraient plus du sexe du locuteur mais plutôt du type de voix (i.e. haute vs. basse, monotone vs. vive). Nous proposons donc, dans une seconde expérience, de tester ces différents quantiles, en combinaison, comme valeurs plancher et plafond.

Expérience 2. —

Dans cette deuxième partie de l'étude, nous avons testé la combinaison des différents quantiles comme seuils plancher et plafond, afin d'évaluer la corrélation entre les valeurs extrêmes obtenues automatiquement par de tels paramétrages et les valeurs de référence, annotées manuellement. Pour cela, pour chaque objet Sound est créé un objet Pitch dont les seuils plancher et plafond sont ajustés aux valeurs par défaut (ici 60-600). Les valeurs quantiles de cet objet Pitch sont obtenues automatiquement, toujours à l'aide de la commande Get quantile... et sont utilisées à la création de nouveaux objets Pitch comme valeurs de seuils plancher et plafond. La combinaison des quantiles q5, q10, q15, q20, q25, q30, q35, q40 comme valeurs plancher et des quantiles q60, q65, q70, q75, q80, q85, q90 et q95 comme valeurs plafond résulte, pour

chaque objet Sound, la création de 68 objets Pitch. A partir de chaque objet Pitch, les valeurs extrêmes de la f_0 sont obtenues par les commandes Get minimum... et Get maximum... et répertoriées dans un tableau. Un ajustement par la méthode des moindres carrés permet d'obtenir les distances entre valeurs extrêmes automatiques et valeurs de référence et d'évaluer quelle(s) famille(s) (e.g. combinaison q5q95, q5q90, q15q60) « élague(nt) » au mieux les valeurs aberrantes [70].

Résultats 2. —

Les résultats obtenus (cf. figures 26 et 27) montrent que les combinaisons dont le plancher est ajusté au quantile 5 permettent une meilleure détection des minima alors que, pour les maxima, l'ajustement des seuils en fonction du sexe du locuteur reste le plus efficace. Il apparaît donc que les valeurs quantiles comme seuils plancher et plafond ne permettent pas, à elles-seules, une bonne détection des valeurs extrêmes de la f_0. Nous proposons donc, dans une troisième expérience, d'allier les valeurs quantiles à des rapports, dans le but d'optimiser les paramètres plancher et plafond.

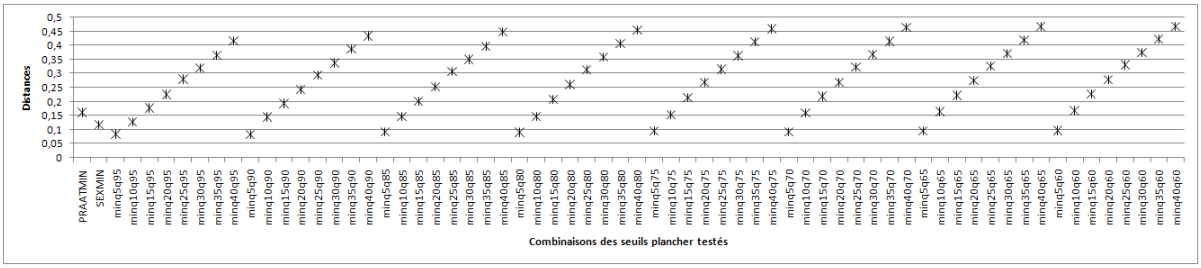


FIGURE 26 – Nuage de points des distances calculées entre valeurs de référence minimales (MMIN) et valeurs minimales obtenues par ajustement des seuils plancher et plafond par défaut (PRAATMIN), par ajustement en fonction du sexe du locuteur (SEXMIN) et par ajustement aux valeurs quantiles (68 combinaisons) - corpus AIX-MARSEC.

70. cf. Tableau des valeurs - CR ROM, ANNEXES_CHAP2 : Table4.

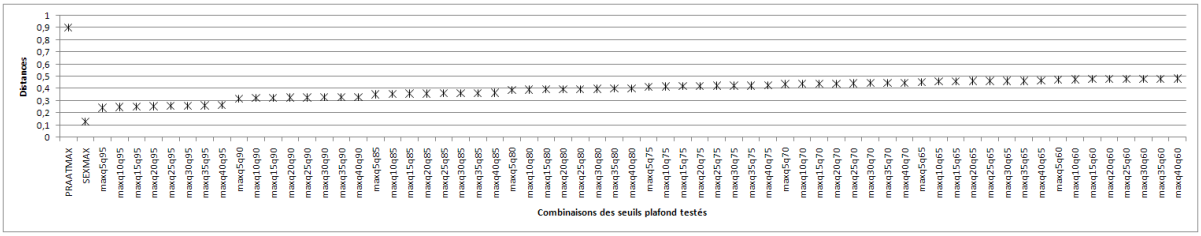


FIGURE 27 – Nuage de points des distances calculées entre valeurs de référence maximales (MMAX) et valeurs maximales obtenues par ajustement des seuils plancher et plafond par défaut (PRAATMAX), par ajustement en fonction du sexe du locuteur (SEXMAX) et par ajustement aux valeurs quantiles (68 combinaisons) - corpus AIX-MARSEC.

Expérience 3. —

A travers cette troisième expérience, nous avons cherché à ajuster les seuils plancher et plafond au registre du locuteur. Pour cela, nous avons cherché à approcher les valeurs de ces seuils aux valeurs minimales et maximales annotées manuellement, une façon adéquate, selon nous, de réduire l'espace tonal des variations de la f_0 et ainsi réduire les valeurs aberrantes aux extrêmes de la courbe. Les valeurs quantiles obtenues dans l'expérience 1 sont donc corrélées aux valeurs de référence minimales et maximales (MMIN et MMAX) pour chaque objet Pitch. Les mesures de ces corrélations sont obtenues par le calcul des rapports des valeurs de référence et des valeurs quantiles (e.g. $\frac{MMIN_i}{q5_i}$). Le rapport optimal résulte du calcul de la médiane [71] des valeurs des rapports. Il est à noter que nous avons réduit, pour le calcul du coefficient optimal, le nombre de locuteurs à celui de 28 [72], afin que le calcul se fasse à partir d'autant de voix de femmes que de voix d'hommes. [73].

71. La moyenne également testée s'est avérée « moins performante » que la médiane, certainement de par le fait qu'elle a pu donner trop de poids à des rapports extrêmes dans le calcul du coefficient optimal.
72. Le choix des locuteurs s'est fait de façon aléatoire.
73. cf. Tableau de valeurs - CR ROM, ANNEXES_CHAP2 : Table5.

Quantiles	Rapports	Quantiles	Rapports
Q5	0,924880635	Q60	2,001519542
Q10	0,874252494	Q65	1,923137084
Q15	0,83390136	Q70	1,850654359
Q20	0,80486943	Q75	1,771653683
Q25	0,780638672	Q80	1,689903204
Q30	0,757159278	Q85	1,575306149
Q35	0,737069402	Q90	1,452368396
Q40	0,71611907	Q95	1,271669208

TABLE 1 – Rapports optimaux calculés pour chaque quantile.

Le tableau 1 présente les rapports optimaux pour chaque quantile. Une fois les rapports optimaux déterminés, le produit des rapports et des quantiles est utilisé comme seuils plancher et plafond dans la création de nouveaux objets Pitch. De la combinaison de ces différents produits, résulte la création de 68 objets Pitch (pour chaque objet Sound [74]), pour lesquels nous récupérons les valeurs minimales et maximales détectées automatiquement [75]. Nous les comparons ensuite aux valeurs minimales et maximales de référence (MMIN et MMAX) afin de définir la combinaison qui permet une meilleure détection des extrêma de la f_0 .

Résultats 3. —

Les distances obtenues pour l'ensemble des combinaisons testées [76], comme paramètres plancher et plafond, sont systématiquement inférieures à celles obtenues pour les seuils ajustés au sexe du locuteur (SEXMIN et SEXMAX), une distance évaluée à 0.08 (rapport) en moyenne pour les minima et 0.09 pour les maxima contre 0.12 et 0.13. Concernant l'estimation des valeurs minimales, les combinaisons dont le plancher est le produit du 5^{me} quantile et du coefficient 0.92, le produit du 10^{me} quantile et du coefficient 0.87 et le produit du 15^{me} quantile et du coefficient 0.83 s'avèrent être les plus adaptées (cf. figure 28). Les combinaisons dont le plafond est le produit des 85^{me}, 80^{me}, 75^{me}, 70^{me}, 65^{me} et 60^{me} quantiles et des coefficients 1.57, 1.69, 1.77, 1.85, 1.92 et 2.00 respectivement, permettent, quant à elles, une meilleure estimation des maxima (cf. figure 29). 18 combinaisons possibles permettent donc une bonne corrélation des valeurs prédites aux valeurs de référence.

74. L'évaluation des seuils plancher et plafond se fait pour la totalité des enregistrements sélectionnés, i.e. pour 53 locuteurs.
75. cf. Tableau de valeurs - CD ROM, ANNEXES_CHAP2, Table6.
76. cf. tableaux 28 et 29

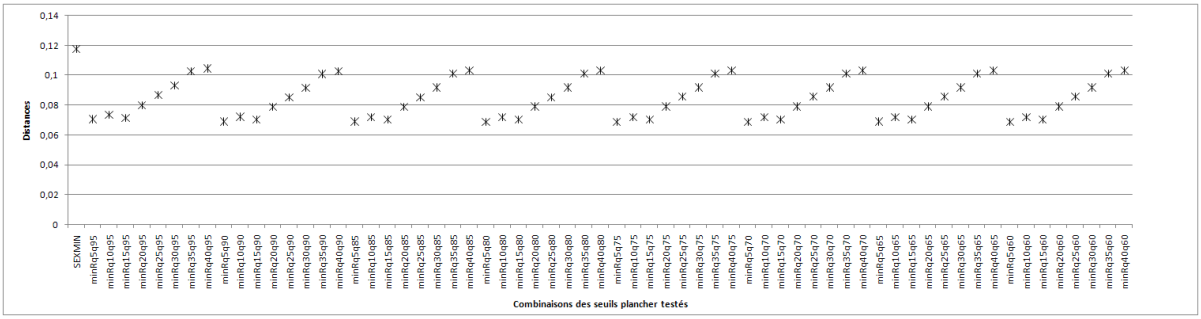


FIGURE 28 – Nuage de points des distances calculées entre valeurs de référence minimales (MMIN) et valeurs minimales obtenues par ajustement des seuils plancher et plafond aux produits des valeurs quantiles et de leur rapport, comparées à la distance calculée entre valeurs de référence minimales et valeurs minimales obtenues par l'ajustement des seuils en fonction du sexe du locuteur (SEXMIN)- corpus AIX-MARSEC.

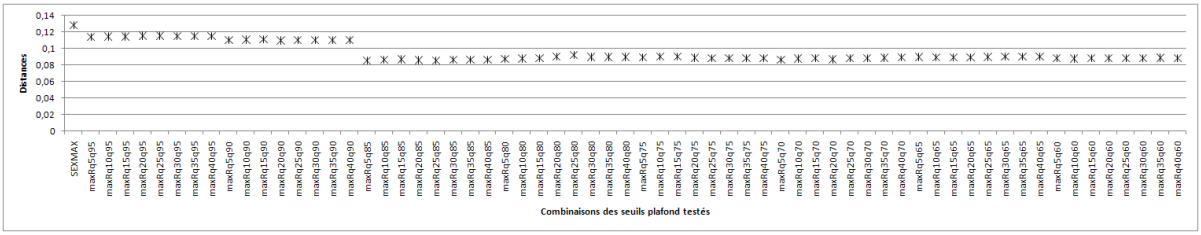


FIGURE 29 – Nuage de points des distances calculées entre valeurs de référence maximales (MMAX) et valeurs maximales obtenues par ajustement des seuils plancher et plafond aux produits des valeurs quantiles et de leur rapport, comparées à la distance calculée entre valeurs de référence maximales et valeurs maximales obtenues par ajustement des seuils en fonction du sexe du locuteur (SEXMAX) - corpus AIX-MARSEC

L'analyse du coefficient de corrélation (i.e. $\sqrt{r^2}$) entre les valeurs de référence et les valeurs prédites permet d'illustrer le gain apporté par l'ajustement des seuils plancher et plafond au registre du locuteur (cf. tableau 2). Alors que le score obtenu par un ajustement des seuils par défaut (75-600) est de 0.17 pour les minima et 0.14 pour les maxima, il est de 0.95 et 0.89 par un ajustement au registre du locuteur (pour les 18 combinaisons). Les coefficients de corrélation d'un ajustement au registre du locuteur sont également supérieurs aux coefficients d'un ajustement en fonction du sexe du locuteur (75-300 pour les hommes ; 100-500 pour les femmes) estimés à 0.86 pour les minima et 0.85 pour les maxima. L'ajustement des seuils par les quantiles permet donc l'obtention de meilleurs coefficients de corrélation et ce, de façon automatique. Il présente donc un double avantage : celui de l'obtention de valeurs plus fiables

aux extrêmes de la courbe de la f_0 et celui d'une méthode « sans coût ».

Seuils Plancher	Coeff Corrélation	Seuils Plafond	Coeff Corrélation
praatmin	0,167630546	praatmax	0,136014705
sexmin	0,863652652	sexmax	0,854341852
minRq5q85	0,955053898	maxRq5q85	0,898333638
minRq10q85	0,951892415	maxRq10q85	0,89634198
minRq15q85	0,950644638	maxRq15q85	0,895482535
minRq5q80	0,955533511	maxRq5q80	0,899205951
minRq10q80	0,952010241	maxRq10q80	0,898624911
minRq15q80	0,950644638	maxRq15q80	0,897927262
minRq5q75	0,955533511	maxRq5q75	0,894681151
minRq10q75	0,952010241	maxRq10q75	0,892896494
minRq15q75	0,950644638	maxRq15q75	0,892926799
minRq5q70	0,955533511	maxRq5q70	0,896163584
minRq10q70	0,952010241	maxRq10q70	0,893636185
minRq15q70	0,950644638	maxRq15q70	0,893659418
minRq5q65	0,955053898	maxRq5q65	0,891286884
minRq10q65	0,951892415	maxRq10q65	0,893258809
minRq15q65	0,950644638	maxRq15q65	0,89326268
minRq5q60	0,955533511	maxRq5q60	0,894429229
minRq10q60	0,952010241	maxRq10q60	0,896373604
minRq15q60	0,950644638	maxRq15q60	0,894280378
Moyqr	0,952663081	Moyqr	0,895153972

TABLE 2 – Tableaux des coefficients de corrélations entre les valeurs de référence et les valeurs prédites pour les 18 combinaisons.

2.3 Etude de réplication

Afin de voir quelle combinaison est la plus « stable » et si les rapports estimés ne sont pas dépendants du corpus utilisé, nous avons décidé de répliquer l'expérience 3 à partir du corpus PFC. Nous avons annoté les extrêma de la courbe de la f_0 des 10 locuteurs (4 hommes et 6 femmes) dans 3 types de production différents, selon le même protocole (décrit en 2.2.1). Puis, nous avons corrélé les valeurs de référence (MMIN et MMAX) aux valeurs prédites par les 68 combinaisons possibles de seuils plancher et plafond ainsi qu'aux valeurs obtenues par l'ajustement des seuils plancher et plafond au sexe du locuteur et par défaut[77]. Les figures

77. cf. Tableau de valeurs - CD ROM, ANNEXES_CHAP2 : Table7.

30 et 31 montrent les distances obtenues pour l'ensemble des ajustements testés [78]. Pour les minima, bien que la distance pour chacune des 68 combinaisons soit inférieure aux distances obtenues avec un ajustement au sexe du locuteur et un ajustement par défaut, il apparaît que les combinaisons dont le plancher est ajusté aux quantiles q5 et q10 sont moins adaptées que celles dont le plancher est ajusté au quantile q15, qui apparaît plus stable dans l'estimation des minima de la f_0 (distance minimale). Quant aux maxima, les distances des combinaisons dont le seuil plafond est ajusté aux quantiles q80, q75, q70, q65 et q60 sont inférieures à celles obtenues avec un ajustement au sexe du locuteur et un ajustement par défaut. Les distances obtenues avec un seuil plafond ajusté au quantile q65 sont les plus petites.

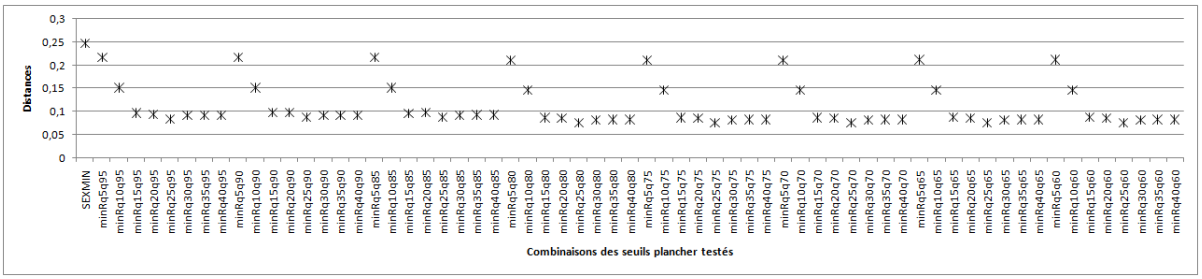


FIGURE 30 – Nuage de points des distances calculées entre valeurs de référence minimales (MMIN) et valeurs minimales obtenues par ajustement des seuils plancher et plafond aux produits des valeurs quantiles et de leur rapport, comparées à la distance calculée entre MMIN et valeurs minimales obtenues par l'ajustement des seuils en fonction du sexe du locuteur (SEXMIN) - corpus PFC.

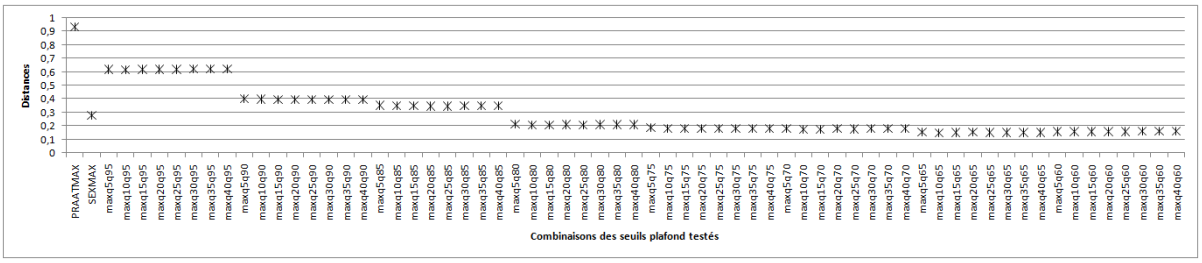


FIGURE 31 – Nuage de points des distances calculées entre valeurs de référence maximales (MMAX) et valeurs maximales obtenues par ajustement des seuils plancher et plafond aux produits des valeurs quantiles et de leur rapport, comparées à la distance calculée entre MMAX et valeurs maximales obtenues par ajustement des seuils en fonction du sexe du locuteur (SEXMAX) - corpus PFC.

78. cf. Tableau de valeurs - CD ROM, ANNEXES_CHAP2 : Table8.

La combinaison, dont le seuil plancher est le produit du 15^{me} quantile et du coefficient 0.83 et dont le seuil plafond est le produit du 65^{me} quantile et du coefficient 1.92, est donc la plus adéquate dans l'estimation des valeurs extrêmes de la f_0 . L'analyse du coefficient de corrélation entre les valeurs de référence et les valeurs prédites permet d'illustrer à nouveau le gain apporté par l'ajustement des seuils plancher et plafond au registre du locuteur. Le tableau 3 et les figures 32 et 33 montrent, pour les deux corpus, une plus grande corrélation entre les valeurs de référence et les valeurs prédites lorsque les seuils plancher et plafond sont ajustés au registre du locuteur (i.e. q15*0.83 et q65*1.92) que lorsqu'ils sont ajustés au sexe du locuteur (75-300 pour les hommes ; 100-500 pour les femmes) ou donnés par défaut (75-600).

	AM	PFC
PraatMin	0,167630546	0,350285598
SexMin	0,863655024	0,775628777
Minq15q65	**0,950631369**	**0,928116372**
PraatMax	0,136014705	0,232379001
SexMax	0,854341852	0,775499839
Maxq15q65	**0,893252484**	**0,845162706**

TABLE 3 – Tableau des coefficients de corrélations entre les valeurs de référence et les valeurs prédites par ajustement des seuils par défaut (PraatMin/PraatMax), en fonction du sexe du locuteur (SexMin/ SexMax) et par ajustement au produit des valeurs quantiles q15 et q65 aux rapports 0.83 et 1.92 - Corpus AIX-MARSEC et PFC.

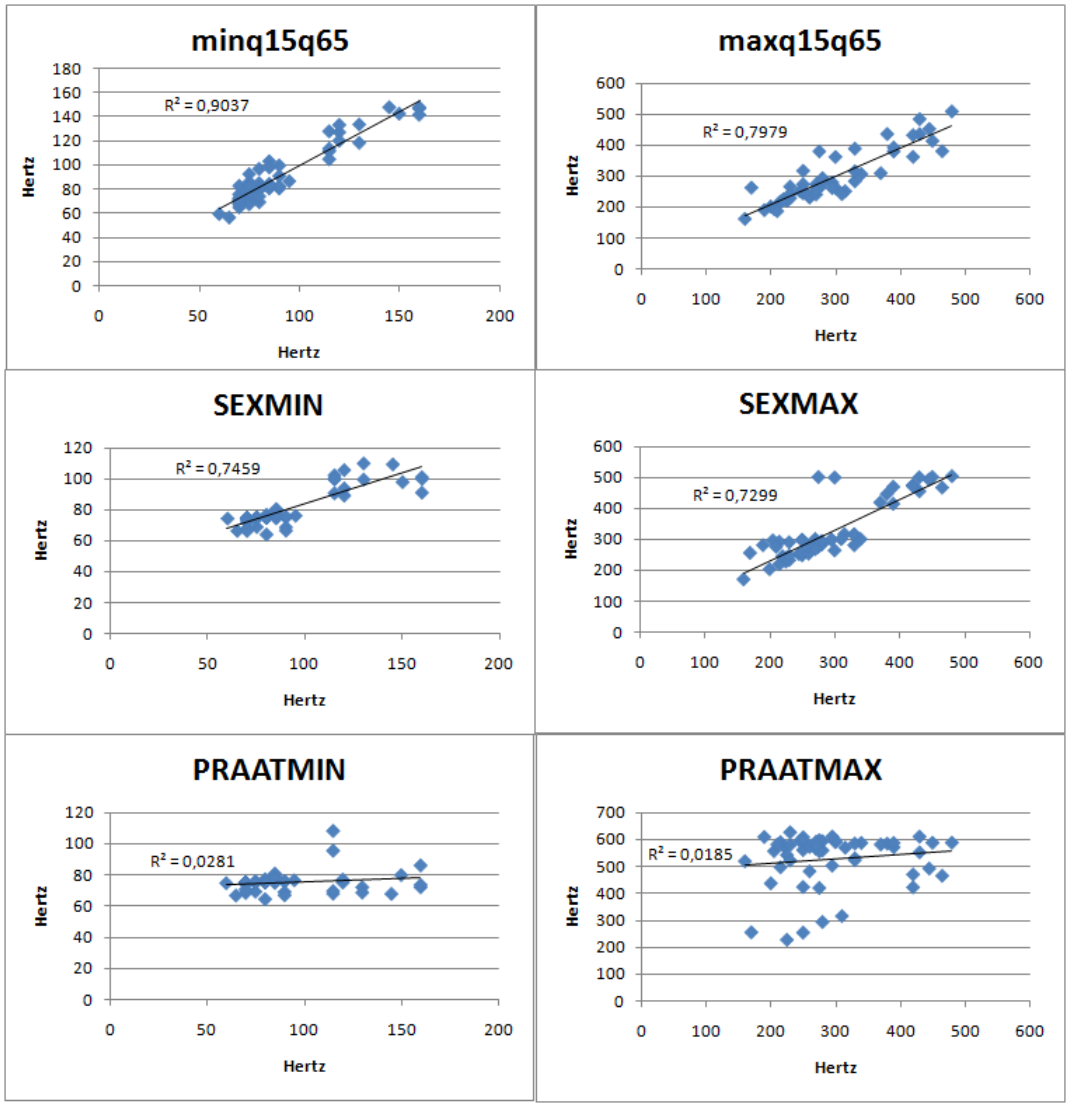


FIGURE 32 – Nuage de points illustrant les corrélations entre valeurs minimales et maximales de référence (MMIN et MMAX) et valeurs minimales et maximales obtenues par ajustement des seuils plancher et plafond aux produits des quantiles q15 et q65 et des rapports 0.84 et 1.92 (minq15q65/ maxq15q65) ; entre MMIN et MMAX et valeurs obtenues par ajustement des seuils en fonctions du sexe du locuteur (SEXMIN/ SEXMAX) ; entre MMIN et MMAX et valeurs obtenues par ajustement des seuils par défaut (PRAATMIN/ PRAATMAX)- Corpus AIX-MARSEC.

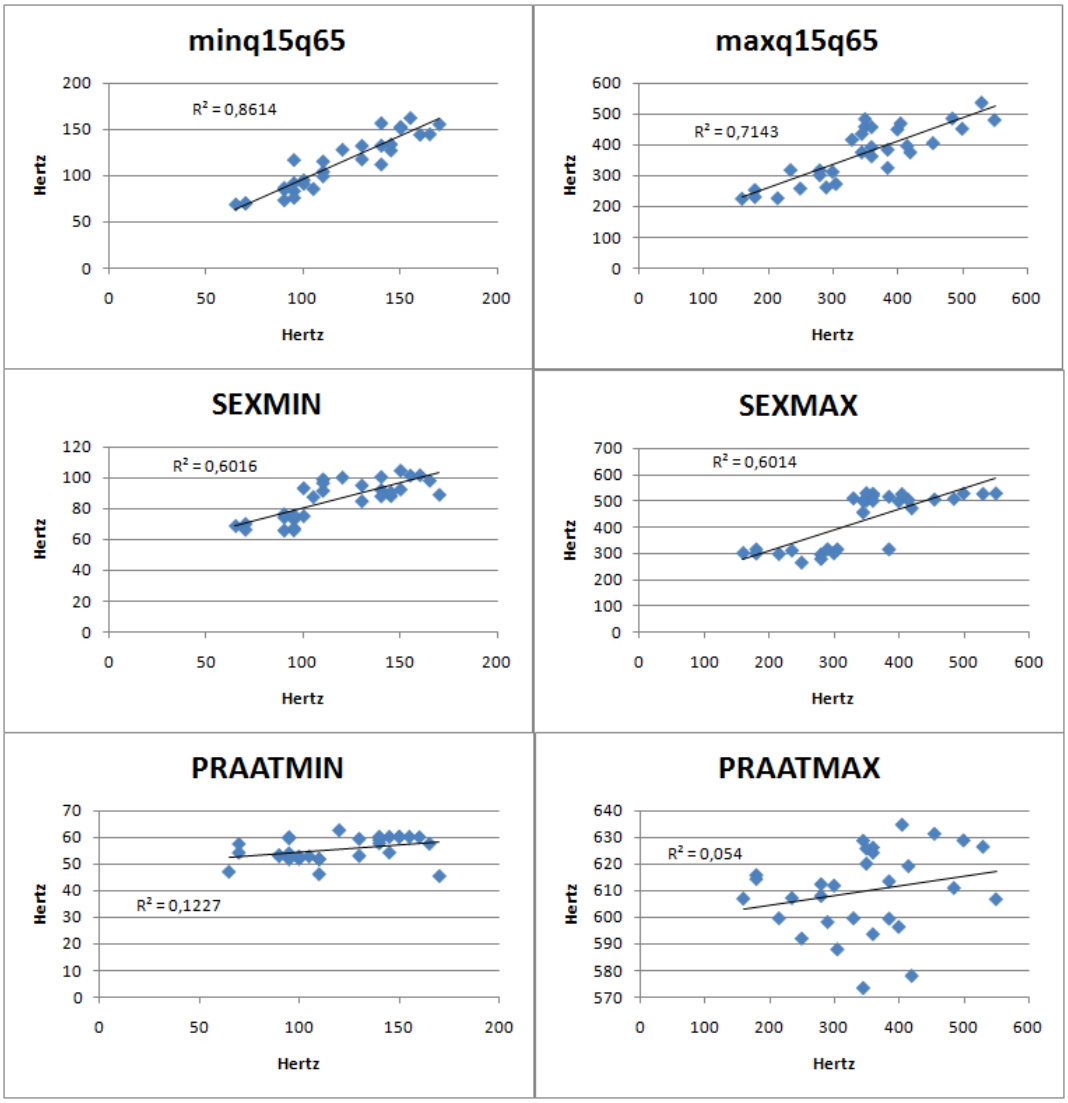


FIGURE 33 – Nuage de points illustrant les corrélations entre valeurs minimales et maximales de référence (MMIN et MMAX) et valeurs minimales et maximales obtenues par ajustement des seuils plancher et plafond aux produits des quantiles q15 et q65 et des rapports 0.84 et 1.92 (minq15q65/ maxq15q65) ; entre MMIN et MMAX et valeurs obtenues par ajustement des seuils en fonctions du sexe du locuteur (SEXMIN/ SEXMAX) ; entre MMIN et MMAX et valeurs obtenues par ajustement des seuils par défaut (PRAATMIN/ PRAATMAX)- Corpus PFC.

2.4 Discussion

Nous avons proposé dans cet exposé une optimisation de la détection des valeurs extrêmes de la f_0, dans le but d'obtenir une mesure fiable du registre. La problématique soulevée en 1 ouvre ainsi la voie à un nouvel ajustement des seuils plancher et plafond dans Praat. Partant de l'hypothèse qu'un ajustement au registre même du locuteur permettrait une meilleure estimation des extrêmes de la f_0, nous avons utilisé les valeurs quantiles afin d'élaguer les valeurs aberrantes qui apparaissent dans une détection erronée de la f_0, et ce, afin de s'approcher des limites du registre du locuteur. Nous avons conclu qu'un ajustement des seuils plancher et plafond aux produits des quantiles 15 et 65 et de leurs coefficients respectifs 0.83 et 1.92 rendaient fiable l'estimation des extrema de la f_0, et ce, bien plus qu'un ajustement manuel au sexe du locuteur. Nous pouvons donc conclure qu'il est à présent possible d'estimer les limites du registre d'un locuteur de façon assez fiable, bien que nous reconnaissions que des abbérations puissent toujours se glisser, notamment dans le cas d'enregistrements (très) bruités.

Maintenant que nous avons traité la problématique de la fiabilité de la mesure de la f_0, nous proposons de nous pencher sur celle du choix du type de mesures à adopter, linguistiques vs. acoustiques, pour le calcul de la hauteur et de l'étendue du registre. Nous rappelons que certaines critiques avaient été avancées à l'utilisation de ces mesures, et que nous avions proposé en 1.5.3 quelques solutions aux problématiques soulevées. Nous avions également suggéré qu'une telle dichotomie (acoustique vs. linguistique) était peut être arbitraire et que ces mesures pouvaient s'avérer complémentaires, si ce n'est comparables. Nous chercherons donc, dans cette section, à répondre à une telle problématique par la comparaison de ces mesures.

3 Du choix du type de mesure : acoustique vs. linguistique

3.1 Choix d'une mesure acoustique

Nous avions vu dans la littérature que plusieurs mesures ont été utilisées pour quantifier la hauteur et l'étendue du registre.

La hauteur est généralement exprimée en termes de moyenne de la f_0 ou médiane de la f_0. La moyenne a été cependant critiquée car elle est sensible aux valeurs aberrantes et parce qu'elle ne peut être « représentative » d'un échantillon lorsque la distribution des données ne suit pas une loi normale. Une solution proposée à cette critique a été apportée au cours du premier chapitre : la transformation des données en log base 2 permettrait de rendre la distribution

à la fois proche d'une distribution normale centrée et réduire le poids des valeurs aberrantes. Pour notre part, nous utilisons la médiane qui se veut plus stable que la moyenne aux valeurs aberrantes et qui permet la division d'un ensemble de données en deux sous-ensembles égaux. Elle est, selon nous, une bonne mesure de hauteur du registre.

L'étendue, quant à elle, a été exprimée en termes de différence des extrema de la f_0, de différence respective des quantiles q90, q95, q97.5 et des quantiles q10, q5, q2.5 ainsi qu'en termes d'écart(s) type(s) autour de la moyenne. La différence entre la valeur maximale et minimale de la f_0 a cependant été critiquée du fait qu'elles pouvaient correspondre à des valeurs aberrantes (résultantes d'une mauvaise détection de la f_0) et parce qu'une telle mesure suppose que la valeur maximale et la valeur minimale sont à égale distance du centre de la distribution. La première partie de ce chapitre offre une possible solution à la problématique des valeurs aberrantes par l'utilisation de seuils spécifiques. Tout calcul de la f_0 sera donc fait une fois ces seuils ajustés. La deuxième critique peut être également solutionnée par une transformation logarithmique des données. Si une telle transformation permet de s'approcher de distributions centrées, nous pourrons utiliser la différence en log base 2 entre les valeurs maximale et minimale de la f_0, comme mesure d'étendue du registre.

3.2 Choix d'une mesure linguistique

Nous avons vu que différentes mesures linguistiques sont utilisées pour décrire la hauteur et de l'étendue du registre. En anglais et en français, la hauteur est mesurée en termes de moyenne des tons bas (Patterson, 2000; Portes & Di Cristo, 2003; T. Rietveld & Vermillion, 2003) ou en termes de moyenne des tons bas finaux (Menn & Boyce, 1982; Liberman & Pierrehumbert, 1984; Patterson, 2000). L'étendue l'est en termes de différence entre la moyenne des tons hauts et la moyenne des tons bas, plus précisément, comme la différence entre la moyenne des pics accentuels non-initiaux et la moyenne des creux post-accentuels (Patterson, 2000; Portes & Di Cristo, 2003) ou encore comme la différence entre la moyenne des pics post-accentuels et la moyenne des creux post-accentuels (I. Mennen et al., 2008).

Nous proposons ici, afin de tester une mesure dite linguistique, d'utiliser le système de notation de l'intonation INTSINT qui permet une obtention des cibles tonales automatique, et ce, quelle que soit la langue donnée. Les possibles mesures de hauteur et d'étendue qui peuvent être effectuées à partir des cibles tonales d'INTSINT (décrit dans la section 1.5.3 du premier chapitre) sont les suivantes : la hauteur peut être mesurée en termes de moyenne des tons médians (M) ou en termes de moyenne des tons bas (B) ; l'étendue peut l'être en termes de différence entre la moyenne des tons hauts et la moyenne des tons bas (T-B). Nous rappelons que les tons T, B et M sont définis de façon absolue, et non relative comme le sont les tons H et L. Ils sont calculés en Hz. Contrairement à ces derniers donc, ils peuvent être utilisés pour

la mesure du registre. Les mesures de hauteur et d'étendue peuvent être également obtenues automatiquement, comme nous l'avons expliqué en 1.5.3, en prenant compte de l'ensemble des points cibles observés, et ce, selon une procédure d'optimisation. Nous utilisons donc ces différentes mesures dans le calcul de la hauteur et d'étendue du registre afin de les comparer aux mesures acoustiques (cf. en 4, un récapitulatif des mesures acoustiques et linguistiques que nous proposons de comparer dans cette section).

	Mesures acoustiques (f_0)	Mesures linguistiques (INTSINT)
Hauteur	Médiane $\log_2(Mediane)$	Moyenne des tons M(Mid) Moyenne des tons B (Bottom) Calcul par optimisation (Key)
Etendue	$\log_2(max/min)$	T - B Calcul par optimisation(Range)

TABLE 4 – Mesures acoustiques et linguistiques utilisées dans le calcul de la hauteur et l'étendue du registre.

3.3 Corpus et base de données

3.3.1 PFC et AIX-MARSEC

Nous avons utilisé pour cette expérience les corpus PFC et AIX-MARSEC décrits en 2.1. Nous avons sélectionné la lecture oralisée des 10 locuteurs du PFC et la production de 53 locuteurs d'AIX-MARSEC, enregistrements dont la qualité permet une étude de la f_0 . Nous avons également étendu notre analyse aux corpus PAC et CID que nous présentons ci-après.

3.3.2 PAC

Le corpus Phonologie de l'Anglais Contemporain, usages, variétés et structures, finalité d'un projet coordonné par J. Durand (Toulouse II & ERSSCNRS) et P. Carr (Montpellier III & ERSS-CNRS), a été mené dans le but de créer une base de données permettant l'analyse comparative des variétés de l'anglais contemporain. Le PAC, soumis au même protocole que le PFC, est ainsi et également représentatif d'un nombre important de locuteurs (hommes et femmes, âgés entre 20 et 70 ans environ, issus de régions diverses du monde anglophone) et de différents types de production (lecture à voix haute d'une liste de mots, lecture d'un passage, conversations guidées et conversations libres).

Pour notre part, nous avons sélectionné 8 locuteurs du Nord de l'Angleterre (Lancashire,

Greater Manchester and West Yorshire), 3 hommes et 5 femmes, agés de 20 à 30 années et avons retenu de leur production la lecture de texte, afin de mener une analyse comparée de nos données pour l'anglais et le français sur un même type de production. Comme pour le PFC, la lecture de texte est en effet une lecture à voix haute d'un passage de type article de journal régional, ne posant aucune difficulté de compréhension.

3.3.3 CID

Le CID, Corpus of Interactional Data, (Bertrand et al., 2007, 2008) est un corpus audio-vidéo de 8 heures, en français, constitué au Laboratoire Parole et Langage, et conçu pour l'analyse multimodale de la langue parlée. L'annotation du CID inclue ainsi la phonétique, la prosodie, la morphologie, la syntaxe, le discours et la mimo-gestualité. Ce corpus s'est avéré avantageux en plusieurs points. Tout d'abord, il relève de la parole authentique. En effet, les sujets participants avaient pour tâche d'évoquer des conflits professionnels ou des situations insolites dans lesquelles ils s'étaient trouvés, résultant en des dialogues riches d'actes de communication. Il est à noté d'ailleurs que les participants pouvaient à tout moment délaisser la consigne qui leur avait été proposée et s'adonner librement à d'autres sujets de conversation. Le CID comprend ainsi de nombreuses séquences de narration, de description, d'argumentation ou d'explication. De plus, le CID nous est paru avantageux de par le temps de parole qu'il représente (8h), une base de données conséquente permettant la conduite d'analyses pertinentes et le développement d'algorithmes performants. Le corpus est d'autant plus intéressant pour nous qu'il offre une transcription orthographique et une annotation phonétique de la parole, ainsi qu'un découpage en unités interpausales et segmentales alignées avec le signal. 16 sujets (10 femmes et 6 hommes) ont donc été enregistrés pour ce corpus. Ils sont tous de langue maternelle française, et issus de diverses régions de France, la moitié d'entre eux étant natifs de la région PACA ou y résidant depuis plus de 20 ans.

Dans notre travail, nous avons sélectionné une partie des données et avons gardé la production de 6 locuteurs (3 hommes et 3 femmes), un total de 30 minutes d'enregistrement.

Nous proposons donc, dans cette analyse, d'étudier les registres de 75 locuteurs. Un tableau synthétique fiche les données en 5.

	AIX-MARSEC	PAC	PFC	CID	TOTAL LOC
LANGUE	51 Anglais	8 Anglais	10 Français	6 Français	75 locuteurs
SEXE	13F, 38H	4F, 4H	6F, 4H	3F, 3H	27F, 48H
STYLE DE PAROLE	Parole authentique	Lectures Oralisées	Lectures Oralisées	Parole authentique	2 styles de parole

TABLE 5 – Informations sur les données utilisées : une synthèse.

Nous proposons tout d'abord, avant de comparer les mesures acoustiques et linguistiques, d'analyser la distribution des données. Cela nous permettra de mieux comprendre la façon dont elles se comportent et ainsi penser à un calcul optimal de hauteur et d'étendue du registre.

3.4 Analyse de la distribution des données

La récolte des données est effectuée à partir d'un script Praat [79]. Pour chaque objet Sound sélectionné des corpus PFC, PAC, CID et Aix-Marsec, est créé un Objet Pitch, dont les seuils plancher et plafond sont ajustés respectivement au produit des valeurs quantiles q15 et du coefficient 0.83 et au produit des valeurs quantiles q65 et du coefficient 1.92. Chaque échantillon de la f_0 (donné en Hz), pour un pas d'analyse à 0.01, est ainsi récupéré et répertorié dans un tableau de données.

La distribution des données est ensuite analysée [80] pour chacun des locuteurs. Pour cela, nous avons mesuré le coefficient de dissymétrie (skewness) qui, lui-même, permet de mesurer l'asymétrie de la densité de probabilité d'une variable aléatoire (la formule est donnée en 2 où $\mu 3$ est le troisième moment centré et σ^3 est l'écart-type).

$$\tau 1 = \mu 3/\sigma^3 \tag{2}$$

Si l'asymétrie d'une distribution est positive, alors la distribution est étalée vers la gauche ; à contrario, si l'asymétrie d'une distribution est négative, alors la distribution est étalée vers la droite. Une distribution proche de 0 suit donc une loi centrée. Pour l'ensemble des locuteurs, le coefficient moyen de dissymétrie mesuré est de 1.03, la valeur maximale étant de 2.35, la valeur minimale de -0.32 (seul coefficient en dessous de 0) ; la distribution générale est donc dans

79. cf. script - CR ROM, dossier SCRIPTS : Get_f0Distribution.
80. cf. Tableau des valeurs - CD ROM, ANNEXES_CHAP2 : Table9.

l'ensemble étalée vers la gauche, ou s'approche d'une distribution normale. Ces résultats, qui vont de pair avec ceux de Liberman et Pierrehumbert (1984) et de Patterson (2000), ne sont pas surprenants puisqu'on sait qu'un locuteur a tendance à parler dans le bas de sa tessiture, se voyant limité par des fréquences basses plancher qu'il ne peut dépasser, ce qui ne semble pas être le cas pour des fréquences plus hautes. Le locuteur ne rejoint pas constamment les mêmes fréquences hautes comme c'est le cas pour les fréquences basses. Par ailleurs, l'asymétrie à gauche est expliquée par le fait que les fréquences ne sont jamais négatives.

Nous avons également mesuré le coefficient d'aplatissement de Pearson (ou kurtosis) afin d'évaluer l'aplatissement ou la pointicité de nos distributions. La formule est donné en 3. Le coefficient d'aplatissement permet en effet de mesurer la disposition des masses de probabilité autour de leur centre. Il est compris entre 1 et $+\infty$; égal à 3, il indique que la distribution suit une loi normale (ou mesokurtique) ; relativement élevé (i.e supérieur à 3), il révèle une distribution pointue en sa moyenne (ou leptokurtique), proche de 1, une distribution aplatie (ou bradykurtique). Nos données relèvent plutôt de distributions pointues en leur moyenne, la moyenne des coefficients d'aplatissement étant de 4.5, le maximum trouvé de 13.15. La valeur minimale est de 2.35 mais elle est unique. 27% des distributions ont en effet un coefficient d'aplatissement proche de 3 et 67% des distributions ont un coefficient supérieur à 3.

$$\beta 2 = \mu 4/\sigma^4 \tag{3}$$

Il apparaît donc clairement que, si l'on souhaite mesurer au mieux l'étendue du registre, il est nécessaire de normaliser les données. Nous pouvons observer dans le tableau 6 que la transformation des données en log base 2 rend en effet les distributions proches de distributions normalisées ; le coefficient moyen de dissymétrie passe de 1.06 à 0.46, le coefficient moyen d'aplatissement de 4.51 à 3.09. Si nous procédons au test de normalité Shapiro-Wilk, afin de tester l'hypothèse selon laquelle nos données logarithmiques suivent une loi normale, nous devons cependant constater que nos données transformées ne suivent pas une distribution « parfaitement » gaussienne, la probabilité critique associée au test étant inférieure à 5% (p-value=6.61e-15). Cependant, parce que les distributions normalisées sont plus centrées que les distributions brutes (cf. comparaison des figures 34 et 35), nous pouvons penser qu'une mesure de l'étendue du registre par la différence entre la valeur maximale et la valeur minimale est assez fiable. De plus, la normalisation des données en log base 2 permet de passer d'une échelle en Hz à une échelle en octave, une échelle qui rend mieux compte des différences d'étendue de registre entre individus (point soulevé en 1.5.2). La médiane est également présupposée une bonne mesure, du fait qu'elle divise la distribution en masses égales. Une telle mesure, non paramétrique, peut donc se faire à partir des valeurs en Hz ou normalisées (log(Hz)).

	KHZ	KLHZ	SHZ	SLHZ
Moyenne	4,51	3,09	1,03	0,46
Ecart Type	1,78	0,81	0,43	0,37
Minimum	2,35	2,15	-0,33	-0,93
Maximum	13,15	7,26	2,36	1,43

TABLE 6 – Moyennes, écart types, valeurs minimales et maximales des coefficients de dissymétrie (S) et d'applatissement (K) calculés à partir de données en Hz (HZ) et normalisées (L).

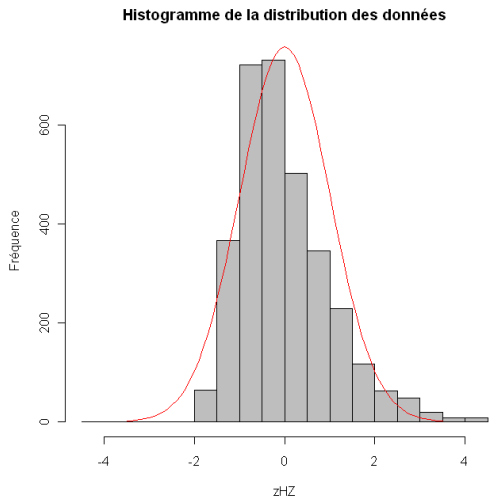


FIGURE 34 – Histogramme de la distribution des données en Hz (locuteur J0201G) en comparaison à une distribution normale ; les données sont transformées en z-score.

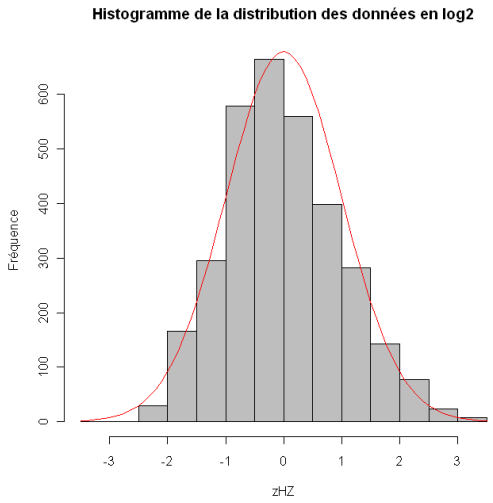


FIGURE 35 – Histogramme de la distribution des données transformées en log base 2 (locuteur J0201G) en comparaison à une distribution normale ; les données sont transformées en z-score.

La distribution des données à présent analysée, nous proposons de comparer les résultats de hauteur et d'étendue du registre, obtenus à partir de mesures acoustiques et de mesures linguistiques dans le but de tester la pertinence de la dichotomie « acoustique vs. linguistique » qui s'impose aujourd'hui dans la littérature.

3.5 Comparaison des données en fonction du type de mesure utilisé

Nous rappelons, dans un premier temps, que la hauteur est obtenue par le calcul de la médiane (MED), la moyenne des tons M (MEANM), la moyenne des tons B (MEANB) et par le calcul d'optimisation d'INTSINT (KEY) (en Hz et en log de base 2) ; l'étendue l'est à partir du calcul de la différence entre le maximum et le minimum en log de base 2 (LOGDMM), la différence entre la moyenne des tons T et la moyenne des tons B (LOGDTB) et par le calcul d'optimisation d'INTSINT (RANGE), également transformés en log de base 2 [81].

Nous procédons à une régression simple afin d'évaluer la relation entre les mesures acoustiques et linguistiques. Nous testons l'hypothèse selon laquelle la relation entre ces variables est linéaire ; si l'hypothèse est avérée, nous pourrons conclure que les mesures acoustiques et linguistiques s'équivalent ; au contraire, si la relation est non-linéaire, nous pourrons penser que ces mesures n'estiment pas le registre de façon comparable, et que seul, un test de perception permettrait de valider quelle mesure permet l'estimation du registre d'un locuteur.

81. cf. Tableau des valeurs obtenues CD ROM - ANNEXES_CHAP2 : Table10.

Nous observons tout d'abord le degré de corrélation entre la médiane (MED) et la moyenne des tons M (MEANM) à partir du nuage de points donné en figure 36. Au vu de cette représentation graphique, la relation entre MED et MEANM semble linéaire. L'ajustement d'un modèle linéaire corrobore l'intuition de la relation linéaire, la valeur du coefficient de détermination étant proche de 1 (R^2=0.9259). Les fortes corrélations entre la médiane (MED) et la moyenne des tons B (MEANB) et MED et la hauteur du registre obtenue par le calcul d'optimisation d'INTSINT (KEY) ($R^2 = 0.9218$ et $R^2 = 0.9475$) valident également l'hypothèse d'une relation linéaire entre les variables acoustiques et les variables linguistiques. Nous pouvons donc conclure que la médiane, la moyenne des cibles tonales M et la moyenne des cibles tonales B sont aussi « efficaces » les unes que les autres dans la mesure de la hauteur du registre.

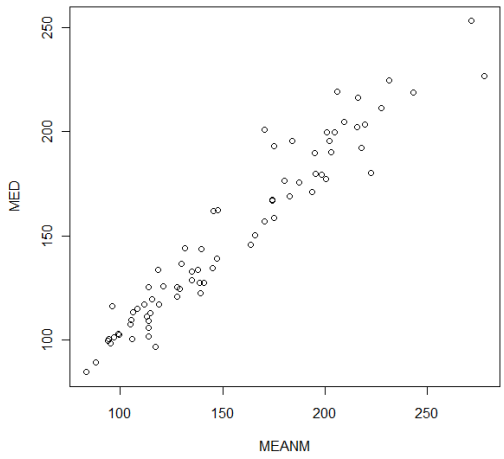


FIGURE 36 – Représentation graphique des valeurs obtenues de hauteur du registre à partir du calcul de la médiane en Hz (MED) en fonction du calcul de la moyenne des cibles tonales M en Hz (MEANM).

Nous observons à présent le degré de corrélation entre la différence entre les valeurs maximale et minimale (LOGDMM) et la différence entre la moyenne des tons T et la moyenne des tons B (LOGDTB) à partir du nuage de points donné en 37. Contrairement aux mesures de hauteur, la mesure acoustique LOGDMM et la mesure linguistique LOGDTB ne semblent pas corrélées. Le modèle linéaire confirme l'intuition de la représentation graphique par un coefficient de détermination proche de 0 ($R^2 = 0.1393$). La tendance est d'ailleurs la même lorsque nous comparons LOGDMM et l'étendue du registre obtenue par le calcul d'optimisation d'INTSINT (RANGE), avec un R^2 à 0.08627.

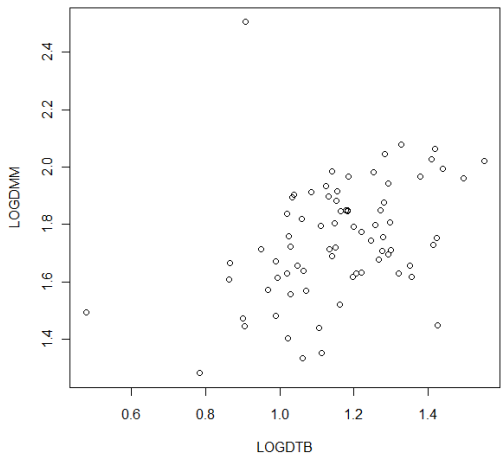


FIGURE 37 – Représentation graphique des valeurs obtenues d'étendue du registre à partir du calcul de la différence entre la valeur maximale et la valeur minimale (transformée en log de base 2) en fonction du calcul de la différence entre la moyenne des cibles tonales T et la moyenne des cibles tonales B (transformée en log de base 2).

Peut donc se poser la question de la raison d'une telle différence. La réponse apparaît très clairement. Alors que la mesure acoustique se base sur *la* valeur maximale et *la* valeur minimale de la distribution pour rendre compte de l'étendue du registre, la mesure linguistique LOGDTB, quant à elle, se base sur une moyennisation *de l'ensemble* des cibles tonales hautes et sur une moyennisation *de l'ensemble* des cibles tonales basses ; la mesure linguistique RANGE est obtenue par optimisation et prend donc également en compte *l'ensemble* des cibles tonales et non *deux* cibles tonales extrêmes. Les valeurs obtenues par calcul de LOGDMM sont donc généralement supérieures à celles obtenues par calcul de LOGDTB et RANGE (cf. figure 38). Reste à savoir maintenant si l'auditeur base son jugement perceptif du registre à partir de deux valeurs extrêmes ou à partir d'un ensemble de valeurs hautes et de valeurs basses, seul un test de perception permettrait de répondre à cette question.

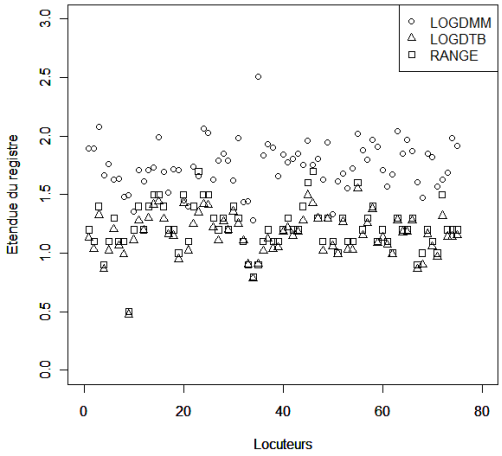


FIGURE 38 – Représentation graphique des valeurs obtenues d'étendue du registre à partir du calcul de la différence entre la valeur maximale et la valeur minimale (transformée en log de base 2 ; LOGDMM), du calcul de la différence entre la moyenne des cibles tonales T et la moyenne des cibles tonales B (transformée en log de base 2 ; LOGDTB) et du calcul par optimisation RANGE (donnée en log de base 2).

Il est tout particulièrement intéressant de noter ici la forte corrélation entre la médiane (MED) et la moyenne des cibles tonales B. Si une telle corrélation est également obtenue entre la médiane (MED) et la moyenne des cibles tonales T (MEANT), nous pourrions proposer une autre mesure d'étendue du registre, une mesure qui résoudrait ainsi la dichotomie acoustique vs. linguistique.

Si nous observons le degré de corrélation entre la variable MED et la variable MEANT, à partir du nuage de points donné en 39, nous pouvons aussi penser à une relation linéaire entre les deux variables. L'ajustement d'un modèle linéaire corrobore notre intuition, avec un coefficient de détermination proche de 1 (R^2=0.9152).

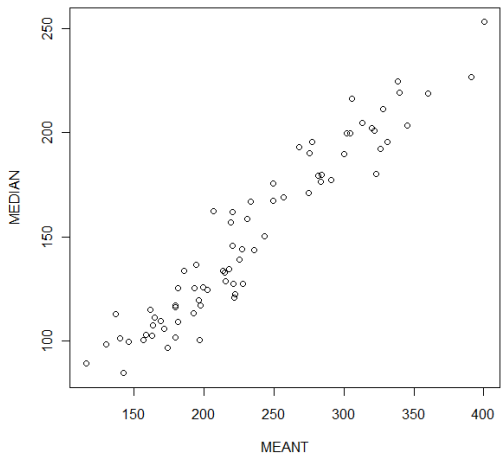


FIGURE 39 – Représentation graphique des valeurs obtenues de hauteur du registre à partir du calcul de la médiane en Hz (MED) en fonction du calcul de la moyenne des cibles tonales T en Hz (MEANT).

Il est donc possible, à partir de la médiane, de prédire les limites du registre d'un locuteur et ainsi son étendue. Nous procédons à une régression linéaire afin d'obtenir les coefficients de de régression et les intercepts nous permettant la prédiction de la moyenne des tons B (MEANB) et la moyenne des tons T (MEANT). MEANB est obtenue par la relation affine donnée en 4.

$$MEANB = 0.741 \times MED - 5.52 \tag{4}$$

Le test de significativité du coefficient de régression donne ici une probabilité critique inférieure à 2e-16. Parce que la probabilité critique de l'intercept est estimée à 0.161, donc non significative, nous proposons un ajustement du modèle sans la constante. La nouvelle relation affine est donnée en 5. Le coefficient de détermination du modèle est également proche de 1 (R^2=0.9934).

$$MEANB = 0.706 \times MED \tag{5}$$

MEANT, quant à elle, est obtenue par la relation affine donnée en 6.

$$MEANT = 1.537 \times MED + 3.75 \tag{6}$$

Le test de significativité du coefficient de régression donne également ici une probabilité critique inférieure à 2e-16. La probabilité critique de l'intercept étant également non significative (p = 0.659), nous effectuons à nouveau un ajustement du modèle sans la constante. La nouvelle relation affine est donnée en 7. Le coefficient de détermination du modèle est également proche de 1 (R^2=0.9937).

$$MEANT = 1.561 \times MED \tag{7}$$

Nous pouvons nous demander si une tranformation en log de base 2 permettrait l'amélioration de la robustesse de ces modèles. Nous procédons ainsi à nouveau à des régressions linéaires dont les relations affines sont données ci-après. Les coefficients de détermination de ces différents modèles sont indiqués entre parenthèses. Bien que la probabilité critique des intercepts soit significative (p=0.00178 et p=0.00567), les relations affines sans constante sont aussi données.

$$log_2(MEANB) = 1.038 \times log_2(MED) - 0.79(R^2 = 0.9264) \tag{8}$$

$$log_2(MEANB) = 0.928 \times log_2(MED)(R^2 = 0.9997) \tag{9}$$

$$log_2(MEANT) = 0.981 \times log_2(MED) + 0.78(R^2 = 0.8999) \tag{10}$$

$$log_2(MEANT) = 1.089 \times log_2(MED)(R^2 = 0.9997) \tag{11}$$

Si les coefficients de détermination obtenus pour les différents modèles de régression linéaire présentés en a1, a2, b1, b2 et c révèlent une certaine robustesse des modèles, ils ne permettent pas pour autant de les comparer. Pour cela, il existe plusieurs critères que l'on peut utiliser, tels que le critère d'information bayésien (BIC) ou encore le critère d'information d'Akaike (AIC) mais ils sont plutôt utilisés pour comparer des modèles emboîtés (Baayen, 2008). Si l'AIC est également estimé dans le cas de modèles indépendants, il ne paraît cependant pas utilisable dans notre cas où les différences de modèle sont simplement dues à une transformation des données. Nous proposons donc d'évaluer nos modèles sur la base de leurs résidus. Les figures 40 et 41 présentent les résidus des 4 modèles proposés pour la prédiction de la moyenne des tons B (MEANB), les figures 42 et 43, ceux des 4 modèles proposés pour la prédiction de la moyenne des tons T (MEANT). D'un côté, une transformation en log ne semble pas apporter d'amélioration à la robustesse des modèles. D'un autre côté, l'ajustement sans constante ne

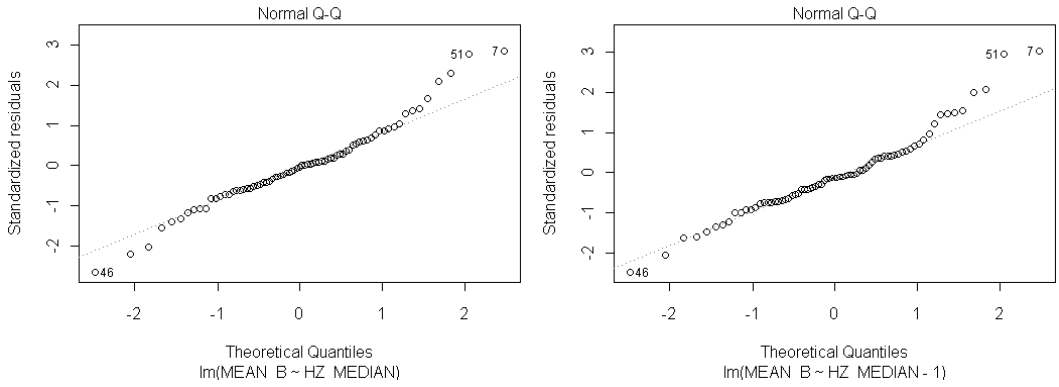


FIGURE 40 – Représentation graphique des résidus du modèle linéaire avec constante (à gauche) et sans constante (à droite) pour la prédiction de MEANB.

diminue pas la robustesse des modèles. On peut donc opter pour les modèles les plus simples, notamment en vertue du rasoir d'Ockham, principe de parcimonie selon lequel lorsque deux théories sont vraisemblables, on favorise celle qui est la plus simple ou la plus parcimonieuse.

Parce que nous cherchons à exprimer l'étendue en octave, nous retiendrons pour l'estimation de MEANB et MEANT les relations affines suivantes, définies en 9 et 11 :

$log_2(MEANB) = 0.928 \times log_2(MED)$, et

$log_2(MEANT) = 1.089 \times log_2(MED)$

Il est à noter que nous prenons les modèles les plus simples, sans constante, puisque cette dernière ne permet pas une amélioration du modèle.

L'étendue du registre (E), à partir de ces mesures, peut donc être calulée comme suit (12) :

$$E = (1.089 - 0.928) \times log_2(MED),\ soit\ E = 0.161 \times log_2(MED) \tag{12}$$

Il est cependant important, avant de valider une telle mesure, de s'assurer qu'il n'existe pas d'interaction avec le sexe du locuteur, la langue ou le type de production. On peut en effet se demander si les pentes des régressions linéaires en fonction des niveaux de chaque facteur (sexe : Homme/Femme ; langue : Anglais/Français ; type de production : lecture/ parole authentique) sont significativement différentes (un exemple est donné en 44 pour la prédiction de la moyenne des tons T (MEANT) en fonction du sexe du locuteur). Nous procédons donc à une ANOVA à deux facteurs pour chacun de ces facteurs. Les ANOVAs effectuées pour la prédiction de MEANT révèlent qu'il n'y a pas d'effet de sexe (pval=0.0917), ni de langue (pval=0.170), ni de

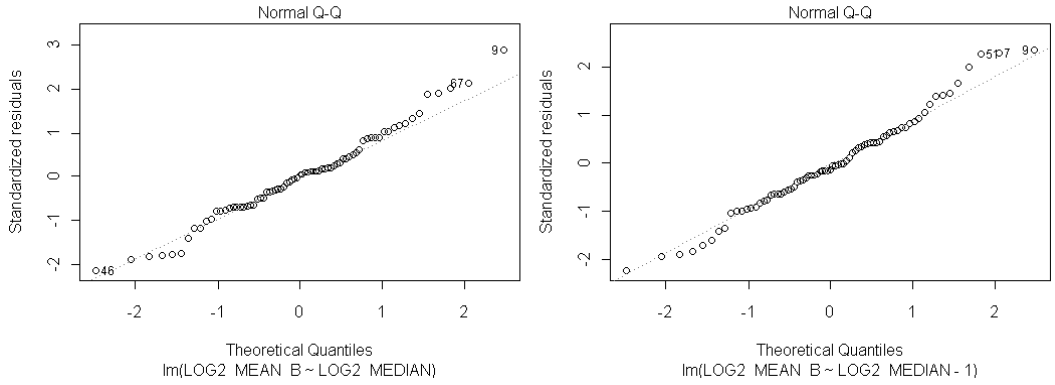


FIGURE 41 – Représentation graphique des résidus du modèle linéaire avec constante (à gauche) et sans constante (à droite) pour la prédiction de MEANB lorsque transformées en log2.

type de production (pval=0.134) ; les ANOVAs effectuées pour la prédiction de la moyenne des tons B (MEANB) rejettent également tout effet de sexe du locuteur (pval=0.381), de langue (pval= 0.274) ou encore de type de production (pval=0.368).

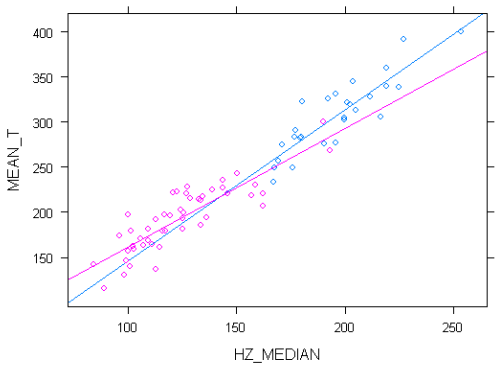


FIGURE 44 – Représentation graphique des régressions linéaires pour la prédiction de MEANT en fonction de médiane et du sexe du locuteur ; les cercles et la régression en bleu correspondent au niveau femme, les cercles et la régression en rose au niveau homme.

3.6 Discussion

Suite à ces constats, la problématique du choix d'une mesure adéquate, ancrée dans la dichotomie « linguistique vs. acoustique », telle que énoncée par Patterson et Ladd (1999) et

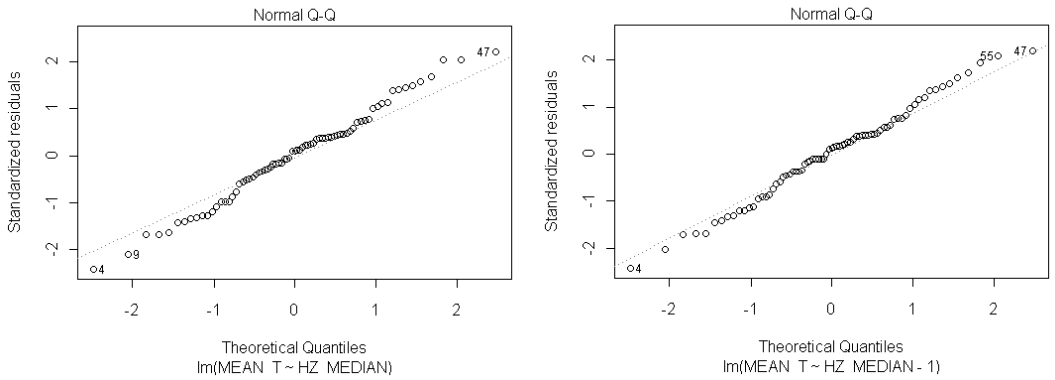


FIGURE 42 – Représentation graphique des résidus du modèle linéaire avec constante (à gauche) et sans constante (à droite) pour la prédiction de MEANT.

Patterson (2000) et reprise par de nombreux auteurs, est sans réel fondement. Le calcul de la hauteur ou de l'étendue du registre à partir de mesures acoustiques ou linguistiques, est sensiblement analogue, puisque les mesures acoustiques et linguistiques sont corrélées entre elles. Bien que nous ayons vu que la différence entre la valeur maximale et minimale de la distribution n'était pas corrélée à la différence entre la moyenne des cibles tonales hautes et la moyenne des cibles tonales basses, un fait pour lequel nous avons d'ailleurs donné des explications, nous avons montré que la différence entre la moyenne des valeurs hautes et la moyenne des valeurs basses pouvait être obtenue à partir de la médiane. Pour notre part, donc, nous choisissons la médiane pour le calcul de la hauteur du registre, et le log2(MEANT/MEANB) pour le calcul de l'étendue.

Outre le fait qu'elles permettent un nouveau calcul de l'étendue du registre, les fortes corrélations obtenues entre la médiane et la moyenne des cibles basses et entre la médiane et la moyenne des cibles hautes, soulèvent un point important, celui de la corrélation entre la hauteur du registre et son étendue. Au vu de cette corrélation, il peut donc être conclu que plus le registre est haut, plus il est étendu, et vice-versa ; plus le registre est bas, plus il est étroit, un fait déjà soulevé par Ladd (1996, p260), dans sa définition du registre. L'auteur expliquait en effet que la difficulté d'admettre deux dimensions au registre venait du fait que ces deux dimensions co-varient.

Enfin, ces corrélations offrent de nouvelles perspectives d'ajustement des seuils plancher et plafond. En effet, s'il est possible de déterminer les limites du registre à partir de la médiane, il devient donc intéressant d'appliquer de telles formules aux seuils plancher et plafond tels que décrits en 2. Nous proposons donc une nouvelle analyse sur ce point.

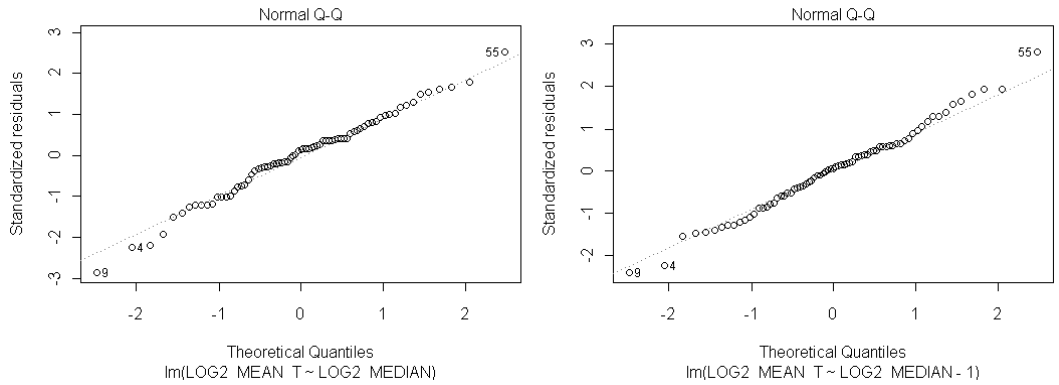


FIGURE 43 – Représentation graphique des résidus du modèle linéaire avec constante (à gauche) et sans constante (à droite) pour la prédiction de MEANT lorsque transformées en log2.

4 Vers un nouvel ajustement des seuils plancher et plafond

4.1 Registre et intervalles musicaux : l'octave

Nous avons vu dans la section précédente que la médiane (MED) est fortement corrélée à la moyenne des cibles tonales hautes (MEANT) et la moyenne des cibles tonales basses (MEANB). Nous rappelons les relations affines définies en 5 et 7 :

$MEANB = 0.706 \times MED$

$MEANT = 1.561 \times MED$

Il est intéressant de noter que le coefficient 0.706 correspond exactement à une demi-octave ($\log_2(0.706) = -0.5$) et que le coefficient 1.561 est à peine supérieur à une demi-octave ($\log_2(1.560) = 0.6$). Nous pouvons donc dire que MEANB et MEANT se trouvent plus ou moins à une demi-octave de la médiane. Nous proposons à travers la figure 45 une représentation graphique de MEANB et MEANT en fonction de la médiane. Les régressions linéaires correspondantes sont tracées en lignes continues et les lignes en pointillés représentent les intervalles +octave, +demi-octave, unison, -demi-octave et -octave par rapport à la médiane. La régression linéaire pour MEANB se confond ainsi à l'intervalle -demi-octave. Au vu de ce graphique, nous pouvons donc dire que MEANB se situe clairement à une demi-octave de la médiane (unison), MEANT entre l'intervalle demi-octave et l'octave.

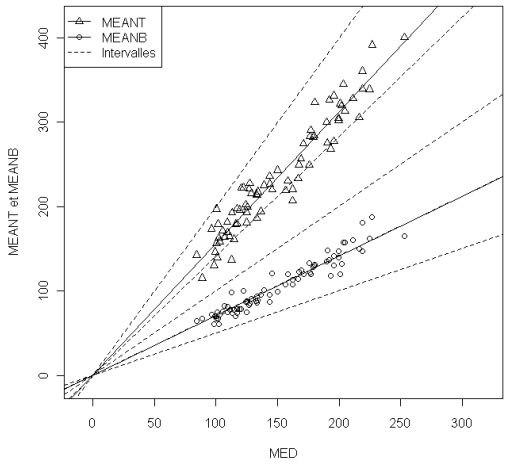


FIGURE 45 – Représentation graphique de MEANB et MEANT en fonction de la médiane ; en lignes continues, les régressions linéaires correspondantes ; en lignes pointillées, les intervalles +octave, +demi-octave, unison, -demi-octave et -octave par rapport à la médiane.

4.2 Ajustements des seuils à l'octave

Ces intervalles musicaux présentent donc un intérêt certain puisqu'ils peuvent être utilisés respectivement dans l'ajustement des seuils plancher et plafond. Nous proposons donc de comparer les valeurs maximales et minimales annotées manuellement (MMIN et MMAX), décrites en 2.2.1, aux valeurs maximales et minimales obtenues après ajustement des seuils plancher et plafond à l'octave et à la demi-octave. Nous testons donc deux combinaisons, l'une dont les seuils plancher et plafond sont ajustés à la demi-octave par rapport à la médiane, l'autre dont les seuils plancher et plafond sont ajustés à l'octave par rapport à la médiane. L'analyse est effectuée sur l'ensemble des fichiers son utilisés en 2.2.2 et 2.3, soit 53 locuteurs du corpus AIX-MARSEC et 10 locuteurs du PFC (sous trois types de production). Afin de comparer les valeurs minimales et maximales obtenues à partir des nouveaux ajustements aux valeurs de référence (MMIN et MMAX), on observe, par la formule décrite en 2.2.3, la distance moyenne entre les valeurs obtenues automatiquement et les valeurs de référence. Les distances obtenues pour les ajustements à l'octave et à la demi-octave sont comparées à celle obtenue pour un ajustement aux produits des quantiles q15 et q65 et de leurs coefficients respectifs. Le tableau 7 donne les distances obtenues pour les valeurs minimales et maximales, selon ces 3 ajustements. Nous observons que les distances obtenues pour les minima avec ajustement du plancher au produit de q15 et du coefficient 0.83 et avec ajustement à la demi-octave sont assez proches ; la distance obtenue avec ajustement à l'octave est, quant à elle, supérieure

aux deux distances ; les distances obtenues pour les maxima, avec ajustement du plafond au produit de q65 et du coefficient 1.92 et avec ajustement à l'octave, sont très proches ; à contrario, la distance avec ajustement à la demi-octave est supérieure aux deux distances. Il apparaît donc clairement, à partir de ce tableau, qu'un ajustement du plancher à la demi-octave convient mieux à l'estimation des minima qu'un ajustement à l'octave ; à contrario, l'estimation des maxima est meilleure lorsque le plafond est ajusté à l'octave supérieure que lorsqu'il est ajusté à la demi-octave supérieure. Les coefficients de corrélation entre les valeurs de référence et les valeurs prédites confirment d'ailleurs de tels résultats : pour les minima, le coefficient est de 0.85 lorsque le plancher est ajusté à l'octave, contre 0.90 lorsque ajusté à la demi-octave ; le coefficient est par contre de 0.80 pour les maxima lorsque le plafond est ajusté à la demi-octave contre 0.87 lorsque ajusté à l'octave. Nous retiendrons donc la demi-octave inférieure (par rapport à la médiane) pour l'ajustement du seuil plancher, l'octave supérieure pour l'ajustement du seuil plafond.

Ajustements	Min	Max
plancher=0.83*q15 plafond=1.92*q65	0.94	0.89
plancher= -Octave plafond= +Octave	0.85	**0.87**
plancher=$\frac{-Octave}{2}$ plafond=$\frac{+Octave}{2}$	**0.90**	0.80
plancher=$\frac{-Octave}{2}$ plafond= +Octave	**0.90**	**0.87**

TABLE 7 – Tableau synthétique des ajustements utilisés pour la détection des valeurs extrêmes (minimale et maximale) de la f_0 et des coefficients de corrélation obtenus pour les minima et les maxima en fonction de ces ajustements.

4.3 Discussion

Nous avons proposé, dans cette section, un nouvel ajustement des seuils plancher et plafond à partir de l'octave et de la demi-octave par rapport à la médiane. Il est cependant à noter que dès lors qu'on mesure la hauteur et l'étendue du registre à partir de la médiane, il n'est pas nécessaire d'ajuster ces seuils plancher et plafond du fait que la médiane est non paramétrique et ne tient pas compte des valeurs aberrantes. En revanche, si on veut maximiser la qualité de détection de la f_0, notamment dans l'étude et la détection automatique de ses extrema, et ainsi réduire le nombre de valeurs aberrantes, il est nécessaire d'ajuster ces seuils.

Le fait que le registre semble compris entre l'octave supérieure et la demi-octave inférieure

nous mène aussi à envisager, outre un ajustement des seuils, une nouvelle échelle de mesure. En effet, nous pourrions envisager une visualisation de la f_0 centrée sur sa valeur médiane et dont les variations seraient délimitées par une ligne plancher qui correspond à la demi-octave inférieure par rapport à la médiane et une ligne plafond qui correspond à l'octave supérieure par rapport à la médiane. A partir d'une telle échelle de mesure, i.e. à partir d'une plage tonale définie, les variations de registre pourraient être plus clairement appréhendées. Ces seuils plancher et plafond justifient par ailleurs la représentation graphique des variations de registre que nous avions donnée dans le premier chapitre de cette thèse, où nous décrivions les limites de la plage tonale comme des limites constantes au sein desquelles varie le registre.

Puisqu'il est à présent possible de mesurer le registre de façon fiable, nous proposons, pour finaliser ce chapitre, une étude succincte des registres des locuteurs des corpus PAC, PFC, AIX-MARSEC et CID, afin de présenter brièvement nos échantillons de données. Nous pourrons ainsi apprécier les ressemblances et divergences qui se dégagent de nos corpus par rapport aux diverses conceptions des fonctions extra-linguistiques des variations de registre.

5 Registre et fonctions extra-linguistiques : Une comparaison en fonction du sexe, de l'origine géographique et du type de production.

5.1 Rappel : les fonctions extra-linguistiques du registre

Nous avons vu au cours de notre introduction que le registre caractérise l'individualité du locuteur (sexe, âge, aspect physique, état de santé, personnalité, statut socioprofessionnel, origine géographique et ethnique du locuteur) et le style discursif qu'il pratique. Pour notre part, nous nous intéressons, dans cette section, aux différences de registre en fonction du sexe et de l'origine géographique du locuteur et en fonction du type de parole (lue vs. authentique).

S'il est clairement établi que le registre est plus haut chez les hommes que chez les femmes (résultant de leurs caratéristiques physiologiques), la question de son étendue ne fait pas consensus. Alors que certains auteurs démontrent que l'étendue du registre des femmes couvre un champ plus large que celle des hommes, d'autres attestent d'une stéréotypisation de ce trait et expliquent que la gamme tonale utilisée par les femmes n'est pas plus étendue que celle des hommes. Une telle différence résulterait notamment de l'échelle de mesure utilisée. Pour notre part, nous avons pu démontrer, dans la section précédente, que la hauteur du registre était fortement corrélée à son étendue. Nous présenterons donc, dans une première partie, les résultats obtenus à partir de nos échantillons de données.

On peut aussi lire que le registre varie en fonction de l'origine géographique du locuteur. Demers (2000) avait rapporté, par exemple, que les hommes québéquois ont un registre plus bas et plus réduit que les hommes français ; S. F. Mennen I. et Docherty (2008) avaient également montré que les femmes anglaises ont un registre plus étendu que celui des femmes allemandes. Nous chercherons donc à voir, dans une deuxième partie, si les locuteurs anglais des corpus PAC et Aix-Marsec ont des registres significativement différents des locuteurs français des corpus PFC et CID.

Dans la littérature, les différences de registre sont aussi observées selon le style de parole (parole préparée/ spontanée, parole formelle/ informelle, monologale/ conversationnelle) : pour les uns, les paroles spontanées et dialogiques seraient caractérisées par un registre plus haut et plus étendu que les paroles préparées et monologales ; pour les autres, les locuteurs utiliseraient un registre plus bas et plus réduit en conversation qu'en monologue joué (*acted monologue*). Nous avions suggéré qu'une telle différence ne reflétait pas des résultats contradictoires mais plutôt laisser envisager que les différences de registre dépendent non seulement du caractère spontané vs. préparé de la parole mais également de son caractère situationnel. Le caractère plus ou moins exalté du locuteur, plus ou moins formel, participerait également à des différences de registre inter-locuteurs. Ayers (1994) suggérait également que ces différences pouvaient refléter l'organisation informationnelle et la dimension hiérarchique propres au style de parole pratiqué. Nous chercherons donc, dans une troisième partie, à partir des corpus PFC et PAC d'un côté, et à partir des corpus Aix-Marsec et CID d'un autre côté, à estimer les éventuelles différences de registre en fonction du style de parole, i.e. lue vs. authentique, respectivement [82].

5.2 Registre et sexe du locuteur

Nous proposons tout d'abord une présentation des différences de registre en fonction du sexe du locuteur. Les boîtes à moustaches [83] de la figure 46 représentent les dispersions de la hauteur (KEY) et de l'étendue du registre (RANGE) en fonction de la variable SEXE. Au vu de ce graphique, il apparaît clairement un effet SEXE. Nous observons que la hauteur du registre se situe aux alentours des 200 Hz pour les femmes, 125 Hz pour les hommes ; l'étendue varie entre 1.2 et 1.25 pour les femmes, entre 1.1 et 1.15 pour les hommes. L'analyse de variance ANOVA permet de conclure à la significativité du facteur SEXE aussi bien pour KEY ($F(1,95)=228$, p-value : $<$ 2.2e-16) que pour RANGE ($F(1,95)=209.9$, p-value : $<$ 2.2e-16).

82. Le Tableau des valeurs à partir desquelles les analyses sont effectuées est donné sur CD ROM - ANNEXES_CHAP2 : Table11.

83. Une boîte à moustaches ou *boxplot* ou encore *diagramme de Tukey* permet de figurer une série statistique quantitative. Les extrémités du rectangle représentent le premier et le troisième quartile de la distribution. La ligne qui coupe le rectangle en deux représente la médiane. Au rectangle sont ajoutés deux segments (moustaches) qui représentent les valeurs égales à 1.5 fois le quartile. Les points au-delà des moustaches révèlent l'asymétrie de la distribution (Baayen, 2008).

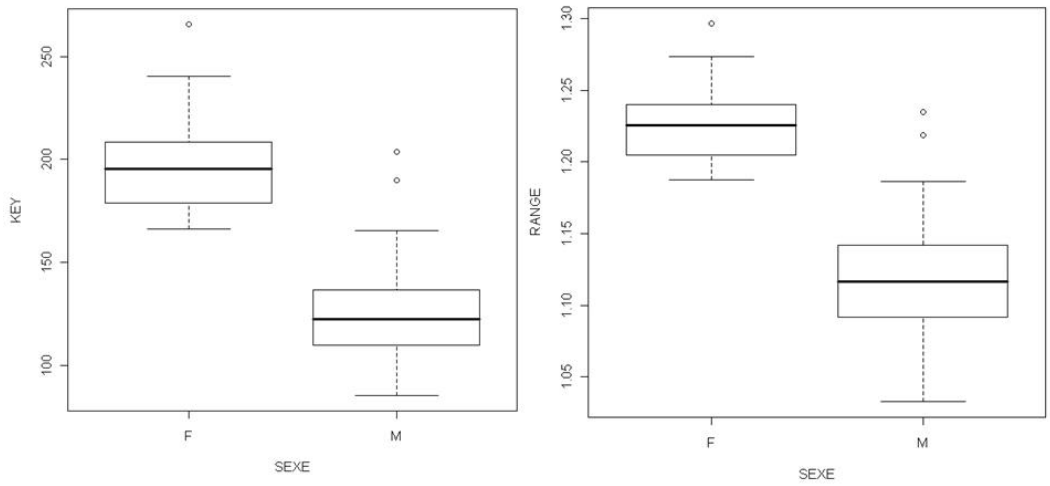


FIGURE 46 – Boîtes à moustaches de la hauteur du registre (KEY ; à gauche) et de l'étendue du registre (RANGE ; à droite) selon les modalités de la variable SEXE. F représente les femmes, H les hommes. Les données sont représentées en Hz pour le KEY, en octave pour le RANGE.

5.3 Registre et langue

Nous présentons à présent les différences de registre en fonction de la langue (LANG). La figure 47 représente les dispersions de la hauteur (KEY) et de l'étendue du registre (RANGE) en fonction de la variable LANG. Au vu de ce graphique, nous pouvons nous poser la question d'un éventuel effet LANG. La hauteur du registre serait plus basse chez les locuteurs anglais que chez les locuteurs français. L'étendue, puisque corrélée à la hauteur, serait du coup plus étroite chez les anglais que chez les français. L'analyse de variance montre une significativité faible du facteur LANG aussi bien pour KEY ($F(1,95)=9.593$, p-value=0.00259) que pour RANGE ($F(1,95)=9.984$, p-value=0.002118).

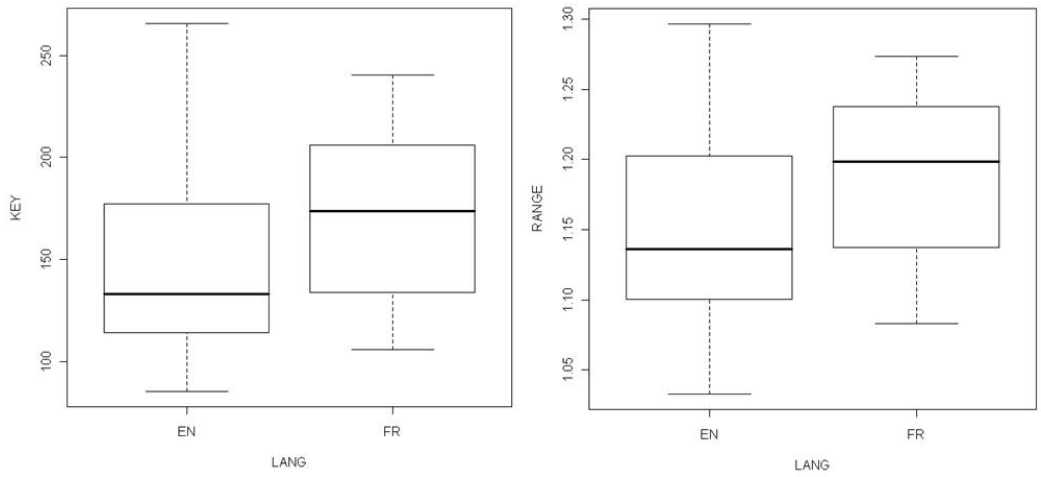


FIGURE 47 – Boîtes à moustaches de la hauteur du registre (KEY ; à gauche) et de l'étendue du registre (RANGE ; à droite) selon les modalités de la variable LANG. EN représente les locuteurs anglophones, FR les locuteurs francophones. Les données sont représentées en Hz pour le KEY et en octave pour le RANGE.

Nous devons cependant rappeler que nos données ne sont pas parfaitement équilibrées. Si l'on comptabilise le nombre d'hommes et de femmes pour les deux corpus anglais d'un côté et les deux corpus français d'un autre côté, nous voyons qu'il y a à 70% d'hommes contre 30% de femmes dans les corpus anglais, alors que les hommes représentent 45% et les femmes 55% dans les corpus français. Nous pouvons donc penser que l'éventuel effet LANG est en fait celui du SEXE. Si l'on équilibre à présent les données (13 femmes et 13 hommes pour les corpus anglais ; 7 femmes et 7 hommes pour les corpus français), l'effet LANG aussi bien pour KEY que pour RANGE n'est plus significatif (KEY : $F(1,95)=0.4742$, p-value$=0.494$; RANGE : $F(1,95)=0.5004$, p-value$=0.4827$). On le voit d'ailleurs très bien sur la représentation graphique donnée en 48. Au vu de nos données donc, il n y a d'effet de la langue parlée sur la hauteur et l'étendue du registre.

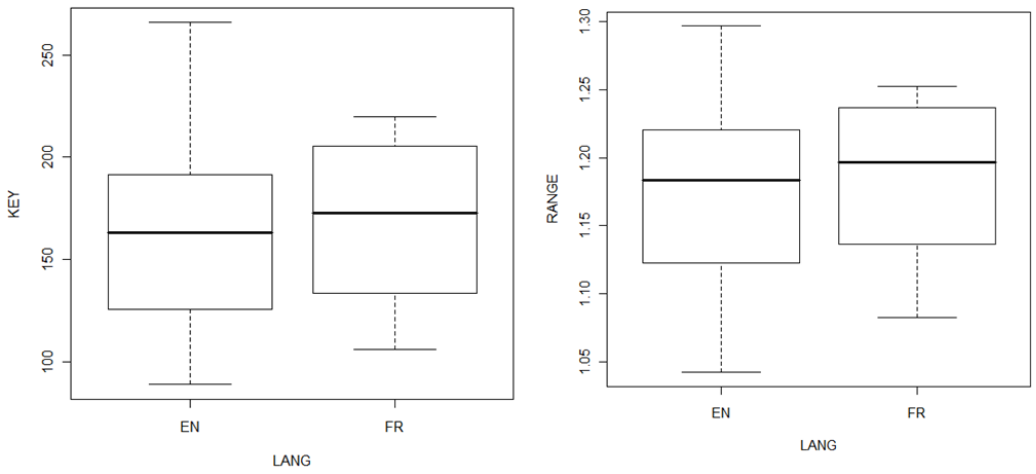


FIGURE 48 – Boîtes à moustaches de la hauteur du registre (KEY ; à gauche) et de l'étendue du registre (RANGE ; à droite) selon les modalités de la variable LANG, établies à partir d'un jeu de données équilibré. EN représente les locuteurs anglophones, FR les locuteurs francophones. Les données sont représentées en Hz pour le KEY et en octave pour le RANGE.

5.4 Registre et type de production

Nous analysons à présent l'effet du type de production (TYPE) sur la hauteur et l'étendue du registre. Nous avons réparti nos données sous deux catégories. La lecture oralisée des corpus PFC et PAC est notée R ; les conversations des corpus PFC et CID et les entretiens radiophoniques du corpus Aix-Marsec sont notées A - pour « parole authentique » (nous rappelons, c'est-à-dire dont le but est de communiquer et non de répondre à une tâche demandée). La représentation graphique en 49 laisserait à penser à un effet de TYPE. La parole authentique serait caractérisée par un registre plus bas et plus étroit que la parole lue, ce qui, d'ailleurs, irait à l'encontre des faits rapportés dans la littérature.

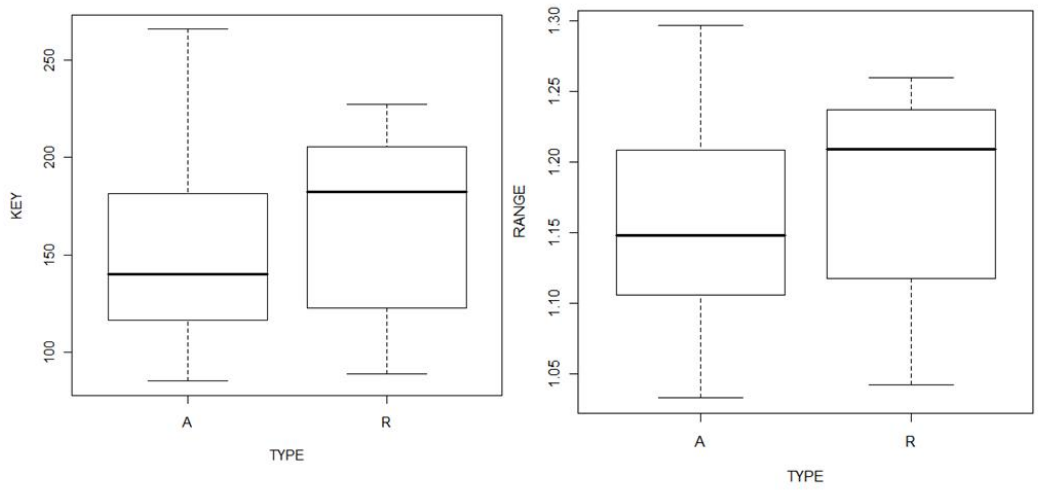


FIGURE 49 – Boîtes à moustaches de la hauteur du registre (KEY ; à gauche) et de l'étendue du registre (RANGE ; à droite) selon les modalités de la variable TYPE. Les données sont représentées en Hz pour le KEY et en octave pour le RANGE.

Cependant, l'analyse de variance conclut un effet non significatif (KEY : F(1,95)=3.143, p-value= 0.07944 ; RANGE : F(1,95)=2.65, p-value=0.1068). D'ailleurs, si l'on veut comparer un éventuel changement de registre en fonction du type de production, il faut plutôt le faire à partir de mêmes locuteurs. Nous proposons donc d'évaluer ces possibles différences sur le corpus PFC puisque les locuteurs sont enregistrés en lecture oralisée (LO), en conversation guidée (CG) et en conversation spontanée (CS). Si l'on catégorise LO sous l'étiquette R et CG et CS sous l'étiquette A, l'analyse de variance ne montre aucun effet du TYPE aussi bien pour la hauteur KEY (F(1,28)= 0.8227 ; p-value=0.3721) que pour l'étendue RANGE (F(1,28)= 0.6692 ; p-value=0.42). Elle ne révèle, de plus, aucun effet de TYPE sur KEY et RANGE si l'on distingue les trois catégories LO, CG et CS (KEY : F(2,27)=0.5909 ; p-value= 0.5608 ; RANGE : F(2,27)=0.4632 ; p-value= 0.6342). Nous pouvons donc conclure, que, à partir de nos données, la lecture oralisée, la conversation guidée et la conversation spontanée ne sont pas caractérisées par des différences de registre significatives. Cependant, il est important de noter que cette analyse n'a été menée qu'à partir de 10 locuteurs ; les résultats obtenus ne peuvent donc être en aucun cas généralisables. Pour répondre à cette problématique, il faudrait partir d'un échantillon de données bien plus important.

5.5 Discussion

Les analyses effectuées à partir des corpus PAC, PFC, Aix-Marsec et CID ont montré que les différences de registre résultent du sexe du locuteur plutôt que de la langue parlée ou encore du type de parole pratiqué. Penser que l'étendue du registre des femmes est plus large pour celle des hommes n'est pas stéréotypisation mais s'explique par la forte corrélation qu'entretient l'étendue avec la hauteur, et par conséquent, par les différences biologiques qu'on constate entre hommes et femmes.

Notre étude ne corrobore donc pas celle de Demers (2000) ou encore de S. F. Mennen I. et Docherty (2008) qui ont observé des différences de registre entre individus de nationalité différente. Pour notre part, nous pensons que ces résultats résultent d'un artefact et ne sont que le reflet du registre « intrinsèque » des locuteurs de ces études.

Par ailleurs, nos résultats ne corroborent pas non plus les travaux révélant des différences de registre entre style de parole. Nous pensons, en revanche, que l'étude des variations de registre intra-locuteurs en fonction des styles de parole révèlerait, elle, les différentes stratégies que les locuteurs peuvent adopter en fonction de la parole qu'ils pratiquent ou en fonction de leurs intentions.

6 Conclusion : une synthèse

L'élaboration de ce chapitre s'appuie sur la problématique de la difficulté de la mesure du registre. Nous avons d'abord listé les problématiques sous-jacentes à sa mesure : (1) la difficulté d'une mesure fiable de la f_0, (2) le choix d'une unité de mesure ou échelle de mesure et (3) le choix d'une mesure acoustique vs. d'une mesure linguistique. Nous avons ensuite cherché à répondre à deux de ces problématiques, celle de la difficulté d'une mesure fiable à partir de la f_0 et celle d'un choix de mesure (acoustique vs. linguistique). La deuxième ne nécessite pas d'étude approfondie puisque les auteurs s'entendent généralement sur l'unité de mesure à utiliser. Les résultats de nos travaux ont cependant révélé qu'une autre échelle de mesure que celles proposées communément pouvait être intéressante pour l'étude du registre et de ses variations.

La première problématique est abordée à partir d'une base de données importante, extraite des corpus Aix-Marsec et PFC. Plus précisément, nous avons proposé une optimisation de la détection des valeurs extrêmes de la f_0, dans PRAAT, afin de rendre plus fiable la mesure du registre. Le travail passe d'abord par une annotation manuelle des extrema de la f_0 pour l'ensemble des locuteurs. Une fois ces extrema annotés, nous les avons corrélés à une détection

automatique des extrema de la f_0. Nous avons montré que la détection de ces extrema sans ajustements des seuils plancher et plafond de la f_0 sont très peu corrélés à l'annotation manuelle. Ce résultat montre qu'en l'état la détection automatique des extrema est sujette à des valeurs aberrantes. Par conséquent, nous avons proposé de comparer un ensemble de possibles seuils plancher et plafond et avons regardé ceux qui permettent une meilleure détection des extrema. L'étude, qui s'appuie sur un ajustement à partir des quantiles de la f_0, a pu montrer que la combinaison, dont le seuil plancher est le produit du 15^{me} quantile et du coefficient 0.83 et dont le seuil plafond est le produit du 65^{me} quantile et du coefficient 1.92, est la plus adéquate dans l'estimation des valeurs extrêmes de la f_0. Elle est aussi bien plus adaptée qu'un ajustement manuel des seuils en fonction du sexe du locuteur. Nous avons conclu qu'il est donc possible d'estimer les limites du registre du locuteur.

Une fois la fiabilité de la mesure du registre établie, nous nous sommes intéressée à la problématique d'une mesure linguistique vs. acoustique du registre. Cette étude s'ancre dans la dichotomie avancée par Patterson (2000) qui soutient que les mesures linguistiques sont préférables aux mesures acoustiques dans l'étude du registre. Elles seraient en effet davantage corrélées au jugement perceptif du locuteur. Par ailleurs, selon lui, les mesures acoustiques seraient sujettes à des erreurs de détection et seraient inadaptées du fait que la distribution de la f_0 ne suit pas une loi normale. Après avoir décrit les différentes mesures acoustiques et linguistiques, nous avons montré, à partir de 4 corpus (PFC, PAC, CID et AM), et pour répondre aux objections portées aux mesures acoustiques, qu'une simple transformation des données suffit à rendre la distribution des données proche d'une loi normale. Ensuite, une fois la certitude de la fiabilité de la f_0 posée et la transformation des données effectuée, nous avons comparé les mesures acoustiques aux mesures linguistiques pour vérifier la réalité de cette dichotomie. Notre étude révèle ainsi que les mesures linguistiques et les mesures acoustiques sont très proches et s'équivalent dans la mesure du registre. Ainsi nous avons conclu que cette dichotomie est sans réel fondement.

Par ailleurs, la comparaison de ces deux mesures nous a mené à un constat intéressant. Nous avons pu démontrer que les cibles basses et hautes du registre sont fortement corrélées à la médiane. Nous avons pu ainsi proposer une mesure de la hauteur et de l'étendue du registre à partir de cette dernière. En outre, ce phénomène suggère que la hauteur et l'étendue du registre co-varient. De plus, ce constat a permis d'envisager un nouvel ajustement des seuils plancher et plafond, i.e. respectivement à partir de la demi-octave inférieure par rapport à la médiane et à partir de l'octave supérieure par rapport à la médiane.

Le chapitre se termine par l'étude du registre et de ses fonctions extra-linguistiques, notamment par une comparaison des registres des locuteurs en fonction de leur sexe, de leur origine géographique et du style de parole qu'ils adoptent. Après un rappel des fonctions extra-linguistiques des variations de registre, nous avons montré qu'il n' y a pas d'effets de langue ou de type de

production sur le registre. Les différences qui ont pu être observées dans la littérature seraient plutôt le reflet des différences biologiques entre hommes et femmes.

Chapitre 3

Détection des variations de registre

Nous nous intéressons à présent aux variations de registre intra-locuteurs. Nous avions soulevé, au cours du premier chapitre, un certain nombre de problématiques autour du concept de registre, notamment celle de la difficulté de sa mesure, que nous avons traitée dans le deuxième chapitre de cette thèse. Une autre problématique qui s'offre à l'étude du registre est celle de la définition de l'empan temporel de ses variations. Nous proposons donc, à présent, de poser plus précisément cette problématique, et ensuite, d'y répondre par la présentation d'expérimentations.

1 Problématique

Nous avions soulevé qu'une description des patrons intonatifs d'une langue, sans la considération des variations globales de registre, semblait, à elle seule, fragile. Or, nous avons également montré la difficulté d'établir le domaine à partir duquel opèrent les variations de registre. A partir de ce qui avait déjà été observé, nous avions remarqué que les variations déclinantes et verticales pouvaient se faire sur un empan temporel différent, selon qu'elles résultent de la structure hiérarchique et de l'organisation informationnelle du discours, des choix (conscients ou inconscients) ou de l'intention du locuteur dans le message qu'il veut faire passer. Les empans temporels des variations de registre, décrits dans la littérature, sont ainsi le syntagme, la proposition, la phrase, le groupe de sens ou groupe de souffle, l'unité intonative, l'énoncé,

le paraton, des domaines ou constituants aussi bien syntaxiques, prosodiques que phonologiques. Les variations déclinantes résulteraient de la modification du registre sur un locus très court (de l'ordre de la syllabe) mais qui servirait de délimitation de domaines plus larges (le syntagme ou encore l'énoncé). Les variations verticales, quant à elles, sont généralement considérées sur un empan temporel global, bien qu'elles puissent opérer sur un empan plus court, de la taille du mot prosodique (Nespor & Vogel, 1983) par exemple.

Nous avons donc cherché à déterminer l'empan temporel des variations de registre, de façon automatique, afin de mesurer l'apport de l'intégration de ces variations dans le système INTSINT et d'étudier la façon dont ces variations indiquent la structure hiérachique du discours, notamment des changements de topique. La détection automatique a nécessité l'élaboration d'un algorithme, ADoReVA, que nous présenterons ci-après et à partir duquel nous avons pu effectuer nos analyses.

Pour cette étude, nous avons utilisé les 4 corpus présentés dans le chapitre 2 de cette thèse : le PFC et le CID pour le français, le PAC et Aix-Marsec pour l'anglais. Nous décrivons ci-après les échantillons sélectionnés de ces données et les annotations effectuées.

2 Base de données et annotations

2.1 Base de données

Nous avons choisi les 4 corpus, dans cette étude, pour une analyse inter-langues et entre types de production. Si, en effet, nous n'avons pas trouvé de différences de registre globales entre types de production dans le deuxième chapitre de cette thèse, nous formulons cependant l'hypothèse que les variations au sein de ces différents types de production se comportent différemment. Nous nous attendons, au vu de la littérature, à ce que les variations de registre en parole lue, résultent uniquement de la structure hiérarchique et l'organisation informationnelle du discours, alors qu'une parole authentique serait caractérisée par des variations de registre reflétant aussi les intentions et émotions du locuteur.

Nous avons donc sélectionné :

- un total de 30 minutes d'enregistrement pour le PFC (6 femmes, 4 hommes), uniquement les lectures oralisées.

- un total de 30 minutes d'enregistrement pour le PAC (3 femmes, 5 hommes), des lectures oralisées également.

- un total de 30 minutes d'enregistrement pour le CID (3 hommes, 3 femmes), des conversations intimes entre amis.

- un total de 30 minutes d'enregistrement pour AIX-MARSEC (5 hommes, 2 femmes), des conversations radiophoniques.

2.2 Annotation des échantillons de données

A partir de nos échantillons de données, nous avons effectué, à partir du lociciel PRAAT, une annotation en groupes clitiques (*clitic group*) ou syntagmes mineurs (*minor phrase*)[84]. L'annotation en groupes clitiques ou syntagmes mineurs, située dans le cadre de la théorie prosodique, est respectivement basée sur les travaux de Nespor et Vogel (1983) et Selkirk (1984). Dans ces travaux, la représentation phonologique, liée au module syntaxique, est exprimée à travers une structure hiérarchisée en constituants prosodiques. Chez Nespor et Vogel (1983), les mots grammaticaux (*function words*) et les mots lexicaux (*content words*) sont reconnus comme des mots prosodiques dans la hiérarchie prosodique. Les mots grammaticaux, associés aux mots lexicaux adjacents, forment ainsi des groupes clitiques. Chez Selkirk (1984), seuls les mots lexicaux sont des mots prosodiques, qui, associés aux mots grammaticaux, forment des syntagmes mineurs. C'est donc sur ces définitions que se base notre annotation. Nous utilisons cette notion de groupe clitique ou syntagme mineur, pour les deux langues, en tant qu'unités prosodiques définies sur la base de critères syntaxiques. De tels domaines permettent en effet la mesure du registre sur une unité assez petite, i.e. en-deça du syntagme phonologique (Nespor & Vogel, 1983; Selkirk, 1984), mais également assez large, i.e. au-delà de la syllabe, pour pouvoir extraire un nombre assez important d'échantillons de la f_0 à partir desquels seront calculées la hauteur et l'étendue. Dans notre annotation donc, tout mot grammatical est regroupé au mot lexical auquel il s'associe, sauf, si les deux éléments sont séparés d'une pause ou si le locuteur porte en emphase volontairement le mot grammatical[85]. Deux exemples de découpage sont données dans les figures 50 et 51.

84. Les annotations sont effectuées à partir du logiciel Praat (Boersma & Weenink, 2009), par la création d'objets TextGrid.

85. Les TextGrids à partir desquels ont été effectuées les annotations et les fichier son correspondants sont donnés sur CD ROM, dossier DATA_CHAP3.

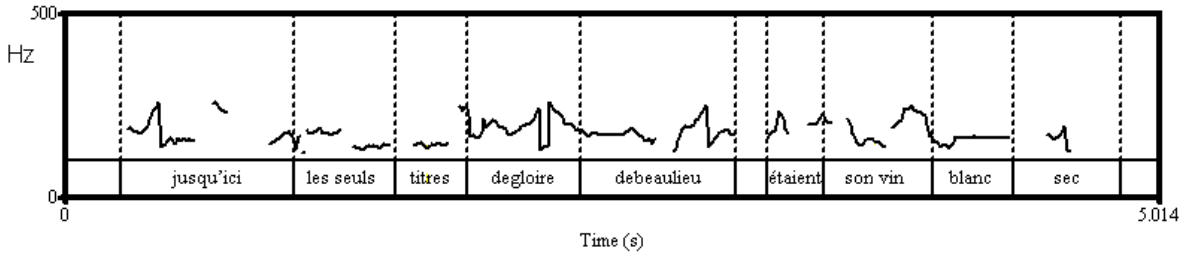


FIGURE 50 – Extrait d'annotation effectuée en groupes clitiques pour le locuteur 13aAC1tw du corpus PFC.

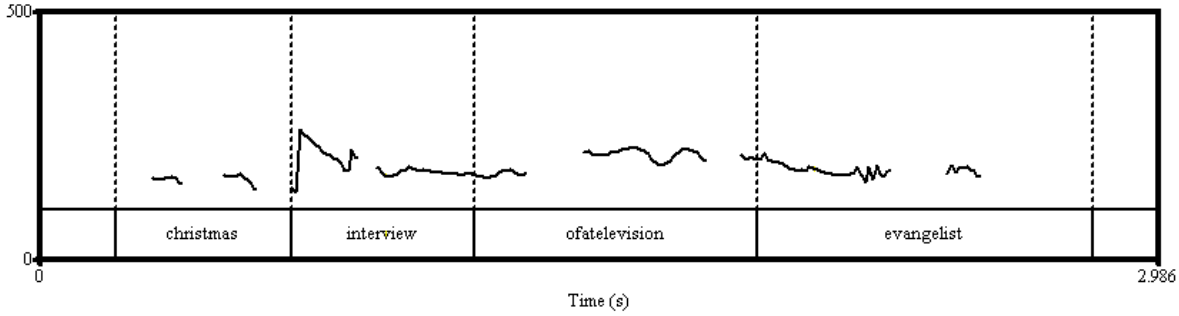


FIGURE 51 – Extrait d'annotation effectuée en groupes clitiques pour le locuteur beckienorm du corpus PAC.

A partir de ces annotations, nous souhaitions évaluer les variations de registre intra-locuteurs. Notamment, nous cherchions à déterminer leur empan temporel et examiner la façon dont elles indiquent des changements de topique. Pour cela, nous avons pensé à l'élaboration d'un algorithme de regroupement hiérarchique (*clustering algorithm*) qui permet à la fois une représentation graphique des variations de registre et leur calcul.

3 ADoReVA : un outil de détection automatique des variations de registre

L'algorithme de regroupement hiérarchique que nous avons développé porte le nom d'ADoReVA qui signifie *Automatic Detection of Register Variations Algorithm*[86]. Il a été spécifiquement conçu pour être implémenté dans Praat, sous forme de plugin. Un plugin permet en effet

86. Nous remercions tout particulièrement Stéphane Rauzy, du Laboratoire Parole et langage, pour son aide dans l'élaboration de cet algorithme et des formules proposées dans ce chapitre.

l'ajout de nouvelles commandes dans Praat sans avoir à recourir directement à la manipulation de scripts. Il est accessible à partir du menu New et est installé dans les Préférences de Praat. Les différents scripts qui le constituent apparaissent alors sous forme de commandes et peuvent donc être lancés directement à partir de ces dernières. Composé de 5 scripts, chacun définissant une étape particulière de l'algorithme, le plugin ADoReVA permet la détection automatique des variations de registre (cf. figure 52). Nous proposons ci-après, de le décrire étape par étape.

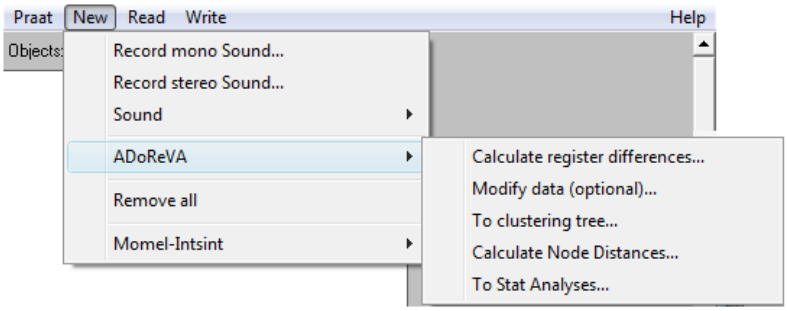


FIGURE 52 – Aperçu du plugin ADoReVA, tel que implémenté dans PRAAT. Les 5 étapes apparaissent sous différents onglets, à droite.

3.1 Etape 1 : Calcul des différences de registre entre unités

Calculate register differences...[87] : La première étape consiste à calculer les différences de registre, i.e. les différences de hauteur et d'étendue entre deux unités. Pour cela, l'algorithme calcule tout d'abord la hauteur (KEY) et l'étendue (RANGE) du registre pour chaque unité. Les formules utilisées sont celles élaborées en section 3.5 du deuxième chapitre. Nous les rappelons ci-après :

$$KEY = MED$$

$$RANGE = 0.161 \times \log_2(MED)$$

où MED est la médiane.

Les objets en entrée sont donc, pour chaque locuteur, un objet TextGrid, dans lequel sont annotées les unités (dans notre cas les groupes clitiques) et un objet Sound. De l'objet Sound est créé un objet Pitch, à partir duquel sont extraits les échantillons de la f_0. Il est à noter que l'Objet pitch est créé à partir de l'objet Sound selon les ajustements des seuils plancher et plafond tels que décrits en section 4 du deuxième chapitre, i.e. respectivement, à la demi-octave

87. Les scripts que nous allons décrire sont tous donnés sur CD ROM - dossiers SCRIPTS, ADOREVA.

inférieure et à l'octave supérieure à la médiane. Une fois la hauteur et l'étendue obtenues pour chacune des unités, l'algorithme compare deux à deux les unités et calcule ainsi la différence de hauteur et d'étendue (DIFFKEY et DIFFRANGE) entre ces deux unités. Les différences obtenues sont donc à chaque fois calculées sur des unités consécutives. Ci-après, les formules utilisées pour ce calcul :

$$DIFFKEY = |\log_2(MEDunit) - \log_2(MEDprevUnit)| \tag{13}$$

$$DIFFRANGE = |0.161 \times \log_2(MEDunit) - 0.161 \times \log_2(MEDprevUnit)| \tag{14}$$

qui revient à l'équation suivante :

$$DIFFRANGE = |0.161 \times \log_2(DIFFKEY)| \tag{15}$$

et,

$$DISTEUCLY = \sqrt{(DIFFKEY^2) + (DIFFRANGE^2)} \tag{16}$$

qui équivaut l'équation suivante :

$$DISTEUCLY = \sqrt{(0.161^2) + (1^2)} * DIFFKEY \tag{17}$$

Dans ces formules, *MEDunit* est la médiane calculée sur l'unité en question et *MEDprevUnit* la médiane calculée sur l'unité précédente. Il est à noter que nous avons ajouté à nos calculs celui de la distance euclidienne (DISTEUCLY), c'est à dire la différence de registre entre deux unités consécutives selon deux paramètres : la hauteur et l'étendue. Dans cette étape est également calculé le registre global du locuteur. Les résultats obtenus sont ainsi rendus sous format tabulaire, comme le montre l'extrait ci-dessous (Figure 53) :

FILENAME	MIN	MAX	SPEAKER_KEY	SPEAKER_RANGE	UNITS	INTSTART	INTEND	INTDUR	HERTZ	KEY	RANGE	DIFFKEY	DIFFRANGE	EUCLY
13aAC1tw	137	408	206	1,24	le premier	0,235	0,835	0,6	307	8,26	1,33	NA	NA	NA
13aAC1tw	137	408	206	1,24	ministre	0,835	1,595	0,76	299	8,22	1,32	0,041	0,007	0,042
13aAC1tw	137	408	206	1,24	ira t il	1,695	2,285	0,59	247	7,95	1,28	0,274	0,044	0,278
13aAC1tw	137	408	206	1,24	à beaulieu	2,285	2,875	0,59	166	7,38	1,19	0,571	0,092	0,579
13aAC1tw	137	408	206	1,24	levillage	4,217	4,867	0,65	284	8,15	1,31	0,771	0,124	0,781
13aAC1tw	137	408	206	1,24	debeaulieu	4,867	5,517	0,65	226	7,82	1,26	0,329	0,053	0,333
13aAC1tw	137	408	206	1,24	est en grand	5,517	6,147	0,63	236	7,88	1,27	0,062	0,010	0,063
13aAC1tw	137	408	206	1,24	émoi	6,147	6,557	0,41	167	7,38	1,19	0,500	0,081	0,506

FIGURE 53 – Extrait des données de sortie obtenues suite à l'étape 1 pour le corpus PFC. La colonne FILENAME donne le nom du fichier, MIN et MAX les valeurs extrêmes du registre du locuteur, SPEAKERKEY et SPEAKERRANGE la hauteur et l'étendue du registre global du locuteur, UNITS l'unité en question, INTSTART, INTEND, INTDUR le début, la fin et la durée de l'unité en question, HERTZ la hauteur de l'unité en Hertz, KEY la hauteur de l'unité en log de base 2, RANGE l'étendue en log de base 2, DIFFKEY, DIFFRANGE les différences de hauteur et d'étendue entre l'unité en question et l'unité précédente et EUCLY la distance euclidienne obtenue entre les deux unités consécutives.

3.2 Etape 2 : Mise en forme des données (optionnel)

Modify data (optional)... : La deuxième étape de l'algorithme consiste en la mise en forme des données. Parce que l'étape 3 fait appel à un format particulier (xml), les données ne doivent pas contenir certains symboles de type « $\leq$ » ou encore « \ ». Si les textGrids sont susceptibles de contenir ces symboles, il est préférable de passer par cette seconde étape. Dans le cas contraire, elle n'est pas nécessaire et l'utilisateur peut directement passer à la troisième étape.

3.3 Etape 3 : Classification ascendante hiérarchique - création de dendrogrammes

To clustering tree...[88] : Une fois les différences de hauteur, d'étendue et la distance euclidienne obtenues entre unités consécutives, l'algorithme de classification ascendante hiérarchique groupe les unités en fonction de leurs différences et de leur distance. Pour chaque locuteur, trois regroupements sont effectués : (1) un regroupement à partir des différences de hauteur, (2) un regroupement à partir des différences d'étendue et (3) un regroupement à partir des distances euclidiennes. Chacun de ces regroupements se forment de la façon suivante : l'algorithme détecte la différence ou la distance la plus petite entre deux unités et effectue leur branchement. Ces unités regroupées forment une nouvelle unité. Ensuite, les différences de hauteur et d'étendue et la distance euclidienne entre cette unité et sa précédente sont à

88. Les dendrogrammes obtenus sont donnés dur CD ROM - dossier DENDRO_CHAP3.

leur tour recalculées. Et ce, de façon itérative, jusqu'à ce qu'il ne reste plus d'unités ou de groupes d'unités à embrancher. Dans une telle procédure donc, plus la différence ou la distance entre deux unités consécutives est petite, plus vite elles sont branchées ; et inversement, plus la différence ou la distance entre deux unités est grande, plus elles sont regroupées tardivement. L'algorithme de regroupement hiérarchique que nous proposons est donc similaire aux algorithmes de regroupement hiérarchique déjà existants, mais il a, à la différence de ces algorithmes, la contrainte de regrouper les unités entre elles en fonction de leur ordonnée temporelle.

L'algorithme génère ensuite une structure arborescente binaire qui prend la forme d'un diagramme à niveaux alignés (*layered icicle diagram*). La sortie au format .xml peut être visualisée à partir d'une feuille de style .xsl. Cette représentation graphique permet ainsi de visualiser les changements de registre intra-locuteurs et ainsi de visualiser la structure hiérarchique et l'organisation relationnelle des unités telles qu'elles sont reflétées par les changements de registre. A partir du dendrogramme, il est donc possible de distinguer des groupes d'unités, à travers des cassures visuelles de l'arborescence. Plus la cassure est grande entre deux unités, plus la différence de registre entre ces deux unités ou groupes d'unités est importante. Pour chaque locuteur, nous obtenons trois dendrogrammes, le premier effectué à partir des différences de hauteur de registre, le deuxième à partir des différences d'étendue et le troisième à partir des distances euclidiennes.

Nous proposons une interprétation visuelle de l'extrait de dendrogramme obtenu à partir des distances euclidiennes pour le locuteur 13aAC1tw (corpus PFC) et représenté en 54. Nous pouvons observer, au bas de la structure arborescente, les unités ou feuilles à partir desquelles sont calculées la hauteur et l'étendue du registre. Les unités sont ensuite regroupées entre elles selon un branchement binaire et l'unité nouvellement créée indique la différence ou distance qui les sépare. Cette différence ($diff$) est la moyenne pondérée des distances au barycentre des deux unités regroupées. En voici la formule :

$$\overline{x} = \frac{w_{k+1}\, x_{k+1} + w_k\, x_k}{w_{k+1} + w_k} \;\; ; \;\; diff = \frac{w_{k+1}\,(x_{k+1} - \overline{x}) + w_k\,(\overline{x} - x_k)}{w_{k+1} + w_k} \tag{18}$$

où x_k et x_{k+1} sont les positions des unités ou groupes d'unités contigües, w_k et w_{k+1} sont les poids associés aux unités, proportionnel au nombre d'unités du groupe. Pour les feuilles, $x_{k+1} = x_k + DIFFKEY$, lorsque le paramètre est la hauteur du registre (KEY).

La distance entre l'unité « le premier » et l'unité « ministre » est de 0.041 alors que celle entre le groupe d'unités « le premier ministre ira t-il à Beaulieu » et le groupe d'unités « le village de Beaulieu est en grand émoi » est de 1.773. Nous voyons d'ailleurs clairement que la rupture est plus grande entre l'unité « Beaulieu » et l'unité « le village » qu'entre l'unité « le

premier » et l'unité « ministre ». Les couleurs, quant à elles, indiquent la hauteur de chaque unité et de chaque groupe d'unités. Plus la couleur est chaude, plus l'unité est énoncée sur de hautes fréquences ; et inversement, plus la couleur de l'unité est froide, plus elle est énoncée sur de basses fréquences. Visuellement, nous pouvons alors apercevoir l'effet de déclinaison que nous avions abordé en 1.3 de notre deuxième chapitre : les unités en début de groupes ont un registre relativement plus haut que les unités en fin de groupes (e.g. « le premier ministre » vs. « à Beaulieu »). L'effet de déclinaison s'observe également à un niveau supérieur, i.e. entre groupes d'unités où le groupe d'unité initial a un registre relativement plus haut que les groupes d'unités qui le suivent (e.g. « le premier ministre ira t-il à Beaulieu » vs. « le village de Beaulieu est en grand émoi »). Si l'on compare d'ailleurs cet extrait de début de texte à l'extrait de fin de texte, représenté en figure 55, on voit très clairement l'effet de déclinaison sur un empan plus large. Alors que l'extrait en début de lecture est marqué par des couleurs chaudes, i.e. par un registre haut, l'extrait en fin de lecture est lui plutôt dominé par des couleurs froides, i.e. par un registre bas. Parce que la hauteur et l'étendue du registre sont corrélées, on peut déjà affirmer visuellement que le début d'un texte est caractérisé par un registre haut et étendu alors que la fin d'un texte est caractérisé par un registre bas et étroit.

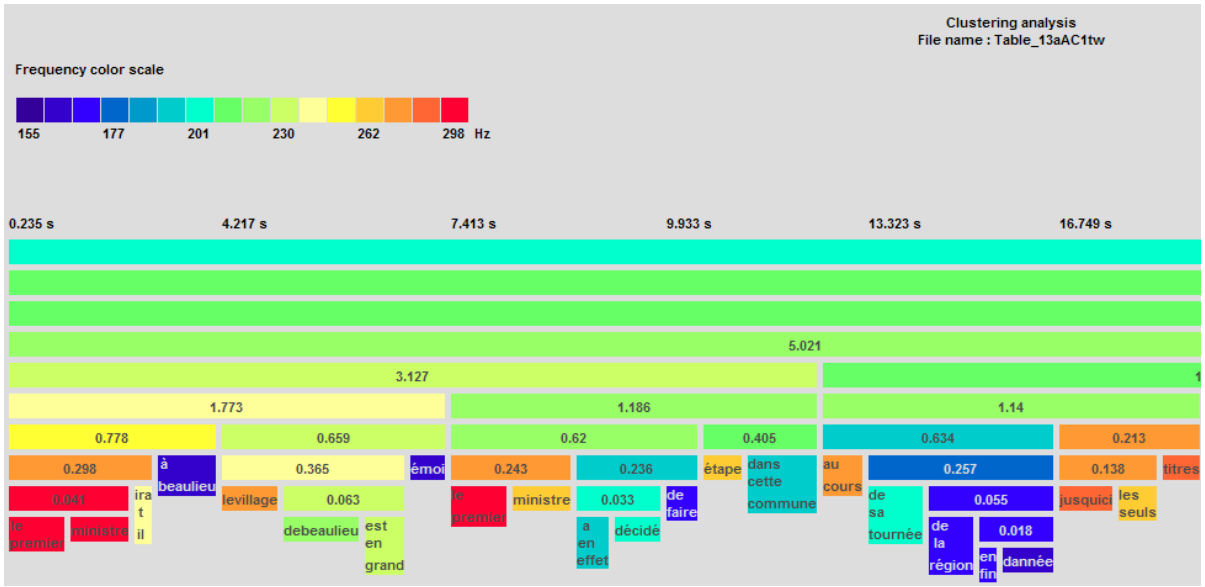


FIGURE 54 – Aperçu du dendrogramme obtenu à partir des distances euclidiennes pour le locuteur 13aAC1tw (corpus PFC). En haut à gauche, une échelle de couleurs indique les fréquences utilisées par le locuteur. En haut à droite est donné le nom du fichier traité. Au dessus du dendrogramme est indiqué le temps en secondes. Les feuilles au bas de la structure arborescente représentent les unités à partir desquelles sont calculées la hauteur et l'étendue.

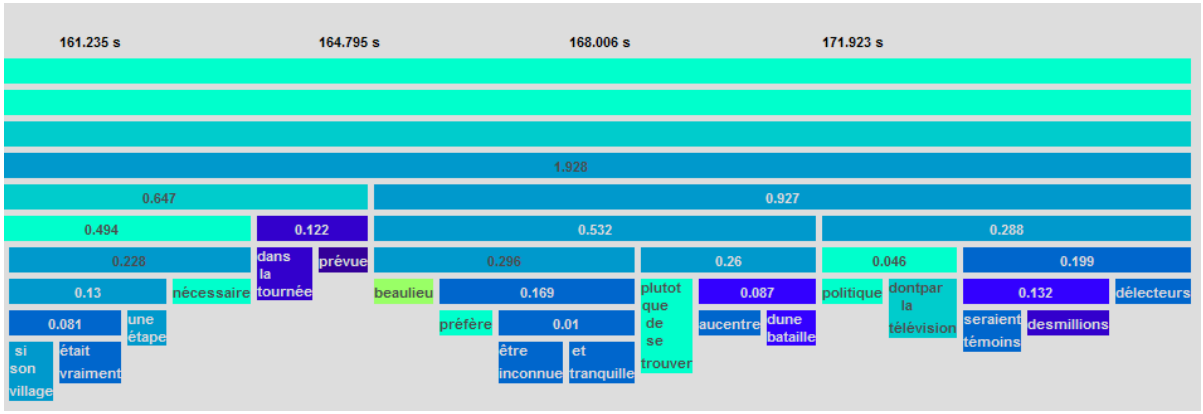


FIGURE 55 – Aperçu du dendrogramme obtenu à partir des distances euclidiennes pour le locuteur 13aAC1tw (corpus PFC). L'aperçu correspond ici à l'extrait de fin de lecture.

Puisqu'il est possible de visualiser des changements de registre, on peut penser qu'une telle structure arborescente nous permettrait d'établir des ruptures de structure, i.e. des changements de registre intra-locuteurs. Parce que les ruptures de l'arborescence peuvent être exprimées par la distance entre les feuilles de la structure, l'étape suivante consiste donc en leur calcul.

3.4 Etape 4 : Calcul des distances entre les feuilles de la structure arborescente

Calculate Node Distances... : A partir de la structure arborescente, l'algorithme calcule la distance entre les noeuds feuilles afin d'obtenir le degré de frontière ou la rupture de registre entre deux unités consécutives. La distance calculée est ici définie comme la valeur de la différence associée au noeud père commun des deux unités (voir formule 18). Plus la distance est grande entre deux unités, plus le degré de frontière ou la rupture entre ces deux unités est fort. Au contraire, une petite distance suggèrera plutôt que les deux unités appartiennent à un même groupe, i.e. qu'elles sont énoncées sur un même registre. Un extrait des données de sortie est affiché en 56. C'est à partir de ces données de sortie que nous chercherons à définir l'empan temporel des variations de registre. Nous aborderons ce point à la section 4 de ce chapitre.

Leaf	LeftDist	RightDist	start	end
le premier	49.166	0.041	0.235	0.835
ministre	0.041	0.294	0.835	1.595
ira t il	0.294	0.768	1.695	2.285
à beaulieu	0.768	1.751	2.285	2.875
levillage	1.751	0.36	4.217	4.867
debeaulieu	0.36	0.062	4.867	5.517
est en grand	0.062	0.651	5.517	6.147
émoi	0.651	3.087	6.147	6.557

FIGURE 56 – Extrait des données de sortie obtenues suite à l'étape 4 pour le locuteur 13aACtw du corpus PFC. La colonne *Leaf* indique les unités traitées, la colonne *LeftDist* la distance calculée à gauche de l'unité, i.e. entre une unité et sa précédente, la colonne *RightDist* la distance calculée à droite de l'unité, i.e. entre une unité et sa suivante ; la colonne *start* donne la valeur temporelle du début de l'unité, la colonne *end* celle de la fin de l'unité.

3.5 Etape 5 : Vers une analyse statistique des données

To Stat analyses... : Cette étape de l'analyse permet une étude des variations de registre et des fonctions qu'elles revêtent. Nous avions en effet expliqué, dans le début de ce chapitre, que nous cherchions à étudier les variations de registre en fonction des changements de topique. A partir des annotations manuelles que nous avons effectuées (décrites dans la section 5.2), l'algorithme crée une table de données qui regroupe à la fois les mesures effectuées dans les étapes précédentes et l'annotation fonctionnelle. Pour l'utilisateur, il suffit de cocher, dès la première étape de l'algorithme, l'option de corrélation entre annotations fonctionnelles et détection automatique des variations de registre, telle que présentée en 57, pour que l'algorithme crée cette base de données. L'étape 5 fournit donc une table à partir de laquelle pourront être menées des analyses statistiques, e.g. à partir de l'environnement de travail R.

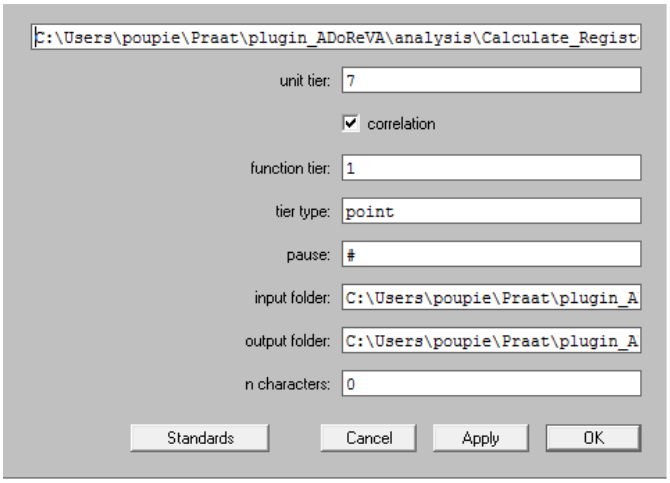


FIGURE 57 – Fenêtre d'ajustement des paramètres du script *Calculate register differences...* utilisé pour l'étape 1. La case *correlation* cochée indique que le TextGrid en entrée contient une annotation fonctionnelle que l'on souhaite corréler à la détection automatique des variations de registre. Le paramètre *function tier* indique le numéro de la *tier* à partir de laquelle est effectuée l'annotation fonctionnelle.

Avant de porter notre intérêt sur la relation entre variations de registre et changements de topique, nous chercherons dans la section suivante, à déterminer l'empan temporel des variations de registre intra-locuteurs, et ce, à partir des distances obtenues entre les noeuds feuilles de la structure arborescente. La détermination de l'empan temporel se fera par l'étude de l'optimisation du codage INTSINT par intégration des variations de registre.

4 Détermination de l'empan temporel des variations de registre

4.1 Intégration des variations de registre dans le système d'intonation INTSINT par ajustement de seuils

Suite à une inspection visuelle des variations de registre à partir d'une structure arborescente et suite au calcul des distances entre les feuilles de cette structure, obtenu à partir de la distance euclidienne (EUCLY), nous cherchons à délimiter un seuil pour lequel l'annotation automatique des cibles tonales est la plus optimale. Pour cela, nous utilisons l'algorithme MOMEL-INTSINT.

L'algorithme MOMEL (*MOdélisation MELodique*), proposé par Hirst et Espesser (1993), permet la modélisation du contour intonatif via une approximation quadratique (*Quadratic Spline*

Function Function). Il repose sur l'hypothèse selon laquelle le contour de la f0 est le résultat de la superposition de deux composants distincts et indépendants : la microprosodie (qui correspond aux variations mélodiques à court terme, propres aux segments) et la macroprosodie (qui caractérise le choix intonatif de l'élocution). Dans cet algorithme, le contour intonatif y est représenté comme une séquence de tons statiques, reliés par des arcs de paraboles (*quadratic spline*), avec des passages par zéro de la première dérivée qui servent de localisation des cibles tonales. Plus précisément, les cibles tonales sont calculées comme suit : la première étape de l'algorithme consiste à éliminer les quelques valeurs aberrantes résultantes des perturbations segmentales (points situés à + de 5 % des deux points avoisinants). On suppose qu'en dehors des ces valeurs, toutes les autres valeurs de f0 sont situées sur, ou en dessous de, la courbe modélisée. Une fenêtre glissante, dont la valeur par défaut est de 300 ms, parcourt ensuite le signal de gauche à droite, et calcule à partir de la courbe de la f0, estimée par un polynôme de second degré, des candidats possibles de localisation des cibles tonales. Les valeurs observées qui sont à plus d'un seuil delta en dessous de la courbe modélisée sont ensuite neutralisées et le polynôme quadratique est calculé de nouveau. Cette étape est répétée tant qu'il existe des valeurs trop éloignées de la quadratique. L'étape suivante consiste à l'extraction des candidats. L'espace temporel est partionné en cherchant des endroits où la différence entre la partie gauche et la partie droite de la fenêtre glissante atteint un maximum local. La moyenne et l'écart type des points dans chaque portion de la partition sont ensuite calculés. Les valeurs supérieures à l'écart type par rapport à la moyenne sont à leur tour supprimées et le calcul de la moyenne des candidats restants (en temps et en fréquence) définit la localisation du point cible [89].

L'algorithme INTSINT (*INternational Transcription System for INTonation*), proposé par Hirst (1987) et repris par Hirst et Di Cristo (1998) et Hirst (2005), permet le codage des patrons intonatifs d'un énoncé selon un alphabet de 8 symboles discrets qui définissent une représentation phonologique de surface de l'intonation. Ils sont dits phonologiques car discrets et de surface car ils sont dérivés directement du signal acoustique. L'interprétation des cibles tonales se fait à partir de l'estimation du registre du locuteur, i.e. à partir de sa hauteur (key) et de son étendue (range). Les codages T(op) et B(ottom) des cibles tonales délimitent le registre du locuteur (valeur haute et valeur basse respectivement), le codage M(id) la valeur moyenne du registre. Ils sont définis de façon absolue, i.e. sans considérer la cible qui les précède. Les cibles H(igher), L(ower), S(ame), U(pstep) et D(ownstep) sont, en revanche, encodées en fonction de la cible qui les précède et sont définies comme étant, respectivement, plus hautes, plus basses, égales, un peu plus hautes et un peu plus basses que la cible qui les précède [90].

89. Pour plus de précisions, voir directement Hirst (1987), Hirst et Espesser (1993), Hirst et al. (2000), et Hirst (2005).

90. Pour plus de précisions, cf. Hirst (1987), Hirst et Di Cristo (1998), Hirst (2005) et Hirst (2007).

La sortie de l'algorithme MOMEL-INTSINT consiste, pour chaque objet Sound, en un objet Pitchtier, représentant les cibles tonales du contour intonatif à partir desquelles est estimé le codage INTSINT, et un objet TextGrid qui indique la valeur et le codage de chaque cible tonale. Un exemple est donnée en 58. Chaque objet TextGrid créé est en effet constitué d'une première rangée ou *tier*, appelée MOMEL, qui donne la valeur en Hz de chaque cible tonale et d'une deuxième rangée, appelée INTSINT, qui indique le codage INTSINT pour chacune des cibles tonales. La troisième rangée estime à nouveau la valeur de la cible tonale, calculée cette fois à partir du codage INTSINT.

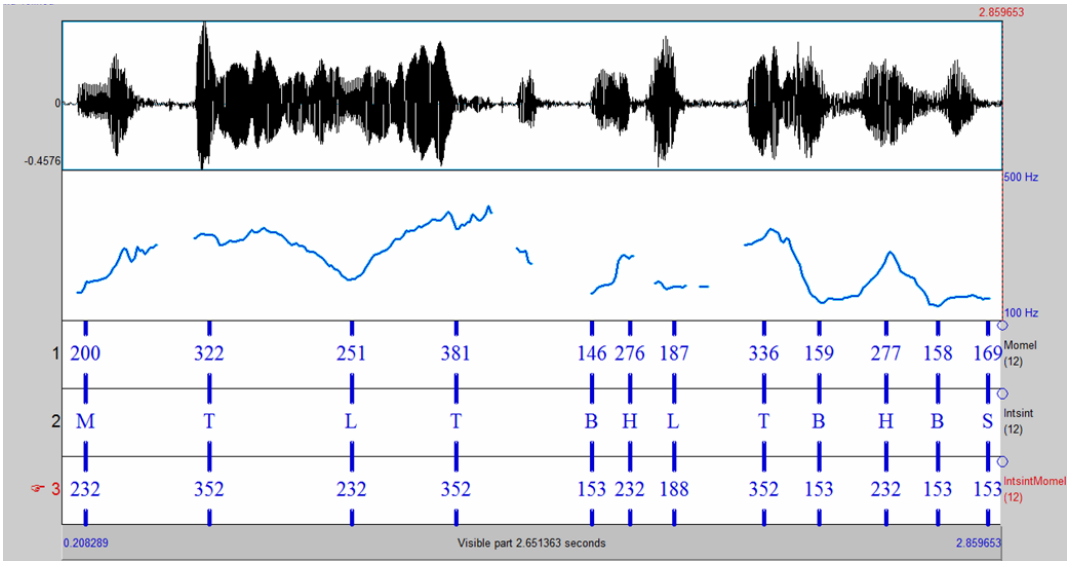


FIGURE 58 – Extrait d'un objet TextGrid tel qu'obtenu à partir de l'algorithme MOMEL-INTSINT. La première rangée donne la valeur des cibles tonales telles que calculées par MOMEL ; la deuxième rangée indique le codage INTSINT pour chaque cible tonale ; la troisième rangée propose une estimation de la valeur des cibles tonales, recalculées à partir du codage INTSINT.

Ce qui nous intéresse donc ici, c'est de voir si l'intégration des variations de registre intra-locuteurs permet une meilleure estimation du codage INTSINT et ainsi une meilleure estimation de la valeur des cibles tonales recalculées. Nous proposons donc, à partir de seuils, et non à partir de la totalité du fichier, de déterminer automatiquement les empans temporels des variations de registre et de faire tourner l'algorithme sur ces empans. Nous rappelons que bien qu'INTSINT prenne en compte la hauteur et l'étendue globale du locuteur pour la génération du codage INTSINT, il ne considère cependant pas les variations de hauteur et d'étendue intra-locuteurs. Or, Hirst (2005) soulignait que si le codage était optimal pour des corpus de phrases courtes lues, il ne pouvait être aussi performant pour des passages plus longs, ou pour de la parole spontanée ou conversationnelle, pour lesquelles les variations de registre sont

nombreuses ; d'où la nécessité de l'estimation de ces variations.

L'estimation des seuils s'est faite à partir de la visualisation de l'arborescence construite pour le paramètre EUCLY. Plusieurs niveaux de regroupement, nous l'avons vu, sont définis. Au vu des différentes représentations arborescentes extraites des corpus PFC, PAC, CID et AIX-MARSEC, et de leurs regroupements internes, il semble intéressant de délimiter les empans temporels des variations de registre à partir des seuils 0.5, 1, 2, 3, 4, 5 et 6. Il est à noter que si les distances dépendent de la longueur totale de la structure arborescente, les seuils, eux, en sont indépendants. Nous avons donc élaboré un algorithme (*Calculate_ intsint_ part* ; disponible sur CD ROM, dossier SCRIPTS) qui permet, à partir d'une table de données telle qu'obtenue à l'étape 4 (et représentée en 56), de récupérer les distances entre chaque feuille de la structure arborescente et ainsi définir le domaine à partir duquel tournera INTSINT. L'algorithme parcourt en effet les distances répertoriées dans la table d'entrée : si la distance est inférieure au seuil, il continue de parcourir la table ; si, au contraire, la distance est supérieure au seuil, il s'arrête et récupère la valeur temporelle finale de l'unité. Le domaine à partir duquel sera lancé INTSINT débute donc par défaut à la première unité, i.e. à 0 secondes, et se termine à la valeur temporelle finale de l'unité marquée par une rupture de registre. La valeur initiale est ensuite remplacée par la valeur finale et l'algorithme parcourt à nouveau la table jusqu'à la prochaine rupture i.e. jusqu'à ce qu'il rencontre une distance supérieure au seuil défini. Le processus est ainsi réitéré jusqu'au traitement total de la table, et donc de l'objet PitchTier. Dans cette configuration, l'algorithme tourne pour chaque objet PitchTier 7 fois, puisque 7 seuils ont été définis dans ce but. Après avoir fait tourner INTSINT sur différents domaines et selon différents seuils, l'algorithme récupère les valeurs MOMEL, INTSINT et IntsintMomel de l'objet TextGrid et les répertorie dans une nouvelle table de données, à partir de laquelle peuvent être menées des analyses statistiques[91]. Un extrait des données de sortie obtenues figure en 8. C'est donc à partir de cette table que nous chercherons à définir le seuil qui permet le codage optimal des patrons intonatifs d'une langue naturelle.

91. Les données de sortie sont sur CD ROM - ANNEXES_CHAP3 :Table1.

FILENAME	TARGETS TIME	INTSINT	MOMEL	IntsintMomel
13aAC1tw	41.123	M	208	205
13aAC1tw	41.315	H	288	270
13aAC1tw	41.445	L	192	178
13aAC1tw	41.597	H	258	252
13aAC1tw	41.831	T	347	357

TABLE 8 – Extrait des données de sortie obtenues par le script *Calculate_intsint_part*. La colonne FILENAME indique le nom du Pitchtier traité, la colonne TARGETS TIME, la localisation temporelle des cibles tonales (en ms), INTSINT, le codage INTSINT, MOMEL, la valeur en Hertz des points cibles, IntsintMomel la valeur recalculée à partir du codage INTSINT des points cibles, en Hz également.

4.2 Détermination d'un seuil pour un codage optimal des patrons intonatifs d'une langue

Afin de déterminer si l'intégration des variations de registre permet une amélioration du codage des symboles INTSINT et ainsi des valeurs recalculées IntsintMomel, nous proposons de comparer les corrélations obtenues à partir des valeurs MOMEL et des valeurs recalculées IntsintMomel des différents modèles testés, i.e. en fonction des 7 seuils définis et sans intégration des variations de registre (i.e obtenues par l'algorithme MOMEL-INTSINT en l'état). Le tableau 9 donne les coefficients de détermination (R^2) obtenus en fonction des seuils et sans seuil, pour l'ensemble des points cibles.

Corpus	0.5	1	2	3	4	5	6	SVR
PFC	0.97	**0.97**	0.96	0.96	0.96	0.96	0.95	**0.94**
PAC	0.98	**0.98**	0.97	0.97	0.97	0.97	0.97	**0.96**
CID	0.97	**0.97**	0.96	0.96	0.95	0.95	0.95	**0.94**
AM	0.98	**0.98**	0.98	0.97	0.97	0.97	0.97	**0.96**

TABLE 9 – Coefficients de détermination ou R^2 obtenus pour la corrélation des valeurs MOMEL et IntsintMomel, à partir des seuils 0.5, 1, 2, 3, 4, 5 et 6 et à partir de l'algorithme MOMEL-INTSINT sans considération des variations de registre (SVR), pour l'ensemble des cibles.

Si nous regardons les coefficients de détermination de la colonne SVR (i.e. obtenus par l'algorithme MOMEL-INTSINT en l'état), nous remarquons que les corrélations sont plus élevées

pour les corpus anglais que pour les corpus français (0.96 vs. 0.94). Si nous nous penchons à présent sur les coefficients de détermination obtenus en fonction des seuils, nous constatons que l'intégration des variations de registre, quel que soit le seuil, augmente le score de la corrélation. Nous remarquons également que plus le seuil est petit, plus le coefficient augmente (e.g. 0.95 pour un seuil à 6, 0.97 pour un seuil à 1 - corpus PFC). En revanche, dépassé le seuil à valeur de 1, le score de la corrélation n'augmente plus (les coefficients sont en effet égaux pour les seuils 0.5 et 1). Nous pouvons donc conclure que l'intégration des variations de registre à un seuil de 1 améliore le calcul des valeurs IntsintMomel, et par là, le codage des patrons intonatifs d'une langue naturelle.

Si l'on regarde à présent de plus près l'effet de l'intégration des variations de registre sur le codage des cibles absolues T, M et B, et ce, à partir des tableaux 10, 11 et 12, nous constatons que le score de la corrélation augmente plus pour les cibles M et B que pour les cibles T (e.g. B : de 0.97 à 0.89 vs. T : de 0.96 à 0.91 - corpus PAC). Cela peut s'expliquer par le fait que la corrélation des T est, à la base, meilleure que la corrélation obtenue pour les M et les B.

Corpus	0.5	1	2	3	4	5	6	SVR
PFC	0.95	**0.95**	0.93	0.92	0.92	0.92	0.92	**0.91**
PAC	0.96	**0.96**	0.94	0.93	0.92	0.91	0.91	**0.9**
CID	0.92	**0.92**	0.88	0.86	0.85	0.85	0.84	**0.83**
AM	0.96	**0.96**	0.94	0.92	0.91	0.91	0.9	**0.9**

TABLE 10 – Coefficients de détermination ou R^2 obtenus pour la corrélation des valeurs MOMEL et IntsintMomel, à partir des seuils 0.5, 1, 2, 3, 4, 5 et 6 et à partir de l'algorithme MOMEL-INTSINT sans considération des variations de registre (SVR), pour les cibles T uniquement.

Corpus	0.5	1	2	3	4	5	6	SVR
PFC	0.92	**0.92**	0.89	0.88	0.88	0.87	0.89	**0.86**
PAC	0.95	**0.95**	0.92	0.89	0.88	0.88	0.88	**0.85**
CID	0.9	**0.9**	0.85	0.84	0.82	0.82	0.8	**0.76**
AM	0.97	**0.97**	0.95	0.93	0.93	0.92	0.92	**0.9**

TABLE 11 – Coefficients de détermination ou R^2 obtenus pour la corrélation des valeurs MOMEL et IntsintMomel, à partir des seuils 0.5, 1, 2, 3, 4, 5 et 6 et à partir de l'algorithme MOMEL-INTSINT sans considération des variations de registre (SVR), pour les cibles M uniquement.

Corpus	0.5	1	2	3	4	5	6	SVR
PFC	0.93	**0.93**	0.92	0.91	0.91	0.89	0.89	**0.81**
PAC	0.97	**0.97**	0.94	0.93	0.92	0.9	0.89	**0.77**
CID	0.93	**0.93**	0.9	0.89	0.88	0.87	0.87	**0.84**
AM	0.97	**0.97**	0.96	0.95	0.95	0.95	0.95	**0.93**

TABLE 12 – Coefficients de détermination ou R^2 obtenus pour la corrélation des valeurs MOMEL et IntsintMomel, à partir des seuils 0.5, 1, 2, 3, 4, 5 et 6 et à partir de l'algorithme MOMEL-INTSINT sans considération des variations de registre (SVR), pour les cibles B uniquement.

Nous pouvons donc conclure qu'intégrer les variations de registre dans des systèmes (semi-) automatique de l'intonation est nécessaire afin d'affiner le codage des patrons intonatifs des langues naturelles, notamment celui des symboles M et B dans l'annotation INTSINT.

4.3 Discussion

La problématique de l'empan temporel des variations de registre est loin d'être résolue. La difficulté d'une telle problématique, nous l'avons dit, repose sur le fait qu'il est difficile de présupposer ces variations du fait qu'elles dépendent de la structure hiérarchique, organisationnelle et informationnelle du discours et des intentions du locuteur. Il est en effet difficile de prédire *le* domaine pour lequel variera le registre. La solution que nous avons proposée est celle d'une détection automatique des variations de registre. Si nous ne partons d'aucun présupposé théorique (i.e. de domaines pré-définis), nous avons en effet plus de chance de relever ces variations. L'algorithme ADoReVA permet en effet une détection automatique des variations de registre, à partir d'unités pré-définies, certes, mais qui, par leur emboitement dans une structure arborescente, ouvrent de nouveaux domaines possibles, d'où la nécessité de partir d'unités relativement étroites afin de saisir un maximum de variations. Cette structure arborescente nous a donc permis de calculer les distances entre les noeuds feuilles de l'arborescence et ainsi définir un seuil à partir duquel peut être élaborée une annotation claire et précise des patrons intonatifs d'une langue naturelle. L'intégration des variations de registre dans des systèmes de l'intonation permet alors de résoudre la problématique du chevauchement et de la délimitation des empans prosodiques à court et à long terme. Nous avons pu ainsi constater une amélioration du codage INTSINT par l'intégration des variations de registre (une intégration qui n'est pas propre au codage MOMEL-INTSINT), définies pour un seuil de 1. Nous avons pu par ailleurs remarquer que l'intégration de ces variations améliorent notamment le codage des symboles M et B. Il serait donc intéressant d'intégrer ces variations dans d'autres modèles

de l'intonation.

Cette section ne permettra donc pas de définir un domaine pour lequel le registre varie. Nous ne pourrons donc trancher entre les divers domaines proposés dans la littérature, i.e. entre le syntagme, la proposition, la phrase, le groupe de souffle, l'unité intonative, l'énoncé ou encore le paraton, qu'ils soient définis selon des critères syntaxiques, prosodiques ou phonologiques. Cependant, il est déjà possible d'envisager de corréler une annotation fonctionnelle en constituants (quels qu'ils soient) à une détection automatique des variations de registre et ainsi de voir à quelles frontières et de quels constituants correspondent les ruptures supérieures au seuil de 1 de la structure arborescente. Pour notre part, nous ne considérons pas qu'il existe *un* domaine pour lequel le registre varie. Comme nous l'avions dit précédemment, le domaine résulte en effet des choix conscients ou inconscients, du moins des intentions du locuteur, il ne peut donc être défini en amont. Nous formulons cependant l'hypothèse que les variations déclinantes résulteraient de la modification du registre sur un empan temporel assez court mais qui servirait de délimitation de domaines plus larges. Les variations verticales, elles, s'observeraient sur des empans divers, qui, même étroits, auraient un effet sur la représentation des patrons intonatifs d'une langue.

Nous adhérons également à la notion de downstep emboité ou de *wheels within wheels model of downstep* proposée par Ladd (1988) et Van Den Berg et al. (1992). Nous avons pu en effet observer un phénomène de déclinaison sur plusieurs niveaux d'emboitement à partir des structures arborescentes. Nous pensons par ailleurs que les structures arborescentes, obtenues à partir d'ADoReVA, permettent la représentation automatique des effets d'emboitement de downstep et/ou de déclinaison, et ce, sur des domaines plus larges. Elles rendent compte ainsi du principe *The Deeper, the Steeper* (Féry & Truckenbrodt, 2005), où la force ou le degré d'abaissement entre deux constituants soeurs est représentée par une distance croissante plus on s'approche de la racine de la structure arborescente. Les premiers résultats d'une analyse en cours sur le corpus Aix-Marsec semblent conforter une telle conception puisqu'ils montrent, en anglais britannique, que la différence de registre entre deux mots situés à la frontière d'unités intonatives majeures [92] est plus importante que celle entre deux mots situés à la frontière d'unités intonatives mineures, également supérieure à celle entre deux mots au sein de l'unité ($F(2,2211)=87.37$; p-val$<$2e-16).

Nous portons à présent notre intérêt sur les fonctions linguistiques que peuvent revêtir les variations de registre, notamment la façon dont elles signalent un changement de topique.

92. L'annotation en unités mineures et majeures est décrite dans le chapitre 4.

5 Registre et fonctions linguistiques : détection automatique et prédiction des changements de topique

Nous proposons tout d'abord un rappel des faits observés dans la littérature.

5.1 Rappel : variations de registre et topicalisation

Nous avions constaté, à partir de la littérature, que les variations de registre indiquaient la structure hiérarchique et organisationnelle du discours. Toute rupture du registre caractériserait la place et le degré de frontières entre deux unités consécutives et ainsi révèlerait la relation des segments dans la structure du discours (Ayers, 1994; Fon, 2002). En lecture de texte, les phrases situées en début de paragraphe sont observées sur un registre plus étendu et plus élevé que les phrases au milieu et en fin de paragraphe. La différence de registre entre ces groupes initiaux, médians et finaux n'est d'ailleurs pas la même, la différence entre les groupes initiaux et médians étant plus importante que celle entre les groupes médians et finaux. En lecture de texte et en parole spontanée ou conversationnelle, les variations de registre communiqueraient donc la structure du topique dans le discours. Un effet de remise à niveau (totale ou partielle) indiquerait ainsi un changement de topique quand l'effet de déclinaison marquerait plutôt la cohésion des unités. En d'autres mots, les unités traitant d'un même topique sont marquées par un registre identique, dont l'étendue diminue et la hauteur s'abaisse progressivement au cours du temps (Brazil, 1980; G. Brown et al., 1980; Yule, 1980; Bruce, 1982; Hirschberg & Pierrehumbert, 1986, 1986; Silverman, 1987; Nakajima & Allen, 1993; Swerts & Geluykens, 1993). Un abaissement final participerait enfin à la clotûre d'un topique majeur où plus le registre est bas et étroit, plus le degré de finalité d'un énoncé est élevé (Hirschberg & Pierrehumbert, 1986).

5.2 Annotation fonctionnelle

Afin de corréler les variations de registre au changement de topique, nous avons effectué une annotation fonctionnelle de ces derniers, à partir de l'annotation en groupes clitiques, décrite en 2, soit 2h de parole (30 minutes par Corpus).

Notre annotation s'ancre dans la tradition pragmatique de l'analyse du discours. A partir de l'annotation prosodique en groupes clitiques, nous décrivons la structure hiérarchique du discours en termes de structure intentionnelle. Nous considérons donc cette structure comme un ensemble d'intentions et de relations entre ces intentions. Inspirée des travaux de Grosz et Sidner (1986), Grosz et Hirschberg (1992) et Hirschberg et Nakatani (1996), et adoptant

l'annotation simplifiée de Fon (2002) et Kong (2004), nous reconnaissons 3 niveaux d'intentions de discours (*discourse segment purpose*), désormais DSP : l'étiquette DSP0 signifie que les deux unités consécutives en question partagent de mêmes intentions de communication, i.e. un même topique, et une même relation avec les unités qui les dominent. Un DSP0 indique donc qu'il n'y a pas de rupture entre deux unités. L'étiquette DSP1 se trouve à un niveau supérieur de la structure hiérarchique intentionnelle par rapport à DSP0. Elle est positionnée entre deux unités si ces dernières partagent un même topique et si une information nouvelle est apportée par la seconde unité. Le DSP1 correspond donc aux catégories ajout (*addition*), clarification (*clarification*) et élaboration (*elaboration*) de Nakajima et Allen (1993). L'étiquette DSP2 se trouve au sommet de la hiérarchie intentionnelle. Elle signifie que deux unités consécutives ne partagent pas un même topique. Elle indique donc un changement de topique.

L'annotation des 3 niveaux d'intention de discours s'est faite à partir de la transcription orthographique afin d'éviter tout problème de circularité. Swerts et Geluykens (1994) expliquait en effet que déterminer la structure du discours en s'appuyant sur des critères prosodiques ne permettait pas de rendre compte correctement du rôle joué par la prosodie dans la démarcation des topiques. Les instructions données aux annotateurs sont celles présentées dans le paragraphe précédent. Ils se sont aussi aidés de la présence vs. absence de marqueurs de liaison (e.g. « de plus ») et du type de marqueurs.

L'annotation effectuée résulte du travail de deux annotateurs (une collègue et moi-même). Nous entendons bien que l'avis d'autres annotateurs auraient pu conforter les choix de l'annotation et ainsi la rendre « plus juste », cependant, elle reste, quel que soit le nombre d'annotateurs, subjective. L'annotation mériterait d'être affinée afin de mieux comprendre le lien entre variations de registre et fonctions linguistiques.

A partir de cette annotation donc, nous avons cherché à voir la façon dont les changements de topique sont caractérisés par les variations de registre. C'est ce que nous présentons dans la section suivante.

5.3 Analyses statistiques

93

Nous avions expliqué que l'algorithme ADoReVA permettait également de corréler des annotations fonctionnelles à la détection automatique des variations de registre. Suite à l'étape 5 donc, nous obtenons la table 59 que nous utilisons pour nos analyses statistiques. Elle permet

93. Les données à partir desquelles ont été effectuées les analyses sont sur CD-ROM, ANNEXES_CHAP3 : Table2.

ainsi d'observer les possibles corrélations entre l'annotation en DSP d'un côté et la hauteur (KEY), l'étendue (RANGE), la différence de hauteur (DIFFKEY), la différence d'étendue (DIFFRANGE), la distance euclidienne (EUCLY), les distances entre les noeuds feuilles obtenues à gauche et à droite pour chacun des paramètres (LEFTDISTK, RIGHTDISTK, LEFTDISTR, RIGHTDISTR, LEFTDISTE et RIGHTDISTE), i.e. à partir de DIFFKEY, DIFFRANGE et EUCLY, de l'autre côté.

FILENAME	UNITS	HERTZ	KEY	RANGE	EUCLY	DIFFKEY	DIFFRAN	LEFTDSP	RIGHTDSP	LEFTDISTK	RIGHTDISTK	LEFTDISTR	RIGHTDISTR	LEFTDISTE	RIGHTDISTE
13aAC1tw	lepremier	307	8.263	1.33	NA	NA	NA	DSP2	DSP0	49.166	0.041	7.918	0.006	49.8	0.041
13aAC1tw	ministre	299	8.222	1.324	0.042012	0.041478	0.006678	DSP0	DSP0	0.041	0.294	0.006	0.047	0.041	0.298
13aAC1tw	iratil	247	7.948	1.28	0.277717	0.274186	0.044144	DSP0	DSP0	0.294	0.768	0.047	0.123	0.298	0.778
13aAC1tw	àbeaulieu	166	7.376	1.188	0.578529	0.571174	0.091959	DSP0	DSP2	0.768	1.751	0.123	0.282	0.778	1.773
13aAC1tw	levillage	284	8.148	1.312	0.781168	0.771236	0.124169	DSP2	DSP0	1.751	0.36	0.282	0.058	1.773	0.365
13aAC1tw	debeaulieu	226	7.819	1.259	0.333044	0.32881	0.052938	DSP0	DSP0	0.36	0.062	0.058	0.01	0.365	0.063
13aAC1tw	estengrand	236	7.881	1.269	0.063127	0.062324	0.010034	DSP0	DSP0	0.062	0.651	0.01	0.105	0.063	0.659
13aAC1tw	émoi	167	7.381	1.188	0.506448	0.500009	0.080502	DSP0	DSP1	0.651	3.087	0.105	0.498	0.659	3.127

FIGURE 59 – Extrait de la table obtenue pour le locuteur 13aACtw du corpus PFC. La colonne FILENAME indique le locuteur, UNITS l'unité en question, HERTZ la hauteur de l'unité en Hz, KEY la hauteur de l'unité en log de base 2, RANGE l'étendue en log de base 2, DIFFKEY, la différence de hauteur entre deux unités, DIFFRANGE, la différence d'étendue entre deux unités, EUCLY, la distance euclidienne entre deux unités, LEFTDSP le DSP à gauche de l'unité, i.e. entre une unité et sa précédente, RIGHTDSP le DSP à droite de l'unité, i.e. entre une unité et sa suivante, LEFTDISTK et RIGHTDISTK la distance à gauche et à droite de l'unité pour le paramètre DIFFKEY, LEFTDISTR et RIGHTDISTR, pour le paramètre DIFFRANGE, et LEFTDISTE et RIGHTDISTE, pour le paramètre EUCLY.

Afin de modéliser la relation entre l'annotation fonctionnelle et les valeurs de hauteur et d'étendue du registre, nous avons mené plusieurs analyses de variance. Nous avons en effet étudier l'effet du facteur DSP sur les variables quantitatives KEY, RANGE, DIFFKEY, DIFFRANGE, EUCLY, LEFTDISTK, LEFTDISTR et LEFTDISTE. Nous proposons de présenter séparément les résultats obtenus pour les corpus PFC, PAC, CID et AM.

5.3.1 PFC : Résultats

Les analyses de variance montrent que la hauteur du registre (KEY), la différence de hauteur entre deux unités (DIFFKEY) et la distance gauche entre les noeuds feuille (LEFTDISTK) sont fortement corrélées aux intentions de discours (DSP) (respectivement : $F(2,1963)=15.58$, p-val=1.927e-07 ; $F(2,1952)=69.36$, p-val$<$ 2.2e-16 ; et $F(2,1963)=110$, p-val$<$ 2.2e-16). Elles montrent également que RANGE, DIFFRANGE, EUCLY, LEFTDISTR et LEFTDISTE sont significativement corrélées à DSP, ce qui n'est pas surprenant puisque RANGE est calculée à

partir de KEY. Plus le DSP est haut dans la structure intentionnelle (e.g. DSP2), plus l'unité qu'il annonce détient un registre haut et étendu, et plus la rupture entre les deux unités qu'il sépare est grande, ce que nous pouvons observer en 60.

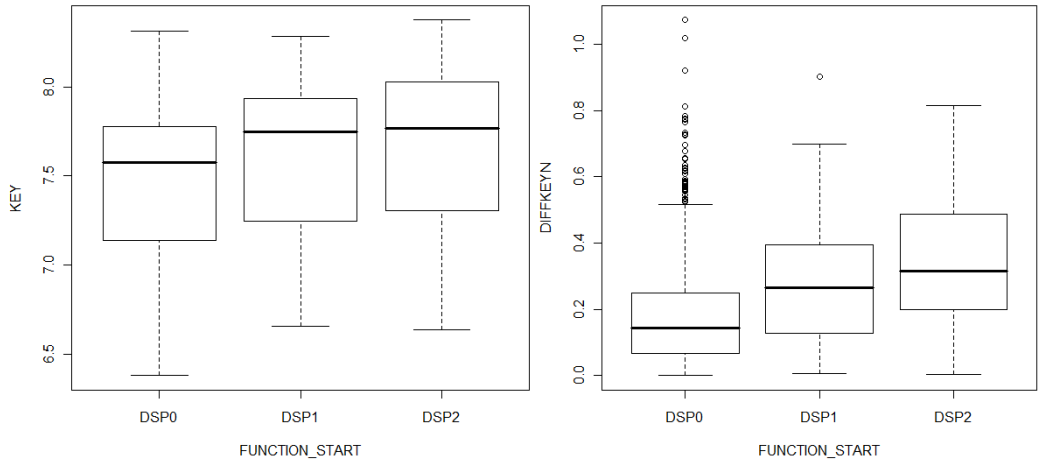


FIGURE 60 – Boîtes à moustaches de la hauteur du registre (KEY) par DSP à gauche ; boîtes à moustaches de la différence de hauteur du registre (DIFFKEY) par DSP à droite - corpus PFC.

Les changements de topique en français et en lecture oralisée sont donc caractérisés par des variations de registre où l'unité qui marque le changement de topique a un registre plus haut et plus étendu que l'unité qui la précède. On observe donc une remise à niveau du registre aux changements de topique.

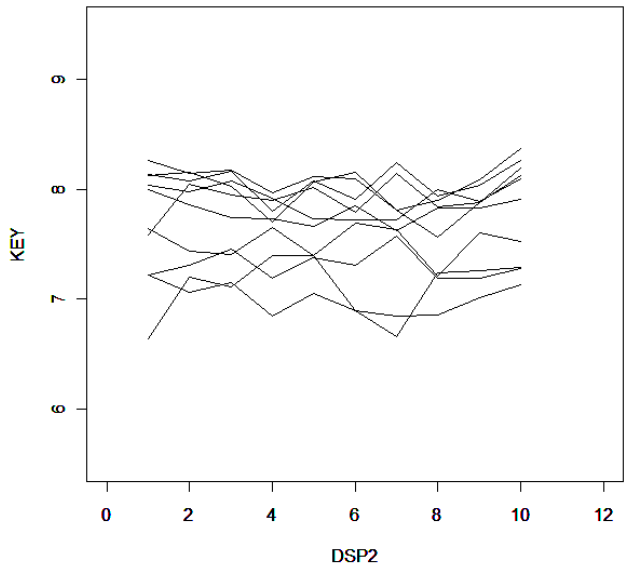


FIGURE 61 – Tracés des DSP2 (changements de topique) pour les 10 locuteurs du corpus PFC.

Au vu du graphique 61, la remise à niveau ne semble pas affectée par la place des DSP2 dans la structure du texte, où un DSP2 en début de texte n'est pas nécessairement plus haut qu'un DSP2 en fin de texte. Nous pouvons par conséquent formuler l'hypothèse qu'il serait plus opportun de rechercher les effets de déclinaison et de remise à niveau partielle au sein d'un domaine délimité par des changements de topique.

5.3.2 PAC : Résultats

Au vu des boîtes à moustaches représentées en 62, on peut se poser la question d'une éventuelle corrélation entre la hauteur du registre (KEY) et les intentions du discours (DSP) et la différence de hauteur entre deux unités (DIFFKEY) et DSP pour le corpus PAC. Bien que la médiane et les quartiles de KEY et de DIFFKEY soient supérieurs pour DSP2 que pour DSP1 et DSP0, les déciles de KEY pour DSP0 et de DIFFKEY pour DSP1 sont, en effet, supérieurs à ceux pour DSP2. Les analyses de variance, pourtant, révèlent que KEY, DIFFKEY et LEFTDISTK sont significativement corrélées à DSP (respectivement : $F(2,2489)=39.78$, $F(2,2481)=155.8$, $F(2,2489)=249.6$; p-val$<$ 2.2e-16). L'étendue étant corrélée à la hauteur, les analyses de variance rapportent également que l'étendue (RANGE), la différence d'étendue entre deux unités (DIFFRANGE), la distance euclidienne (EUCLY), la distance entre les noeuds feuilles calculée à partir du paramètre étendue (LEFTDISTR) et la distance entre les

noeuds feuilles calculée à partir du paramètre EUCLY (LEFTDISTE) sont fortement corrélées à DSP.

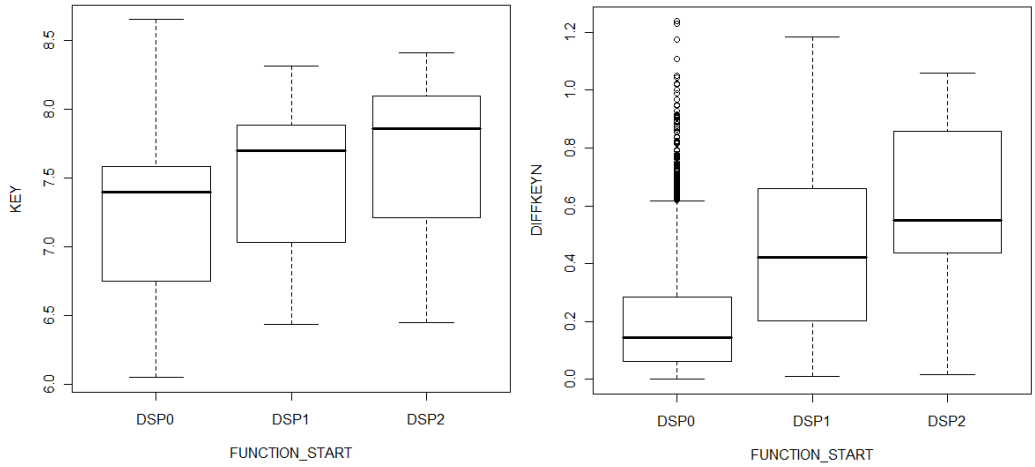


FIGURE 62 – Boîtes à moustaches de la hauteur du registre (KEY) par DSP à gauche ; boîtes à moustaches de la différence de hauteur du registre (DIFFKEYN) par DSP à droite - corpus PAC.

On peut donc conclure, à partir de ces analyses, que, en anglais et en lecture oralisée, les changements de topique sont aussi caractérisés par une rupture du registre, où l'unité annonçant un nouveau topique a un registre plus haut et plus étendu que sa précédente. Au vu du graphique 63, il est intéressant de noter que la structure intentionnelle est similaire en termes de registre pour l'ensemble des locuteurs. Les trois premiers changements de topique semblent marqués par une remise à niveau totale alors que les deux derniers seraient caractérisés par une remise à niveau partielle. Nous pouvons formuler l'hypothèse selon laquelle l'abaissement du registre au niveau des deux derniers changements de topique résulterait de la structure même du texte et indiquerait que le locuteur est sur le point de finir sa lecture. Une autre hypothèse serait que ces deux derniers changements de topique ne se situent pas au même niveau de la structure intentionnelle, i.e. à un niveau inférieur (de type DSP1) par rapport aux autres changements de topique.

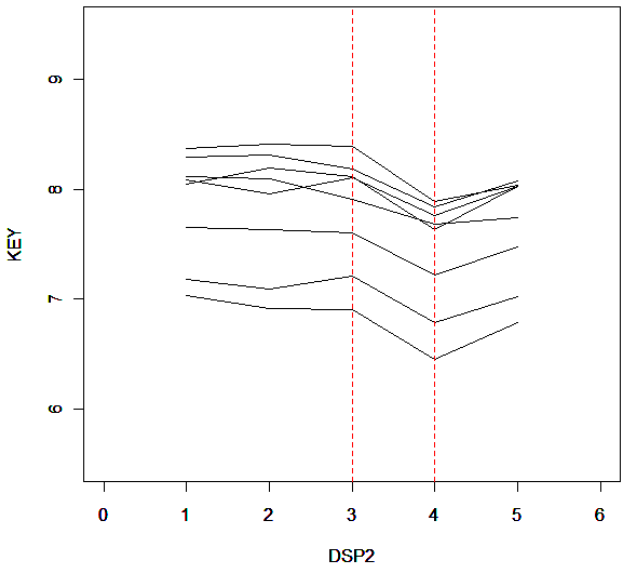


FIGURE 63 – Tracés des DSP2 pour les 8 locuteurs du corpus PAC.

5.3.3 CID : Résultats

Si nous observons la médiane et les quartiles de la hauteur (KEY) et de la différence de hauteur (DIFFKEY) pour DSP0, DSP1 et DSP2 à partir des boîtes à moustaches données en 64, il semble que les niveaux DSP0 et DSP1 ne soient pas caractérisés par des variations de hauteur, et donc, par des variations d'étendue. Les analyses ANOVA confirment nos prédictions. Elles montrent que KEY n'est pas corrélée à DSP1 (p-val=0.01729) et que l'effet de DSP2 sur KEY est faible (p-val= 0.0096). L'effet de DSP1 sur DIFFKEY est également faible (p-val=0.0349) alors que DSP2 est fortement corrélé à DIFFKEY (p-val< 2.2e-16). L'analyse de variance pour LEFTDISTK révèle que LEFTDISTK n'est pas corrélée à DSP1 (p-val=0.154) mais qu'elle est fortement corrélée à DSP2 (p-val< 2.2e-16). L'étendue (RANGE) étant corrélée à la hauteur, nous comprenons aisément qu'elle n'est pas corrélée à DSP1 et très peu à DSP2, que la différence d'étendue (DIFFRANGE) et la distance euclidienne (EUCLY) sont faiblement corrélées à DSP1 et fortement corrélées à DSP2 et que LEFTDISTR et LEFTDISTE ne sont pas corrélées à DSP1 et significativement corrélées à DSP2.

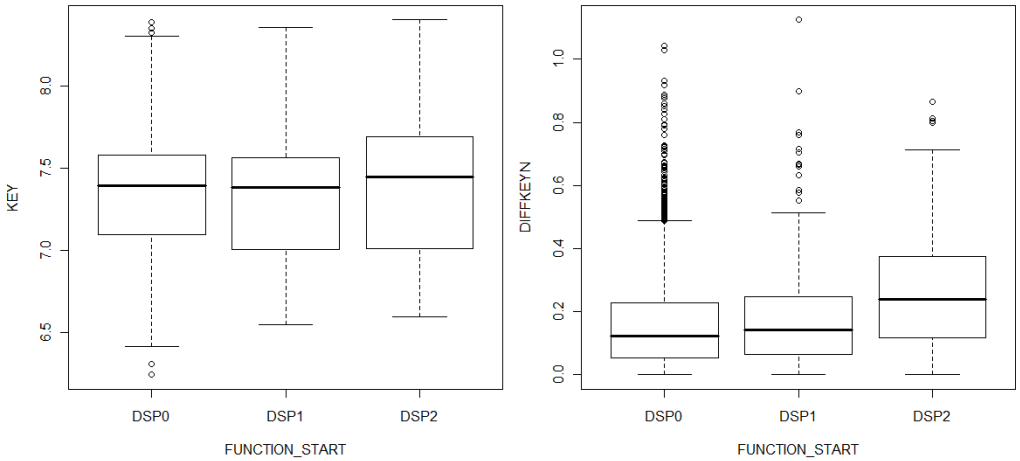


FIGURE 64 – Boîtes à moustaches de la hauteur du registre (KEY) par DSP à gauche ; boîtes à moustaches de la différence de hauteur du registre (DIFFKEYN) par DSP à droite - corpus CID.

En français et en parole conversationnelle donc, les changements de topique (DSP2) sont caractérisés par des variations de hauteur et d'étendue de registre. Les intentions de niveau inférieur (DSP1) ne le sont pas. On n'observe en effet aucune différence significative entre les niveaux DSP1 et DSP0 en termes de registre. Nous pouvons formuler l'hypothèse qu'en parole conversationnelle, les ruptures de registre sont uniquement utilisées à des hauts niveaux de la structure intentionnelle.

5.3.4 AM : Résultats

Au vu des boîtes à moustaches qui figurent en 65, il semblerait que la hauteur du registre (KEY) ne soit corrélée à DSP et que seul DSP2 ait un effet sur la différence de hauteur (DIFFKEY). Les analyses ANOVA révèlent que DSP1 est faiblement corrélée à KEY (p-val=0.00823) quand DSP2 l'est fortement (p-val=9.7e-06). Il est à noter, en revanche, que le registre est plus bas pour DSP1 que pour DSP0, ce qui va à l'encontre d'une augmentation du registre en fonction du niveau de l'intention. DIFFKEY est également faiblement corrélée à DSP1 (p-val=0.0143) et fortement à DSP2 (p-val<2e-16). DSP1 n'a pas d'effet sur LEFTDISTK (p-val=0.826) quand DSP2 en a (p-val=5.41e-12). Les analyses ANOVA montrent également que l'étendue (RANGE), la différence d'étendue (DIFFRANGE) et la distance euclidienne (EUCLY) sont faiblement corrélées à DSP1 et fortement à DSP2 et que la distance entre les noeuds feuilles obtenues avec le paramètre RANGE (LEFTDISTR) et EUCLY (LEFTDISTE)

ne sont pas corrélées à DSP1 et fortement à DSP2.

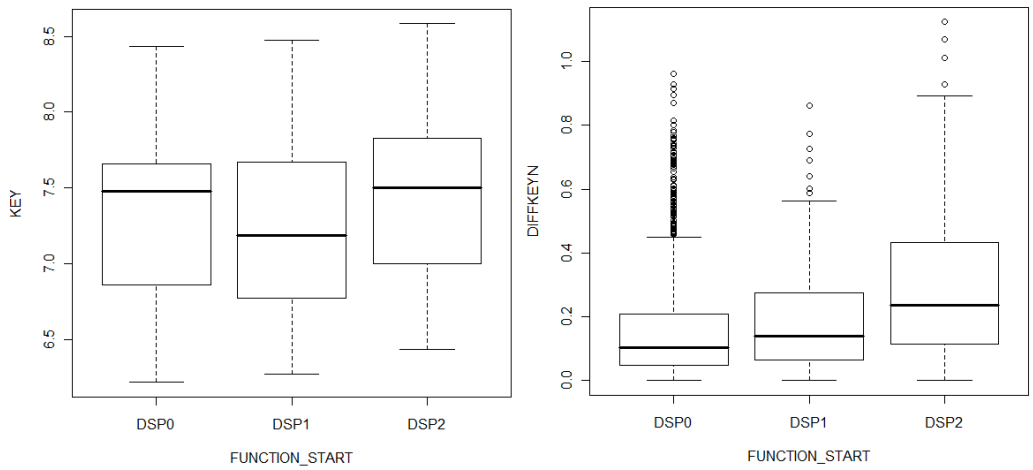


FIGURE 65 – Boîtes à moustaches de la hauteur du registre (KEY) par DSP à gauche ; boîtes à moustaches de la différence de hauteur du registre (DIFFKEYN) par DSP à droite - corpus AM.

En anglais et en parole authentique (i.e. dont le but est de communiquer), les analyses montrent que les changements de topique (DSP2) sont indiqués par des ruptures de registre, où l'unité qui suit la rupture est caractérisée par un registre plus haut et plus étendu que celle qui la précède. Les niveaux d'intention DSP1, en revanche, ne sont pas marqués par des variations de registre. Nous pouvons formuler également ici l'hypothèse qu'en parole authentique, seuls les hauts niveaux de la structure intentionnelle sont caractérisés par des ruptures de registre.

Au vu de ces résultats, et à partir de nos échantillons de données, nous pouvons conclure qu'en français et en anglais, les changements de topique, en lecture oralisée de texte et en parole authentique et conversationnelle, sont marqués par des changements de registre, plus précisément, par une remise à niveau de la hauteur et de l'étendue du registre. Les niveaux d'intentions DSP1, également caractérisés par des variations de registre en lecture oralisée, en revanche ne le sont pas en parole conversationnelle ou authentique.

Avant de conclure cette section, nous portons notre intérêt sur les résidus que nous avions pu observer dans les différents diagrammes de Tukey ou boîtes à moustaches, et ce pour les différents corpus. On peut se demander en effet la raison de leur présence, un point que nous proposons de traiter dans la section suivante.

5.4 Etude des résidus : formulation d'hypothèses

Si nous nous intéressons à nouveau aux figures 60, 62, 64 et 65, nous pouvons observer, dans la fenêtre droite (DIFFKEY), un certain nombre de résidus au niveau des DSP0. Dans la fenêtre gauche (KEY), les déciles des boîtes à moustaches obtenues pour les DSP0 révèlent que la répartition des DSP0 englobe celles des DSP1 et DSP2. On peut donc se poser la question de la pertinence de nos modèles. Les figures 66, 67 et 68 présentent les résidus des modèles KEY, DIFFKEY et LEFTDISTK respectivement pour le corpus PFC. Les « QQplot » utilisés pour la représentation et l'analyse des résidus de nos modèles montrent une non-linéarité de nos échantillons de données pour les trois modèles obtenus à partir du corpus PFC, un fait qui s'avère identique pour les trois autres corpus. Quel que soit le corpus donc et quel que soit le modèle, nos données ne suivent pas une loi normale.

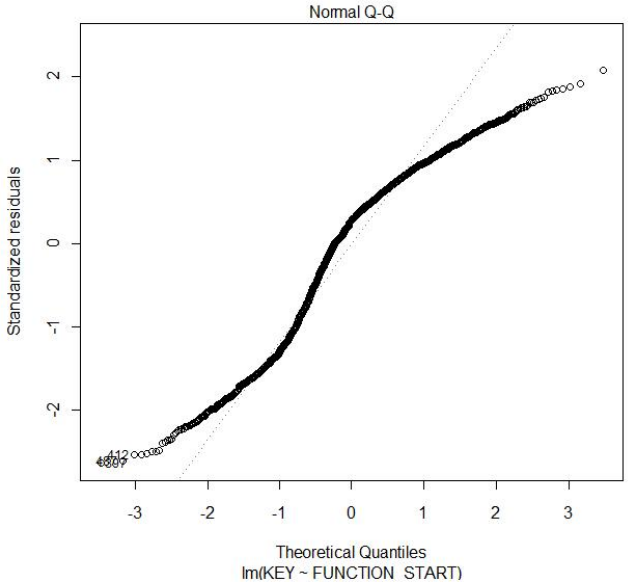


FIGURE 66 – Représentation graphique des résidus du modèle KEY - corpus PFC.

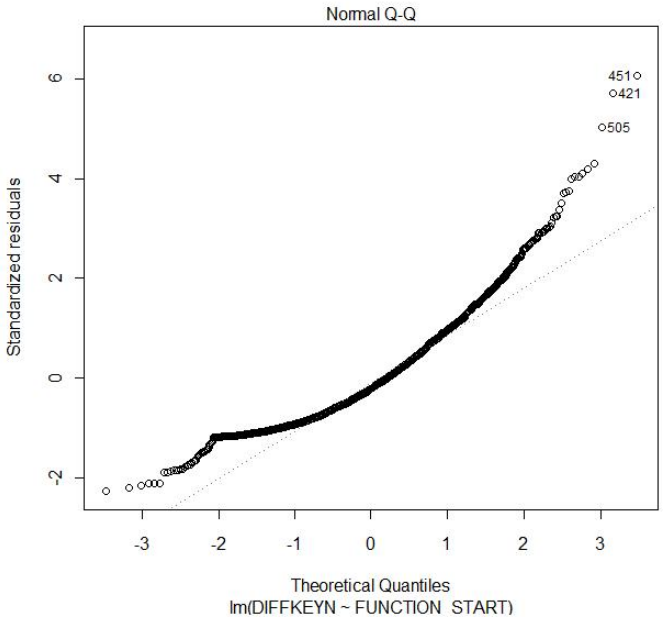


FIGURE 67 – Représentation graphique des résidus du modèle DIFFKEY - corpus PFC.

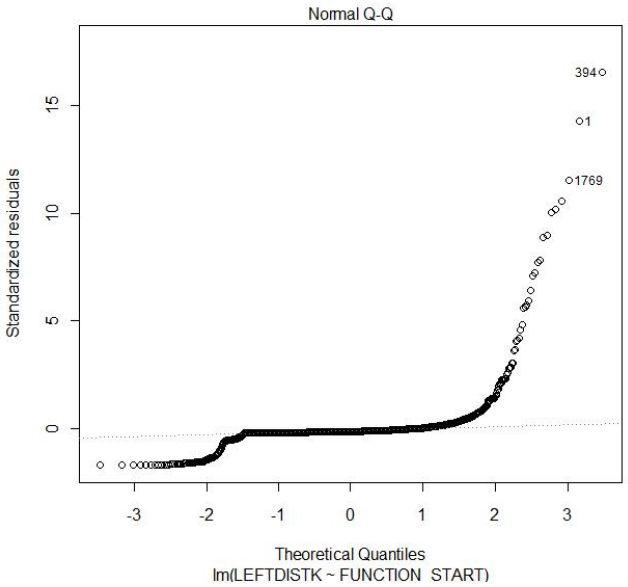


FIGURE 68 – Représentation graphique des résidus du modèle LEFTDISTK - corpus PFC.

Doit-on pour autant réfuter les conclusions portées dans les sections précédentes ? Nous ne le

pensons pas. Nous pensons plutôt que l'analyse des résidus révèle un point intéressant : si la répartition des données pour DSP0 englobe celle pour DSP1 et DSP2, cela montre que des variations de registre ont lieu en dehors des changements de topique. Sachant que nous avons porté notre attention sur les changements de topique uniquement, l'annotation DSP0 devient alors, en termes de registre, l'étiquette « fourre-tout ». Elle indique une continuité de topique soit, mais elle devient par conséquent également l'étiquette d'autres annotations fonctionnelles possibles, comme celle de l'information nouvelle, du focus, du changement d'état émotionnel du locuteur, soulevés dans l'introduction de cette thèse. Sachant que les variations de registre n'indiquent pas uniquement des changements de topique, la catégorie DSP0 se trouve donc certainement porteuse de variations revêtant d'autres fonctions.

Puisque les variations de registre sont quelque part corrélées aux changements de topique pour l'ensemble des corpus, nous pouvons nous poser la question de savoir si, à partir de la détection automatique des variations de registre que permet l'algoritme ADoReVA, il est possible de prédire des changements de topique. Nous traiterons donc ce point dans la section suivante. Nous pouvons cependant d'ores et déjà émettre l'hypothèse qu'une telle prédiction sera difficile de par le fait que d'autres variations de registre, revêtant des fonctions différentes, risquent d'interférer dans le score de la prédiction.

5.5 Prédiction des changements de topique à partir de la détection automatique des variations de registre

Nous avons montré précédemment qu'il existait une corrélation entre les variations de registre et la structure intentionnelle du discours. Nous étudions ici la possibilité d'utiliser cette corrélation afin de prédire la position temporelle des marqueurs de discours à partir des variations du registre. Pour ce faire, nous utilisons des classifieurs bayésiens qui permettent, en fonction de la valeur d'un paramètre discret ou continu (e.g. la différence de hauteur de registre entre deux unités), d'attribuer une classe particulière (e.g. présence d'une frontière DSP2). Les classifieurs nécessitent un apprentissage sur corpus pendant lequel le modèle statistique est extrait des données. Après cette phase d'apprentissage, le classifieur est utilisé pour prédire les classes compte tenu de la valeur du paramètre observé. Le pouvoir de prédiction d'un classifieur est généralement mesuré par des scores de précision, rappel et F-Mesure. L'évaluation du classifieur est effectuée à partir d'un jeu de données pour lequel les classes ont été observées. Le principe consiste alors à comparer les classes prédites par le classifieur aux classes réellement observées, en formant la matrice de confusion qui associe, à chaque couple classe observée i - classe prédite j, le nombre d'occurrences N_{ij} rencontrées dans l'échantillon d'évaluation. Les scores de précision, rappel et F-Mesure pour chaque classe k sont obtenus à partir de la matrice de confusion par la formule suivante :

$$\texttt{rappel} = \frac{N_{kk}}{\sum_j N_{kj}} \; ; \; \texttt{precision} = \frac{N_{kk}}{\sum_i N_{ik}} \; ; \; \texttt{Fmesure} = \frac{2 \times \texttt{rappel} \times \texttt{precision}}{\texttt{rappel} \times \texttt{precision}} \tag{19}$$

Plus ces scores sont proches de l'unité, meilleur est le pouvoir de prédiction du classifieur.

Nous avons considéré 2 classes pour la prédiction des changements de topique : l'une correspondant aux unités étiquetées DSP0, l'autre aux unités DSP1 et DSP2. Ce choix a été effectué compte tenu de l'importante assymétrie de la distribution de nos classes. Les DSP0 à eux seuls représentent en effet plus de 90% des données. L'entrainement des classifieurs binaires se fait à partir de nos échantillons de données (i.e. à partir des corpus PFC, PAC, CID et AM). Les paramètres d'entrée pour les différents classifieurs sont la hauteur de registre (KEY), la différence de hauteur entre deux unités (DIFFKEY) et la distance entre les noeuds feuilles (LEFTDISTK) et leurs combinaisons. Les paramètres RANGE, DIFFRANGE, LEFTDISTR, EUCLY et LEFTDISTE ne sont pas utilisés puisqu'ils sont calculés à partir de KEY via la formule $RANGE = 0.161 \times \log_2(KEY)$. L'entrainement des classifieurs, nous l'avons dit, permet la définition d'un seuil qui délimite 2 classes. Une unité sera donc attribuée à la classe 1 si la valeur du paramètre est inférieure au seuil défini ; inversement, elle appartiendra à l'autre classe si la valeur est supérieure au seuil.

L'évaluation des classifieurs est obtenue pour chaque échantillon par la construction de matrices de confusion (i.e. les classes observées vs. les classes prédites). Les scores des F-mesures associés à la prédiction des DSP1|DSP2 sont donnés en 13 [94]. Les scores des classifieurs basés uniquement sur un paramètre montrent que les paramètres DIFFKEY et LEFTDISTK permettent l'obtention de scores plus élevés. La combinaison à deux paramètres la plus optimale est tout naturellement la combinaison DIFFKEY et LEFTDISTK. Le tableau révèle enfin que la combinaison des 3 paramètres n'augmente pas le score obtenu par la combinaison DIFFKEY|LEFTDISTK.

94. Le détail des matrices est sur CD ROM, ANNEXES_CHAP3 : Table3.

Corpus	K	DF	DST	K|DF	K|DST	DF|DST	K|DF|DST
PFC	0.25	0.28	0.3	0.28	0.31	0.31	0.33
PAC	0.3	0.31	0.33	0.32	0.32	0.33	0.33
CID	0.19	0.27	0.27	0.27	0.27	0.27	0.27
AM	0.31	0.38	0.34	0.38	0.34	0.38	0.38

TABLE 13 – Scores des F-mesures pour les classifieurs basés sur les paramètres hauteur (KEY(K)), différence de hauteur (DIFFKEY(DF)), distance entre les noeuds feuilles (LEFTDISTK(DST)), sur la combinaison des paramètres KEY et DIFFKEY, KEY et LEFTDISTK, DIFFKEY et LEFTDISTK, et sur la combinaison des trois paramètres KEY, DIFFKEY et LEFTDISTK.

Cependant, les scores obtenus restent faibles. Une des raisons de la faible prédiction est que la catégorie DSP0, comme nous l'avons dit, comprend des variations de registre aussi importante que pour la catégorie DSP2, des variations qui revêtent d'autres fonctions et donc qui abaissent le score de prédiction. Une autre difficulté dans la prédiction des DSP relève de l'assymétrie de la distribution de nos classes. A partir de classes non équilibrées, il est en effet difficile de prédire les intentions du discours. Mêmes si les taux de prédiction des segments de discours sont dans l'ensemble 4 fois supérieurs à une prédiction au hasard, ils produisent, toutefois, un taux d'erreur d'environ 65%, et, par conséquent, ne permettent pas l'annotation automatique des changements de topique.

5.6 Discussion

L'étude menée corrobore les observations de la littérature quant aux fonctions linguistiques des variations de registre (Brazil, 1980; G. Brown et al., 1980; Yule, 1980; Bruce, 1982; Hirschberg & Pierrehumbert, 1986, 1986; Silverman, 1987; Nakajima & Allen, 1993; Swerts & Geluykens, 1993). Nous avons pu en effet montrer, en français et en anglais, en lecture oralisée et en parole authentique et conversationnelle, que les changements de topique étaient corrélés aux variations de registre. Une unité portant ainsi un nouveau topique est caractérisée par un registre plus haut et plus étendu que celle qui la précède, et la différence de registre entre les deux unités consécutives s'avère être importante. En revanche, lorsque deux unités consécutives partagent un même topique, elles ont un registre équivalent.

Nous avons également montré que les 3 niveaux définis de la structure intentionnelle, i.e. DSP0, DSP1 et DSP2, étaient marqués par des registres différents (plus on monte dans la structure intentionnelle, plus la différence entre deux unités consécutives est grande) en lecture oralisée. En revanche, en parole authentique et conversationnelle, nous avons pu voir que le

niveau DSP1, en fonction des variations de registre, n'était pas significatif. Nous avions alors formulé l'hypothèse que, en parole authentique et conversationnelle, les variations de registre se faisaient plutôt à un haut niveau de la structure intentionnelle. Pour valider un tel fait, il serait intéressant d'affiner l'annotation effectuée des changements de topique et ainsi apprécier une éventuelle gradation des variations de registre en fonction de la structure intentionnelle, en lecture oralisée et en parole authentique.

Cependant, il est à rappeler que les modèles des analyses de variance présentaient des résidus non linéarisés. Cela montre, qu'il est difficile, dès lors que l'on dépasse des corpus de phrases, de neutraliser les variations de registre existantes pour en extraire seulement celles que l'on cherche à étudier. Nous avons en effet montré que l'étiquette DSP0 de notre annotation était une étiquette « fourre-tout », représentant à la fois une continuité de topique mais également toute autre variation de registre existante se situant à ce niveau. La prédiction automatique des fonctions linguistiques et para-linguistiques des variations de registre n'est donc pas près d'être résolue. Elle mériterait en amont, une annotation précise des différentes fonctions linguistiques et para-linguistiques que revêtent les variations de registre afin que chaque étiquette reflète clairement le dessein de l'annotateur. Pour autant, la prédiction restera difficile dans le sens où de mêmes variations de registre peuvent revêtir des fonctions différentes (e.g. un registre plus haut et plus étendu à un changement de topique mais également dans le cas d'emphase) et inversement, dans le sens où de mêmes fonctions linguistiques ou para-linguistiques peuvent être exprimées par des variations de registre différentes, résultant tout simplement du caractère intrinsèque de la parole du locuteur. Si un locuteur peut en effet exprimer des changements de topique par des variations de registre, un autre peut le faire par l'utilisation d'autres paramètres prosodiques, tels que le débit de parole, l'intensité, etc. Parce que les locuteurs peuvent utiliser des stratégies différentes, la prédiction devient difficile. Cependant, on peut envisager que l'intégration de l'ensemble des paramètres, que peut utiliser un locuteur pour exprimer des changements de topique, pourra améliorer le score de prédiction. Nous avons notamment montré (De Looze & Rauzy, 2009) que les scores atteignaient les 75% pour le corpus PFC dès lors que l'on intégrait, dans le classifieur, la pause comme paramètre de prédiction.

6 Conclusion : une synthèse

Dans ce chapitre, nous nous sommes intéressée à la problématique des variations de registre intra-locuteurs, notamment à celle de leur chevauchement et de leur délimitation avec les variations prosodiques à plus court terme et à celle des fonctions qu'elles revêtent. Pour cela, nous avons élaboré un algorithme (ADoReVA), implémenté sous forme de plugin dans PRAAT, à partir duquel les variations de registre entre unités consécutives sont obtenues automati-

quement et représentées sous forme de structures arborescentes binaires, où les ruptures de l'arborescence marquent un changement de registre.

La première étude était donc celle de l'intégration des variations de registre dans le système d'annotation des patrons intonatifs MOMEL-INTSINT. A partir du calcul de la distance des feuilles de l'arborescence, et suite à la définition d'un seuil, nous avons pu montrer que l'intégration des variations de registre, dans un tel système, améliorait significativement l'annotation des patrons prosodiques qu'il propose. Nous avions par conséquent suggéré que la détection automatique des variations de registre devait être intégrée dans d'autres systèmes d'analyse de l'intonation.

Notre deuxième étude a ensuite consisté en la corrélation d'une annotation fonctionnelle en termes de changement de topique aux variations de registre détectées automatiquement. Nous avons pu montrer que les variations de registre étaient en effet corrélées à la structure intentionnelle du discours où un changement de topique est annoncé par une rupture du registre, l'unité portant le nouveau topique ayant un registre plus haut et plus étendu que sa précédente. Suite à ces résultats, nous avons essayé de prédire les changements de topique à partir des variations de registre, mais les scores de prédiction se sont avérés relativement faibles. Nous avons estimé que de tels scores résultaient de la non-neutralisation d'autres variations de registre et révélaient les différentes stratégies que peut utiliser un locuteur dans l'annonce d'un changement de topique.

Nous nous intéressons à présent à l'empan temporel des variations de tempo et aux fonctions qu'elles revêtent.

Chapitre 4

Détection des variations de tempo

1 Problématique

La difficulté qui se pose à la modélisation de l'organisation temporelle de la parole est multiple. Elle repose à la fois sur le choix d'une unité optimale qui capture au mieux les phénomènes temporels et sur la prise en compte des différents facteurs qui influencent les durées inhérentes des segments. Implémenter l'influence de ces facteurs dans des modèles de durée segmentale nécessite en amont de déterminer le domaine sur lequel ils opèrent et le locus de leur effet. Pour notre part, nous nous intéressons à l'empan temporel des variations de tempo et proposons une détection automatique de ce dernier.

Détecter les variations de tempo nécessite en amont de se poser la question de sa mesure, et donc celle de ses composantes, à savoir le débit d'élocution et les pauses. Pour le débit d'élocution, cela requiert le choix d'une unité de mesure à partir de critères de sélection précis. Pour notre part, nous avons déterminé 3 critères : la possibilité d'une étude inter-langues, la variance temporelle et la façon dont l'unité rend compte du tempo perçu. Nous avons conclu dès lors que, parmi les unités choisies dans la littérature, le phonème semblait une unité adéquate pour la mesure du tempo et de ses variations. Nous avons également soulevé les objections qui pouvaient être portées à une quantification du débit d'élocution en termes de nombre de segments par unité de temps ou en termes de durée. Puisque les durées moyennes des phonèmes peuvent être très différentes, nous avons expliqué que le nombre de phonèmes par unité de temps peut être biaisé par la prédominance de phones longs ou courts et que, par conséquent, ce type de mesure ne peut estimer correctement le débit d'élocution. La durée moyenne d'une unité ne semble pas plus satisfaisante, du fait qu'elle dépend du

nombre de segments qui la compose et de son caractère accentuel. Nous avons donc suggéré de mesurer le débit et ses variations comme la différence entre les valeurs prédites (i.e. les valeurs moyennes calculées pour chaque type de phonème à partir de l'ensemble des données) et les valeurs observées. Le choix de cette mesure est aussi motivé par le fait qu'elle reflète assez bien la façon dont les auditeurs perçoivent les variations du débit d'élocution (Campbell, 1988). Quant aux pauses, il s'agit de déterminer leur type (i.e. silencieuse vs. remplie) et fixer un seuil minimum de durée. Pour notre part, nous considérons, dans cette étude, seulement les pauses silencieuses ; nous ne fixons en revanche aucun seuil limite.

L'intérêt que nous portons aux variations de tempo ne s'arrête pas à la détection automatique de leur empan temporel. Nous nous intéressons aussi aux fonctions extra-linguistiques et linguistiques qu'elles revêtent. Notamment, nous nous intéressons à la façon dont les variations de tempo informent de l'identité du locuteur, du style de parole qu'il pratique ou encore à la façon dont elles indiquent des changements de topique.

Ces deux études, celle de l'empan temporel des variations de tempo et des fonctions qu'elles revêtent, a nécessité l'élaboration d'un algorithme de détection automatique des variations de tempo, ADoTeVA, que nous présentons ci-après.

Puisque nous abordons la problématique de l'empan temporel des variations de tempo à travers la présentation de notre algorithme, nous proposons, avant cette présentation, de rappeler les domaines utilisés dans la littérature pour l'étude du débit d'élocution.

2 Rappel : les domaines du débit d'élocution

Nous avons soulevé dans la littérature que plusieurs domaines ont été utilisés pour mesurer le débit d'élocution et ses variations. Ces domaines ont généralement été établis à partir de critères syntaxiques : la phrase, la proposition ou encore le syntagme (groupe verbal, nominal, prépositionnel) (Lobacz, 1976; Eefting, 1991; G. Fant et al., 1991; Campbell & Sagisaka, 1992; Ohno et al., 1996; Verhasselt & Martens, 1996). Un autre domaine que nous avons également mentionné et qui d'ailleurs est largement répandu dans l'étude du débit d'élocution, est l'extrait de parole inter-pausal (Miller et al., 1984; Eefting & Rietveld, 1989; T. Crystal & House, 1990; Walker et al., 1992; Beinum & Donzel, 1996; Tsao & Weismer, 1997). Nous avons vu aussi que l'unité intonative, quand elle ne correspond pas à l'extrait inter-pausal mais se situe plutôt en-deça de ce dernier, permet aussi d'estimer les variations du débit d'élocution.

Bien que ces domaines aient permis l'analyse des variations du débit dans de nombreuses études, il semble cependant, dans quelques travaux de « réplication », que, finalement, ces domaines ne conviennent pas à la mesure du débit d'élocution. Par exemple, alors que l'extrait

de parole inter-pausal a été largement utilisé dans la littérature, Dankovicova (1999), dans son étude, montre qu'en réalité, le débit d'élocution n'est pas constant au sein de ce domaine et que, par conséquent, il ne convient pas à l'étude des variations de débit. L'auteur propose en revanche l'unité intonative.

Ces divergences, que l'on peut trouver, dans la littérature, sont le résultat, à notre sens, de plusieurs facteurs. Tout d'abord, les auteurs ne se basent pas sur des corpus de même type (e.g. lecture de phrases isolées, lecture de textes, conversations spontanées). Or, certains ont montré que le domaine des variations du débit dépend du type de production. Cedergren et Perreault (1994) expliquent, par exemple, que la phrase n'est pas appropriée pour des corpus de parole spontanée, pour laquelle les faux départs et les répétitions sont nombreux, alors qu'elle l'est pour des corpus de phrases ou de textes lus. De plus, les auteurs n'étudient pas les mêmes langues et n'utilisent pas les mêmes unités de mesure. Or, selon l'unité que l'on va choisir, selon la complexité syllabique de la langue, par exemple, les variances vont être plus ou moins importantes et dispersées sur des domaines différents. Enfin, parce que le débit est soumis à l'influence de nombreux facteurs, notamment ceux mentionnés dans la problématique de ce chapitre, le domaine qui avait pu sembler adéquat dans une étude peut s'avérer inadapté dans l'autre, du fait de l'interaction et de l'influence de ces facteurs.

L'élaboration d'un algorithme de regroupement hiérarchique, qui permet la détection automatique des variations de tempo, semble intéressante. En effet, un tel algorithme représente sous forme de structures arborescentes les variations de tempo sur plusieurs niveaux. L'emboitement de ces niveaux serait en effet un moyen de visualiser les changements de tempo sur un large éventail de domaines et ainsi ne pas restreindre leur analyse à un domaine particulier.

3 ADoTeVA : un outil de détection automatique des variations de tempo

A l'image d'ADoReVA, algorithme de détection automatique des variations de registre présenté dans le chapitre 3 de cette thèse, ADoTeVA (*Automatic Detection of Tempo Variations Algorithm*) est un algorithme de regroupement hiérarchique permettant à la fois la mesure des tempos de différents locuteurs et la détection automatique des variations de tempo intra-locuteurs. Conçu pour être implémenté dans Praat, sous forme de plugin, il est accessible à partir du menu New, une fois installé dans les Préférences de Praat. La figure 69 donne un aperçu de l'implémentation du plugin sous praat et de sa composition. L'algorithme est composé de deux parties : la première (*part 1 : inter-analysis*) permet l'étude des tempos inter-locuteurs, la deuxième (*part 2 : intra-analysis*) l'analyse des tempos intra-locuteurs. Elles peuvent donc être lancées séparément. Chaque commande que décline chacune des par-

ties, et que l'on peut observer dans la figure sous forme d'onglets, correspond à un script, i.e. à une étape particulière de l'algorithme. Les trois premières étapes des analyses inter- et intra-locuteurs correspondent sensiblement aux mêmes scripts, nous les commentons donc conjointement.

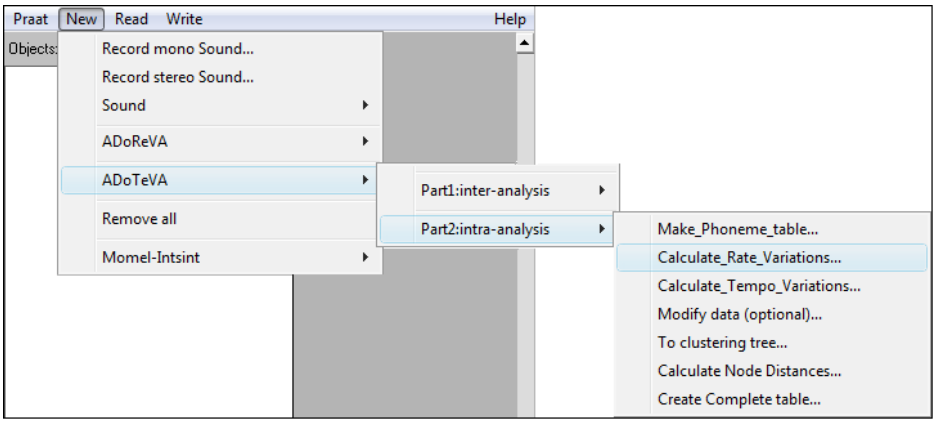


FIGURE 69 – Aperçu du plugin ADoTeVA, tel qu'implémenté sous PRAAT. Les 7 étapes de *part 2 : intra-analysis* apparaissent sous différents onglets, à droite.

3.1 Etape 1 : Création d'une table de phonèmes et de leur durée moyenne

Make_ Phonemes_ TableTotal[95] [96] :

Cette première étape consiste en la constitution d'une base de données dont on se servira comme référence pour la mesure du débit d'élocution inter- et intra-locuteurs. Pour l'étude des différences inter-locuteurs, l'algorithme relève les phonèmes de l'ensemble du corpus utilisé (à partir d'objets TextGrids donnés en entrée et annotés en phonèmes) et calcule la durée moyenne de chacun de ces phonèmes. Pour l'étude des variations intra-locuteurs, l'algorithme liste les phonèmes et calcule leur durée moyenne pour chaque locuteur. Chaque table contenant les phonèmes listés et leur durée respective sera donc utilisée comme référence dans la mesure du débit d'élocution. Une seule table servira de référence pour l'étude des différences inter-locuteurs, quand plusieurs tables, chacune référant à un locuteur en particulier, seront utilisées dans l'étude des variations intra-locuteurs. Un extrait de ces tables obtenues est donné dans le tableau 14.

95. Ce premier script a été écrit par Daniel Hirst.
96. Les scripts que nous allons décrire sont sur CD ROM, dossiers SCRIPTS, ADOTEVA.

ROW_LABEL	DURATION
@	0.0761254564275832
@U	0.11035712530671413
A :	0.13099999446296112
D	0.03687499818315772
I	0.05481327479176294
I@	0.09466665236265263
N	0.10210525587232079
O :	0.11749999610677159
OI	0.17249999999999166

TABLE 14 – Extrait des données de sortie obtenues par le script *Make_ Phonemes_ Table*. La colonne ROW_LABEL liste les phonèmes, la colonne DURATION indique leur durée respective (en secondes), pour le locuteur J01SP1M - Corpus AM.

3.2 Etape 2 : Calcul des différences et des variations de débit inter- et intra- locuteurs

Partie 1 - *Calculate_ Speakers_ Rate* : Pour cette étape de l'algorithme, les données d'entrée sont un objet TextGrid pour chaque locuteur dans lequel sont annotés les phonèmes et une table de référence comprenant les durées moyennes des phonèmes calculées pour l'ensemble des locuteurs. L'algorithme calcule le débit d'élocution intrinsèque de chaque locuteur comme la moyenne des rapports entre la somme des durées observées des phonèmes, i.e. obtenues à partir de l'objet TextGrid, et la somme des durées des phonèmes prédites, i.e. à partir de la table de référence.

Partie 2 - *Calculate_ Rate_ Variations*[97] : Dans l'étude des variations de débit intra-locuteurs, les données d'entrée sont un objet TextGrid dans lequel sont annotés les phonèmes sur une couche (*tier*) et les unités à partir desquelles on souhaite étudier les variations de débit sur une autre couche, et une table de référence, comprenant les durées moyennes des phonèmes calculées pour le locuteur en question. Il est à noter que, comme pour ADoReVA, le choix des unités dans cet algorithme revient à l'utilisateur. Ici nous avons utilisé, comme dans le chapitre 3, les groupes clitiques, bien que nous reconnaissons le choix de cette unité discutable. Nous reviendrons sur ce point dans la discussion finale du chapitre. L'algorithme calcule d'abord le débit d'élocution pour chaque unité comme le rapport entre la somme des durées observées des phonèmes, i.e. obtenues à partir de l'objet TextGrid, et la somme des durées des phonèmes prédites, i.e. à partir de la table de référence. Une fois le débit (RATE) obtenu pour chacune

97. Ce script a été élaboré conjointement par Daniel Hirst et nous-même.

des unités, l'algorithme compare deux à deux les unités et calcule ainsi la différence de débit (DRATE) entre ces deux unités. Les différences obtenues sont donc à chaque fois calculées sur des unités consécutives. Ci-après, la formule utilisée pour ce calcul :

$$DTEMPO = |RATE_{unit} - RATE_{prevUnit}| \tag{20}$$

où $RATE_{unit}$ est le débit d'élocution calculé sur l'unité en question et $RATE_{prevUnit}$ le débit calculé sur l'unité précédente. Les résultats obtenus sont ainsi rendus sous format tabulaire, un extrait est donné dans le tableau 15.

NAME	NPHO	UNITS	START	END	PRED	OBS	RATE	DRATE
J01SP1M	7	Paddy'not	0.105	0.445	0.471	0.440	0.934	NA
J01SP1M	8	long'/after	0.445	1.005	0.615	0.590	0.959	0.025
J01SP1M	3	that |	1.005	1.285	0.184	0.400	2.171	1.212
J01SP1M	5	the'tears	1.285	2.005	0.443	0.870	1.962	0.209
J01SP1M	6	of grief |	2.005	2.745	0.502	0.800	1.594	0.368

TABLE 15 – Extrait des données de sortie obtenues par le script *Calculate_ Rate_ Variations*, pour le locuteur J01SP1M - Corpus AM. La colonne NAME indique le nom du fichier, NPHO le nombre de phonèmes dans l'unité, UNITS l'unité en question, START et END le début et la fin de l'unité en question, PRED le débit d'élocution prédit (à partir de la table de référence), OBS le débit observé (à partir de l'objet TextGrid), RATE le débit calculé pour l'unité et DRATE la différence de débit entre l'unité en question et l'unité précédente.

3.3 Etape 3 : Calcul des différences et des variations de tempo inter- et intra- locuteurs

Partie 1 - *Calculate_ Speakers_ Tempo* Partie 2 - *Calculate_ Tempo_ Variations*

Cette troisième étape intègre dans le calcul du tempo celui des pauses. Elle consiste en effet en la création d'une table de données qui répertorie les durées des pauses détectées et leur nombre. Pour l'analyse inter-locuteurs, la table contient la durée moyenne des pauses pour chaque locuteur, l'écart type par rapport à la moyenne, le rapport entre la moyenne des pauses obtenue par locuteur et la moyenne des pauses obtenue pour l'ensemble des locuteurs, le rapport entre l'écart type obtenu par locuteur et l'écart type obtenu pour l'ensemble des locuteurs, le nombre de pauses, la durée totale de l'enregistrement et le rapport entre le nombre de pauses et la durée totale. Un extrait de la table des données est présenté dans le tableau 16.

NAME	μP (sec)	σP	$R\mu P$	$R\sigma P$	RATE (sec)	RR	NP	DUR	RNP
A0101B	0.273	0.345	0.679	1.106	0.07282	0.93953	35	58.6417	0.59
A0202G	0.314	0.302	0.780	0.968	0.07706	0.99421	43	57.3902	0.74
A0301B	0.380	0.371	0.946	1.190	0.08482	1.09438	28	60.532	0.46
A0501G	0.438	0.358	1.088	1.148	0.07831	1.01031	30	52.3978	0.57
A0601B	0.366	0.250	0.910	0.802	0.07484	0.96556	27	59.4534	0.45
A0701G	0.259	0.351	0.644	1.124	0.07320	0.94446	37	56.3188	0.65

TABLE 16 – Valeur obtenue de tempo pour chaque locuteur : NAME indique le locuteur ; μP, la durée moyenne des pauses ; σP, l'écart type par rapport à la moyenne ; $R\mu P$, le rapport entre la durée moyenne des pauses obtenue par locuteur et la durée moyenne obtenue pour l'ensemble des locuteurs ; $R\sigma P$, le rapport entre l'écart type obtenu par locuteur et l'écart type obtenu pour l'ensemble des locuteurs ; RATE (sec), la durée moyenne des phonèmes ; RR, le rapport entre la durée moyenne des phonèmes obtenue par locuteur et la durée moyenne des phonèmes obtenue pour l'ensemble des locuteurs ; NP, le nombre de pauses ; DUR la durée totale de l'enregistrement ; et RNP, le rapport entre le nombre de pauses et la durée totale.

Pour l'analyse intra-locuteurs, la table de données répertorie les durées de chaque pause, les rapports entre les durées de ces pauses et la durée moyenne des pauses obtenue par locuteur, le nombre de pauses et la durée totale du fichier.

Les étapes 4, 5, 6 et 7 ont été développées pour l'analyse du débit d'élocution intra-locuteurs.

3.4 Etape 4 : Mise en forme des données (optionnel)

Modify data (optional)... : Cette troisième étape consiste en la mise en forme des données, i.e. à un « nettoyage » des éventuels symboles pouvant inférer dans l'affichage .xml des données de sortie de l'étape 5. Les données de sortie ne doivent pas par exemple contenir les symboles « ' » ou encore « < ». Si les TextGrids contiennent ces symboles, il est donc préférable de passer par cette quatrième étape. Dans le cas contraire, elle n'est pas nécessaire et l'utilisateur peut directement passer à l'étape suivante.

3.5 Etape 5 : Classification ascendante hiérarchique - création de dendrogrammes

To clustering tree... :

Dans cette étape, l'algorithme de regroupement hiérarchique groupe les unités en fonction de leurs différences de débit. Pour chaque locuteur, l'algorithme détecte la différence la plus petite entre deux unités et effectue leur regroupement. Ces unités regroupées forment une nouvelle unité, pour laquelle le débit et la différence de débit avec l'unité précédente sont recalculés. La différence est obtenue par la moyenne pondérée des distances au barycentre des deux unités regroupées (cf. formule 18 dans le chapitre 3 de cette thèse). La procédure est itérative et ce, jusqu'à ce qu'il ne reste plus d'unités ou de groupes d'unités à embrancher. L'algorithme regroupe d'abord deux unités consécutives dont la différence de débit est la plus petite. A contrario, plus la différence entre deux unités est grande, au plus tard s'effectue leur embranchement. Pareillement à ADoReVA, l'algorithme de regroupement hiérarchique que nous proposons est donc similaire aux algorithmes de regroupement hiérarchique existants mais il a la contrainte de regrouper uniquement des unités consécutives, i.e. d'effectuer leur embranchement en respectant leur ordonnée temporelle.

L'algorithme génère ensuite une structure arborescente binaire sous forme de diagramme à niveaux alignés (*layered icicle diagram*). Le dendrogramme obtenu permet alors la visualisation des changements de débit d'élocution intra-locuteurs et ainsi la structure hiérarchique et l'organisation relationnelle des unités telles qu'elles sont reflétées par les changements de débit. Il est en effet possible, à partir de la structure arborescente, de distinguer des groupes d'unités, à travers des cassures visuelles de l'arborescence. Plus la cassure est grande entre deux unités, plus la différence de débit entre ces deux unités ou groupes d'unités est importante. Nous proposons un extrait du dendrogramme obtenu pour le locuteur J01SP3M (corpus AM) en 70.

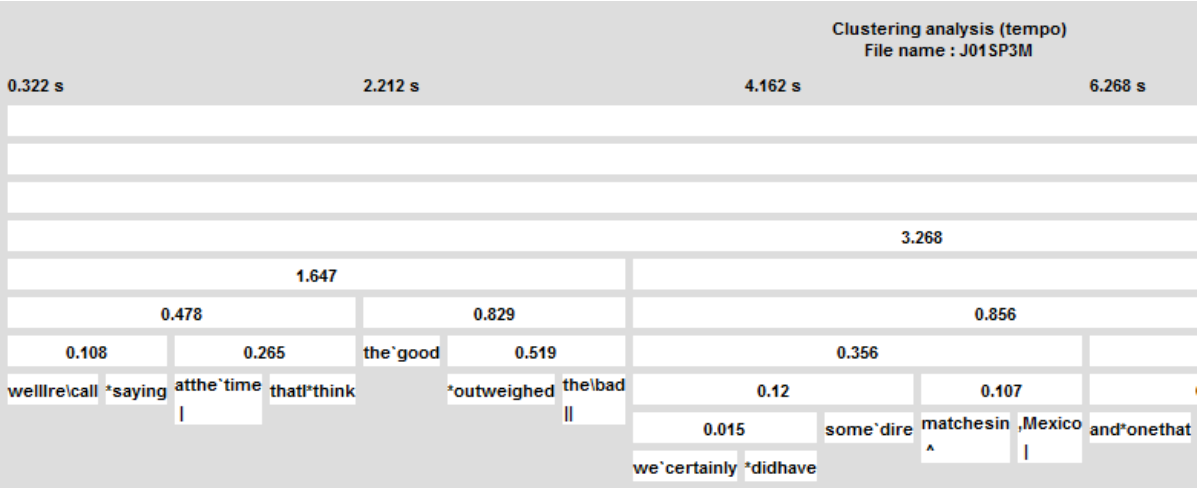


FIGURE 70 – Aperçu du dendrogramme obtenu à partir des différences de débit pour le locuteur J01SP3M (corpus AM). En haut à droite est donné le nom du fichier traité. Au dessus du dendrogramme est indiqué le temps en secondes. Les feuilles au bas de la structure arborescente représentent les unités à partir desquelles sont calculées le débit d'élocution.

Les feuilles de la structure arborescente correspondent aux unités à partir desquelles les variations de débit ont été calculées. Nous pouvons observer une cassure importante entre l'unité intonative majeure (indiquée, dans le corpus Aix-marsec, par le symbole ||) « well I recall ... outwheighed the bad » et l'unité « we certainly did have ... one that... » qui la suit. Nous pouvons également voir une autre cassure, un peu moins importante mais qui signalerait aussi un changement de débit, entre l'unité intonative mineure (marquée du symbole |, dans le corpus Aix-Marsec) qui termine par « Mexico » et celle qui commence par « and one that ». Les unités intonatives majeures et mineures (*major and minor tone-units*) ont été annotées dans ce corpus à partir de critères phonétiques tels que la durée de la pause à la frontière de l'unité ou encore l'allongement de la syllabe qui précède la frontière. Elles ne sont pas déterminées à partir de critères syntaxiques. Les unités majeures correspondent toutefois en général à des phrases, les unités mineures à des propositions ou des syntagmes. La différence majeure entre les unités intonatives majeures et mineures est la durée de la pause à leur frontière (plus longue pour les unités majeures) (B. Williams, 1996, p51). Le taux d'accord inter-annotateurs est de 63 % (Pickering, Williams, & Knowles, 1996, p66).

Ces cassures relèvent ici un point important, celui d'un plus fort changement de débit d'élocution à de hauts niveaux de la structure prosodique. En effet, la différence entre le groupe « well I recall ... outwheighed the bad » et le groupe « we certainly did have ... one that... » est estimée à 3.268 alors qu'elle n'est estimée qu'à 0.856 entre l'unité « Mexico »et « one that ». Les analyses statistiques effectuées viennent d'ailleurs confirmer l'intuition graphique.

Nos résultats montrent en effet que la différence de débit d'élocution entre deux mots à la frontière d'unités intonatives majeures est plus importante que celle entre deux mots situés à la frontière d'unités intonatives mineures, elle-même supérieure à celle entre deux mots au sein de l'unité ($F(2, 3234)=74.9$; p-val<2e-16). Ces résultats laissent présager qu'il n'existe pas *un* domaine de variations du tempo mais *des* domaines et que ces domaines sont emboités. Ainsi, comme pour les variations de registre, nous défendons l'idée qu'une étude de l'empan temporel des variations de tempo ne peut s'effectuer sans la considération de l'emboitement de ces effets. Cet aspect montre que l'étude de l'empan temporel est complexe, et qu'en l'état, nous ne pouvons le déterminer. Nous pensons cependant que l'intégration de ces variations, par une détection automatique qui prend en compte leur emboitement, permettrait d'améliorer la modélisation du rythme de la parole et aussi, en retour, de déterminer leur empan temporel.

Puisque les ruptures visuelles retracent les variations de débit, il devient donc intéressant de calculer ces ruptures, i.e. les distances entre les feuilles de la structure arborescente.

3.6 Etape 6 : Calcul des distances entre les feuilles de la structure arborescente

Calculate Node Distances... : La sixième étape consiste à calculer les distances entre les noeuds feuilles de la structure arborescente. Les noeuds feuilles correspondent aux unités à partir desquelles ont été calculées les variations de débit. Le degré de rupture entre deux unités consécutives est ainsi obtenu. Plus la distance est grande entre deux unités, plus la différence de débit entre ces deux unités est importante. Inversement, une petite distance entre deux unités signifie que les deux unités sont énoncées sur un même débit. La table de données de sortie, obtenue suite au lancement de ce script, pour le locuteur J01SP1M (corpus Aix-Marsec) figure dans le tableau 17.

LEAF	LEFTDIST	RIGHTDIST	START	END
Paddy‘not	114.208	0.025	0.105	0.445
long‘/after	0.025	1.486	0.445	1.005
that |	1.486	0.209	1.005	1.285
the‘tears	0.209	0.472	1.285	2.005
of grief |	0.472	2.519	2.005	2.745

TABLE 17 – Extrait des données de sortie obtenues par le script *Calculate Node Distances...*, pour le locuteur J01SP1M - Corpus AM. La colonne LEAF indique les unités traitées, la colonne LEFTDIST la distance calculée à gauche de l'unité, i.e. entre une unité et sa précédente, la colonne RIGHTDIST la distance calculée à droite de l'unité, i.e. entre une unité et sa suivante ; la colonne START donne la valeur temporelle du début de l'unité, la colonne END celle de la fin de l'unité.

3.7 Etape 7 : Vers une analyse statistique des données

Create Complete Table... : Dans cette dernière étape, l'algorithme regroupe sous forme de tableau les fiches signalétiques des locuteurs (sexe, langue, type de production qu'il pratique) et les variations de tempo inter-locuteurs obtenues lors des étapes précédentes ; il regroupe aussi, dans un autre tableau, les annotations fonctionnelles effectuées (en termes de changements de topique) et les variations de tempo intra-locuteurs obtenues. La table de données regroupe plus précisément, d'un côté les différentes valeurs obtenues de débit d'élocution, de différence de débit et de distance entre les noeuds feuilles, et de l'autre côté les annotations fonctionnelles effectuées. Une autre table regroupe aussi d'une part, la durée des pauses et leur écart par rapport à la moyenne et, d'autre part, les annotations fonctionnelles effectuées.

Nous nous intéressons à présent aux variations de tempo et aux diverses fonctions qu'elles revêtent. Dans un premier temps, nous présenterons une analyse des différences de tempo en fonction du sexe du locuteur et du style de parole qu'il pratique et, dans un second temps, une étude des variations de tempo intra-locuteurs comme indices de changement de topique.

4 Tempo et fonctions extra-linguistiques : Une comparaison en fonction du sexe et du type de production

4.1 Rappel : les fonctions extra-linguistiques du tempo

Nous avons mentionné au cours de notre introduction que les variations de tempo, i.e. les variations du débit d'élocution, de la durée et du nombre de pauses, caractérisent à la fois l'individualité du locuteur et le style discursif qu'il pratique. En effet, certaines études ont pu révéler, par exemple, des différences de tempo entre hommes et femmes (Byrd, 1992; Verhoeven et al., 2004; Jacewicz et al., 2009). Cette différence résulterait, selon eux, de la vitesse d'élocution et non du nombre de pauses. D'autres études, en revanche, infirment ces résultats et montrent au contraire que les femmes ont un débit plus rapide que celui des hommes, une différence ici attribuée à une réduction du temps de pause, plus élevée chez les sujets femmes (Saint-Bonnet & Boe, 1977). D'autres encore ne révèlent aucun effet du sexe du locuteur sur le débit d'élocution. Il est donc difficile, au vu de la littérature, de certifier le lien entre les variations de tempo et le sexe du locuteur.

Par ailleurs, les variations de tempo dépendraient du style discursif que le locuteur pratique. Une parole conversationnelle serait ainsi marquée par un débit d'élocution plus lent, un plus grand nombre de pauses et une plus grande durée des pauses qu'une lecture oralisée (Goldman-Eisler, 1968; Grosjean & Deschamps, 1972b; Silverman et al., 1992; Ayers, 1994; Hirschberg, 2000). En outre, la description oralisée d'images se distinguerait des interviews radiophoniques par des pauses plus courtes et moins nombreuses (Grosjean & Deschamps, 1972b). Mais on peut se demander ici si le fait que ces différents types de production sont marqués par des pauses plus ou moins longues, plus ou moins fréquentes, des débits plus ou mois lents, ne dépend pas finalement d'autres facteurs. En effet, Grosjean et Deschamps (1972b) suggèrent que, plus que le style de production, les variations de débit reflètent le niveau cognitif de l'opération verbale requis pour chacun de ces modes. Il serait donc possible que lorsqu'on étudie les variations de tempo en fonction du type de production, on en revient en fait à étudier l'influence d'autres facteurs sur les variations de tempo. En effet, la parole spontanée, par exemple, serait marquée par des pauses plus longues du fait du niveau cognitif qu'elle implique ; elle serait aussi marquée par un débit d'élocution plus variable qui révèle, non pas du caractère spontané de cette parole, mais en réalité des intentions du locuteur, certainement plus présentes en parole authentique.

Nous proposons ici, à partir de notre base de données, une étude succincte des variations de tempos en fonction des facteurs sexe et style de parole.

4.2 Base de données

Pour cette étude, nous utilisons uniquement les enregistrements des corpus Aix-Marsec et CID, pour lesquels une transcription phonémique est disponible [98]. La durée des enregistrements pour chaque locuteur est d'environ 1 minute pour les locuteurs anglais, 3 minutes pour les locuteurs français. Parmi les locuteurs anglais, nous comptons 16 femmes et 33 hommes, parmi les locuteurs français, 3 femmes et 2 hommes. L'analyse en fonction des styles de parole est effectuée sur le corpus Aix-Marsec. Le tableau 18 indique le nombre d'enregistrements par catégorie :

Catégories	Nb Enreg.
Commentaires	9
Bulletins d'information	7
Paroles publiques	4
Emissions religieuses	2
Reportages	12
Fictions	5
Dialogues	4
Propagande	2
Divers	3

TABLE 18 – Nombre d'enregistrements par catégorie, i.e. style de parole, sélectionnés - corpus Aix-Marsec.

Les données de sortie de l'Etape 3 d'ADoTeVA (analyse inter-locuteurs) sont utilisées pour nos analyses [99].

4.3 Analyses statistiques

4.3.1 Interdépendance des composantes ?

D'abord, nous analysons la relation entre les composantes du tempo : le débit d'élocution, la durée des pauses et le nombre de pauses. Au vu du graphique 71, il est difficile d'établir une relation linéaire entre le débit et la durée des pauses. Le modèle de régression linéaire donne une probabilité significative pour la pente ($F(1,51)=9.807$, p-val=0.002), ce qui indiquerait une relation linéaire significative entre le débit d'élocution et la durée des pauses. Plus le

98. Les TextGrids utilisés pour cette étude sont sur CD ROM - dossier DATA_CHAP4
99. Ces données sont sur CD ROM, ANNEXES_CHAP4 : Table1.

débit est lent, plus les pauses seraient longues. En revanche, le coefficient de corrélation n'est pas très élevé (R^2=0.1448), ce qui nous laisse à penser que si la relation est linéaire entre les variables débit et durée des pauses, elle n'est que très faible. En effet, seulement 14% de la variabilité des durées des pauses est expliquée par le débit d'élocution.

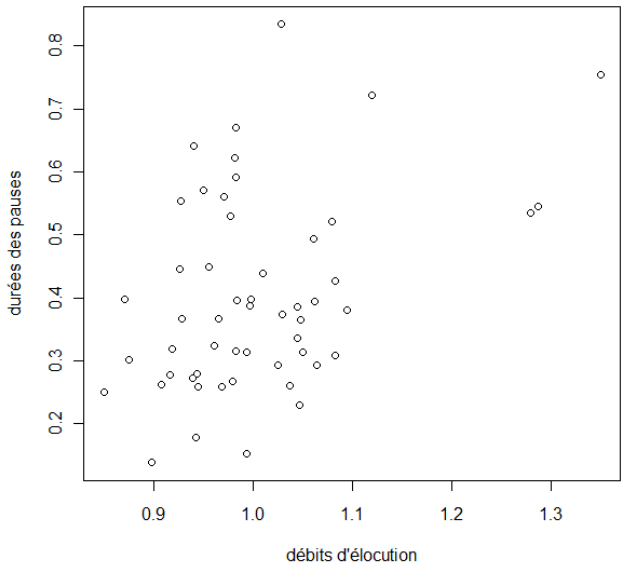


FIGURE 71 – Représentation graphique des débits d'élocution en fonction des durées des pauses.

La relation entre le nombre de pauses et la durée des pauses semble davantage linéaire au vu du graphique 72. Le modèle linéaire donne une probabilité significative pour la pente (F(1,51)=16.32, p-val=0.0001). Plus le nombre de pauses augmente, plus les pauses seraient courtes. En revanche, le coefficient de corrélation est ici aussi peu élevé ((R^2=0.1448), ce qui indiquerait une faible corrélation des variables 'durée des pauses' et 'nombres de pauses'.

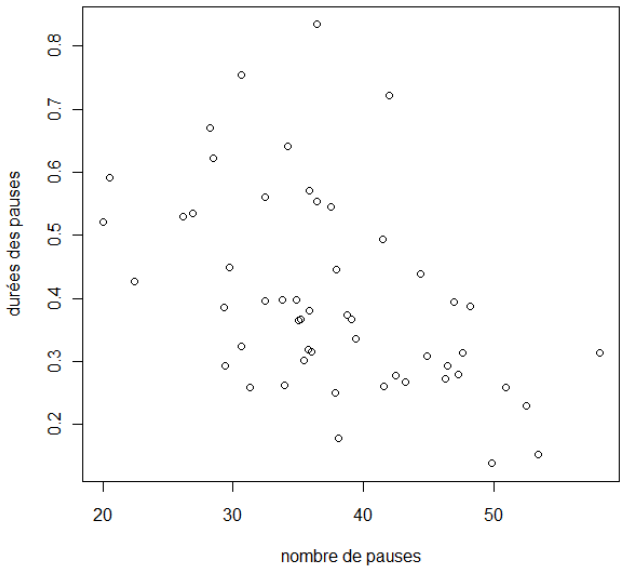


FIGURE 72 – Représentation graphique du nombre de pauses en fonction des durées des pauses.

Au vu du graphique 73, il est clair que le débit d'élocution et le nombre de pauses ne sont pas corrélés. La probabilité associée est supérieure à 5% ($F(1,51)=0.858$, p-val=0.358), ce qui indique qu'il n'existe pas de relation linéaire entre ces deux variables.

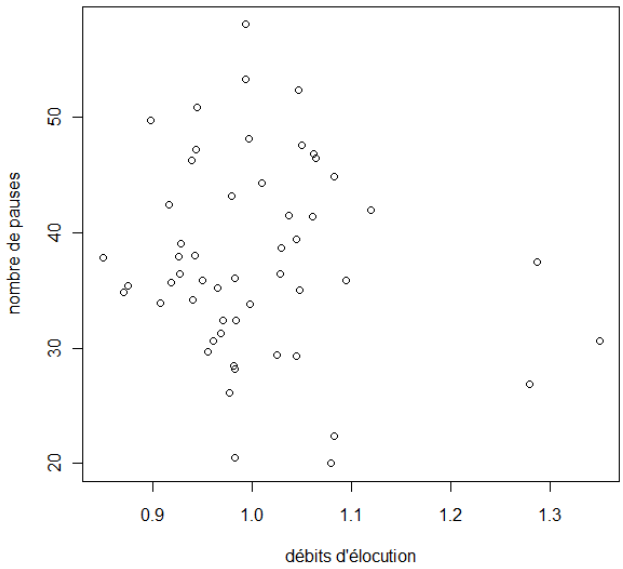


FIGURE 73 – Représentation graphique des débits d'élocution en fonction du nombre de pauses.

Au vu de ces résultats, la tendance serait que plus le débit est lent, plus les durées des pauses sont longues et moins le nombre de pauses est élevé. Si les données suivent une tendance, on ne peut cependant pas affirmer que les trois composantes du tempo sont fortement corrélées. On peut donc en déduire que ces composantes ne sont pas interdépendantes et que l'une peut varier sans que les autres varient à leur tour.

4.3.2 Tempo et genre du locuteur

Nous cherchons à présent à analyser l'effet du sexe du locuteur (SEXE) sur les différentes composantes du tempo. Au vu des graphiques 74, 75 et 76, il est difficile d'établir visuellement un effet du sexe du locuteur sur le débit d'élocution, sur la durée des pauses ou encore sur le nombre de pauses. Nous observons en effet un chevauchement important des boîtes à moustaches obtenues pour les populations hommes et femmes. Les modèles linéaires montrent que l'effet du sexe du locuteur n'est pas significatif, et ce, sur aucune des composantes du tempo (durée des pauses : $F(1,51)=1.217$, p-val=0.275 ; débit : $F(1,51)=2.693$, p-val=0.107 ; nombre de pauses : $F(1,51)=1.905$, p-val=0.173).

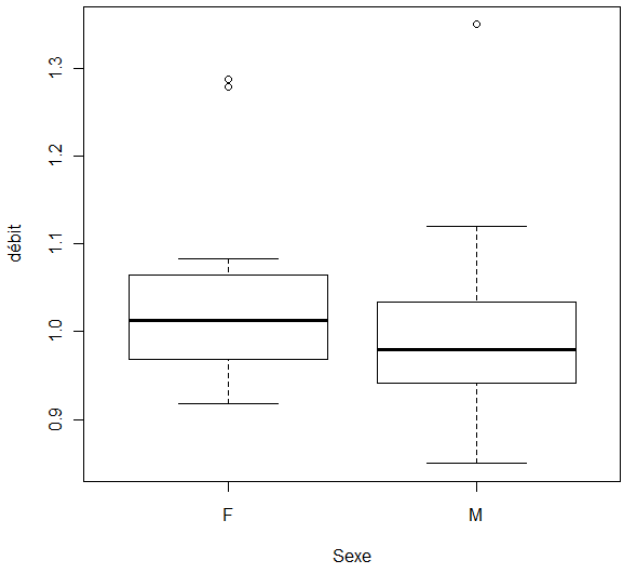


FIGURE 74 – Boîtes à moustaches des débits d'élocution selon les modalités de la variable SEXE. F représente les femmes, M les hommes.

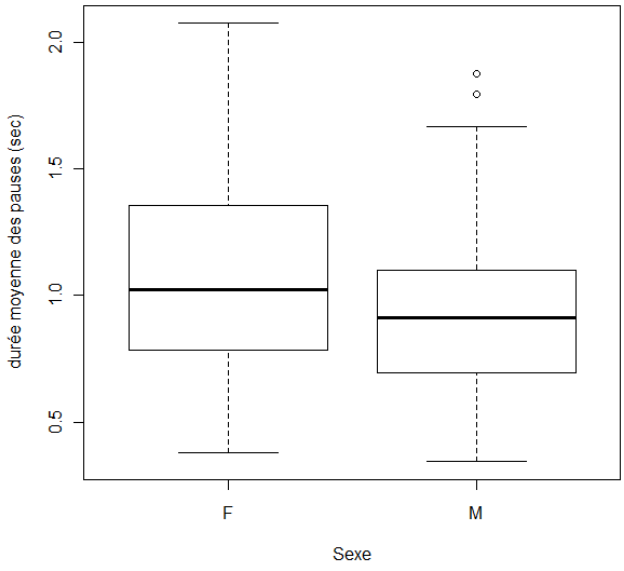


FIGURE 75 – Boîtes à moustaches des durées des pauses selon les modalités de la variable SEXE. F représente les femmes, M les hommes.

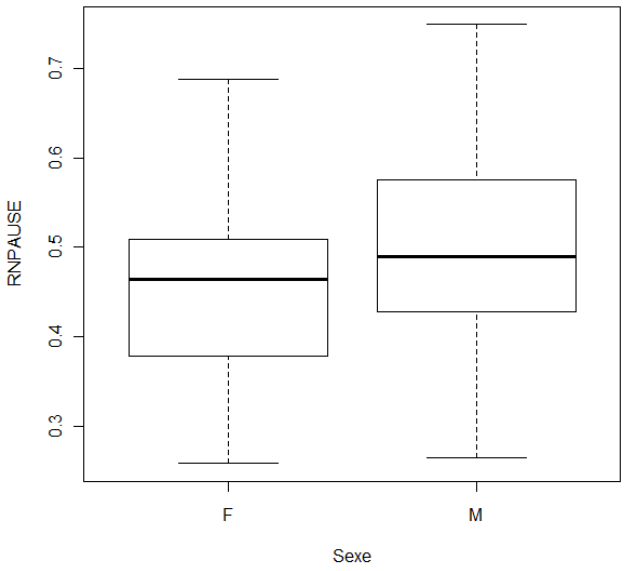


FIGURE 76 – Boîtes à moustaches du nombre de pauses selon les modalités de la variable SEXE. F représente les femmes, M les hommes.

Par ailleurs, si nous nous intéressons aux moyennes des débits d'élocution, des durées des pauses et du nombre de pauses obtenues pour les hommes d'un côté et pour les femmes de l'autre, nous voyons que ces moyennes sont similaires. Le débit moyen des femmes est de 1.03, celui des hommes de 0.99 ; la durée moyenne des pauses pour les femmes est de 435ms, pour les hommes de 385ms ; le nombre de pauses est de 35 pour les femmes et 38 pour les hommes. En revanche, on note une grande variabilité au sein de chaque groupe. L'écart type des débits d'élocution pour les femmes est de 0.103, pour les hommes de 0.089 ; l'écart type des durées moyennes des pauses est de 0.164 pour les femmes et de 0.150 pour les hommes ; l'écart type du nombre de pauses est de 8.41 pour les femmes, 8.3 pour les hommes. Ou encore, le débit des locuteurs varie entre 0.84 (le débit le plus rapide) et 1.34 (le débit le plus lent), la durée des pauses varie entre 138ms et 834ms et le nombre de pauses varie entre 20 et 58 pauses. Les locuteurs de cette base de données ont donc bien des débits d'élocution différents (i.e. intrinsèques) mais cette différence ne résulte pas du facteur sexe. Ces différences sont d'ailleurs plus marquées au niveau des pauses qu'au niveau du débit d'élocution, où la variance des durées moyennes des pauses est de 0.38, celle du nombre de pauses de 0.22 et celle du débit de 0.09.

4.3.3 Tempo et type de production

Nous observons à présent l'effet du type de production (TYPE) sur le débit d'élocution, la durée des pauses et le nombre de pauses. Il est à noter que nous nous en tenons, pour cette étude, à une analyse graphique des données, vu que certaines catégories (e.g. *émissions religieuses*, *propagandes*) ne sont pas représentatives d'un nombre assez important de locuteurs. Par ailleurs, pour une analyse fine du tempo en fonction des styles de parole, il est, à notre sens, plus juste d'étudier les mêmes locuteurs pratiquant divers styles de parole, afin que l'effet intrinsèque du tempo, i.e. le tempo propre au locuteur, n'interfère pas dans les résultats obtenus. Nous pensons cependant, à partir de boîtes à moustaches, voir se dégager certaines tendances, qui, bien-entendu, mériteraient d'être vérifiées par des analyses statistiques.

Au vu du graphique 77, seule la catégorie *émissions religieuses* semble être marquée par un débit différent, i.e. plus lent. Les débits de la catégorie *informations* semblent aussi plus rapides que ceux des catégories *dialogues*, *fictions*, *paroles publiques* et *propagandes*. En revanche, les médianes des boîtes à moustaches pour les catégories *dialogues*, *divers* et *fictions* ou encore pour les catégories *paroles pupliques*, *propagandes* et *reportages* sont proches et leurs quartiles se chevauchent, ce qui laisse à penser que les débits de ces catégories sont similaires.

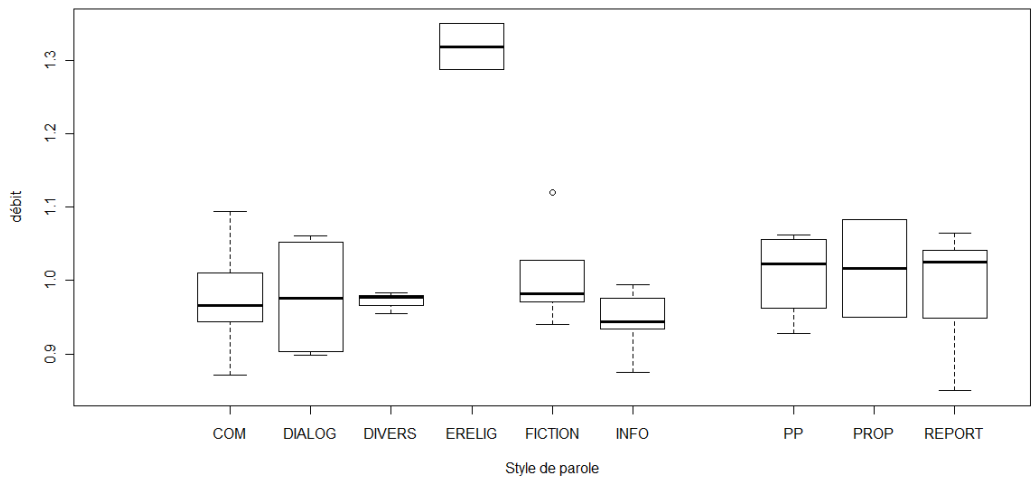


FIGURE 77 – Boîtes à moustaches des débits de parole selon les modalités de la variable TYPE. COM renvoie à la catégorie *commentaires*, DIALOG à la catégorie *dialogues*, DIVERS à la catégorie *divers*, ERELIG à la catégorie *émissions religieuses*, FICTION à la catégorie *fictions*, INFO à la catégorie *informations*, PP à la catégorie *paroles publiques*, PROP à la catégorie *propagandes* et REPORT à la catégorie *reportages*.

Le graphique 78 représente les boîtes à moustaches des durées des pauses selon les modalités de variable TYPE. Au vu du graphique, la catégorie *informations* semble marquée par une grande variabilité. Elle se différencie des catégories *divers*, *émissions religieuses* et *fiction*, marquées, elles, par des pauses plus longues. Les durées des pauses des catégories *commentaires* et *dialogues* semblent similaires. Elles paraissent en revanche plus courtes que celles des catégories *divers*, *émissions religieuses*, *fictions*, *propagandes* et *paroles publiques*. Les catégories *émissions religieuses* et *fictions* semblent marquées par des pauses de mêmes durées, ainsi que les catégories *paroles publiques* et *propagandes*. De plus, les durées des pauses de la catégorie *divers* semblent plus courtes que celles des catégories *émissions religieuses* et *fictions*. Les catégories *divers*, *émissions religieuses* et *fiction* seraient marquées par des pauses plus longues que les catégories *paroles publiques*, *propagandes* et *reportages*. Enfin les pauses de la catégorie *paroles publiques* paraissent plus longues que les pauses de la catégorie *reportages*.

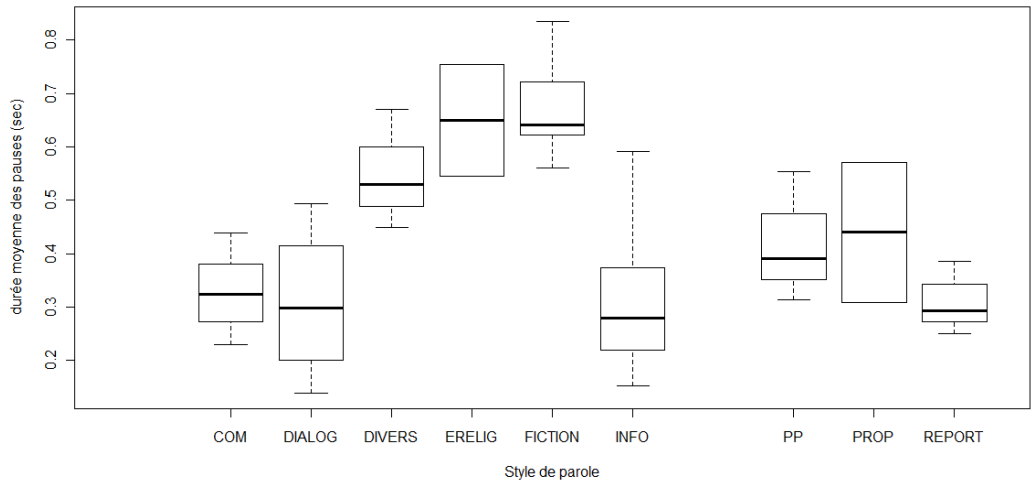


FIGURE 78 – Boîtes à moustaches des durées des pauses selon les modalités de la variable TYPE. COM renvoie à la catégorie *commentaires*, DIALOG à la catégorie *dialogues*, DIVERS à la catégorie *divers*, ERELIG à la catégorie *émissions religieuses*, FICTION à la catégorie *fictions*, INFO à la catégorie *informations*, PP à la catégorie *paroles publiques*, PROP à la catégorie *propagandes* et REPORT à la catégorie *reportages*.

Le graphique 79 représente les boîtes à moustaches du nombre de pauses selon les modalités de variable TYPE. Au vu du graphique, les catégories *commentaires*, *dialogues*, *paroles publiques*, *propagandes* et *reportages* seraient marquées par un nombre de pauses semblable. Les catégories *émissions religieuses* et *fictions* compteraient aussi un même nombre de pauses. En revanche, le nombre de pauses de la catégorie *dialogues* est plus élevé que celui des ca-

tégories *fictions* et *émissions religieuses*. Les *émissions religieuses*, quant à elles, comptent moins de pauses que les catégories *paroles publiques* et *propagandes*. La catégorie *informations* est à nouveau marquée par une grande variabilité. Enfin, la catégorie *divers* se détache de l'ensemble des catégories avec un nombre de pauses assez réduit.

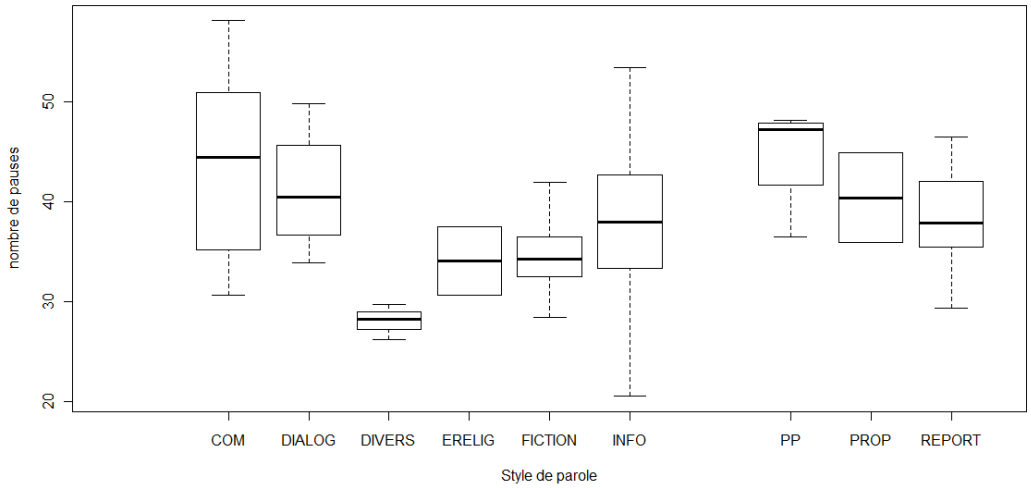


FIGURE 79 – Boîtes à moustaches du nombre de pauses selon les modalités de la variable TYPE. COM renvoie à la catégorie *commentaires*, DIALOG à la catégorie *dialogues*, DIVERS à la catégorie *divers*, ERELIG à la catégorie *émissions religieuses*, FICTION à la catégorie *fictions*, INFO à la catégorie *informations*, PP à la catégorie *paroles publiques*, PROP à la catégorie *propagande* et REPORT à la catégorie *reportages*.

Au vu de ces résultats, nous proposons une représentation graphique des débits, des durées des pauses et du nombre de pauses en fonction des types de production, sur une échelle graduée (figure 80). Les *émissions religieuses* et les *fictions* seraient marquées par des débits plus lents, des pauses plus longues et moins nombreuses ; les *reportages*, les *dialogues*, les *commentaires* et les *informations* seraient marqués par des débits plus rapides, des pauses plus courtes et plus nombreuses (excepté pour la catégorie *reportages*, pour laquelle le nombre de pauses est plutôt réduit) ; au centre de cette gradation des variations de tempo, se trouveraient les catégories *propagandes* et *paroles publiques*. Il est à noter toutefois que le nombre de pauses pour la catégorie *propagandes* est supérieur à celui de la catégorie *paroles publiques* et le plus élevé de tous.

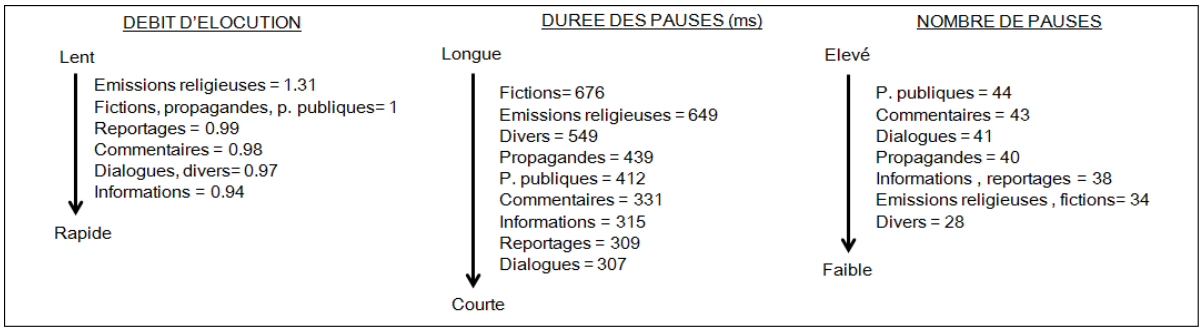


FIGURE 80 – Représentation graphique des débits, des durées des pauses et du nombre de pauses en fonction des types de production.

4.4 Discussion

Nous nous sommes intéressée dans cette étude aux fonctions extra-linguistiques que revêtent les variations de tempo, notamment la façon dont elles renseignent sur le sexe du locuteur et sur le style de parole qu'il pratique. Nous avons pu montrer tout d'abord qu'il n'existait pas de différence significative de tempo entre hommes et femmes dans notre corpus. Ces résultats confirment donc ceux de Lass et Sandusky (1971), Malecot et al. (1972), Tsao et Weismer (1997), Binnenpoorte et al. (2005) et Robb et al. (2004). Contrairement à l'étude de Saint-Bonnet et Boe (1977), les durées des pauses des femmes ne sont pas plus courtes que celles des hommes. On doit cependant noter que cette étude se base sur 10 locuteurs et que les résultats obtenus reflètent peut être tout simplement les tempos intrinsèques des locuteurs, qui se trouvent, dans cette étude, être plus rapides pour les femmes que pour les hommes. Les auteurs avaient d'ailleurs souligné que cette « différence [ne pouvait pas être] très significative ici, compte tenu du petit nombre de locuteurs », mais avaient finalement conclu qu'elle pouvait l'être du fait « [qu'elle avait] été relevée comme telle sur de plus vastes échantillons ». Dans notre étude, en tous cas, malgré un nombre de locuteurs assez élevé, aucune différence de tempo n'est relevée. Par ailleurs, nos résultats contredisent ceux de (Verhoeven et al., 2004) qui montrent une différence significative des débit d'élocution en fonction du sexe du locuteur. Dans cette étude, les hommes ont en effet un débit significativement plus rapide que celui des femmes ($p<0.0001$). Au vu du nombre de locuteurs de cette étude (i.e. 160 locuteurs : 80 femmes et 80 hommes), il est difficile ici d'attribuer ces résultats à un nombre trop réduit de données. En revanche, ce qui différencie cette étude de la nôtre est la mesure utilisée. En effet, les auteurs mesurent le débit d'élocution en termes de syllabes par seconde alors que nous le mesurons comme le rapport entre la somme des valeurs observées et la somme des valeurs prédites. On peut donc se demander si les résultats obtenus ne sont pas le reflet des mesures utilisées. Au vu de l'étude de (Tsao & Weismer, 1997), il semblerait que non, puisque les

auteurs mesurent le débit en termes de syllabes par minute et ne trouvent aucune différence significative entre les débits d'élocution des hommes et des femmes. Si des différences ont pu être trouvées entre hommes et femmes, il semblerait qu'elles résultent plutôt des tempos intrinsèques des locuteurs.

Si aucune différence n'est rapportée dans notre étude entre hommes et femmes, il semblerait en revanche, au vu des premières analyses, que les différences de tempo permettent de distinguer des styles de parole mais aussi permettent le regroupement de types de production similaires. Par exemple, nous avons pu observer que les fictions et les émissions religieuses se distinguent des autres catégories par des débits d'élocution plus lents et des durées de pause plus lentes. Ou encore, les reportages, les commentaires et les informations se différencient des autres catégories par des débits d'élocution plus rapides et des pauses plus courtes. Il est intéressant de noter que ces trois catégories qui peuvent être regroupées sous une même catégorie, e.g. « rapport d'informations » sont marquées par de mêmes tempos. De la même manière, les paroles publiques et les propagandes sont marquées par des débits et des durées de pauses similaires. Les variations de tempo, notamment la combinaison des variations de débit d'élocution et des durées des pauses, semblent donc permettre la distinction des styles de parole et leur regroupement. Ici, par exemple, il semble que 3 groupes de catégories se distinguent : (1) les fictions et les émissions religieuses, (2) les paroles publiques et les propagandes, (3) les dialogues, les reportages, les commentaires et les informations.

Cette étude révèle aussi que les composantes du tempo ne sont pas ou peu interdépendantes. Il n y a que peu de corrélation entre débit d'élocution et durée des pauses, entre durée des pauses et nombre de pauses, et aucune corrélation entre débit d'élocution et nombre de pauses. Cela suggère que la distinction des types de production peut se faire par le débit d'élocution, par la durée des pauses, par le nombre de pauses, par la combinaison du débit et de la durée des pauses ou encore par la combinaison de la durée des pauses et du nombre de pauses.

Il est toutefois à noter que ces observations nécessitent d'être complétées par des analyses statistiques et qu'elles ne peuvent être généralisables, du fait que les données ne sont pas équilibrées. Par ailleurs, une étude des tempos de locuteurs pratiquant divers styles de parole serait peut être plus juste dans le sens où le débit intrinsèque du locuteur n'interfèrerait pas dans les analyses.

Nous portons à présent notre intérêt sur la façon dont les variations de tempo peuvent indiquer des changements de topique.

5 Tempo et fonctions linguistiques : détection automatique et prédiction des changements de topique

5.1 Rappel : variations de tempo et topicalisation

Nous avons relaté de la littérature que les variations de tempo, ici le débit d'élocution et la durée des pauses participeraient à la signalisation du discours et indiqueraient des changements ou des continuations de topicalisation. En anglais et en français, des pauses plus longues précèderaient les segments initiaux de discours introduisant un nouveau topique (Lehiste, n.d.; G. Brown et al., 1980; Silverman, 1987; Swerts & Geluykens, 1993; J. Pijper & Sanderman, 1994; Fon, 2002; C. Smith, 2005; Ouden et al., 2009). De plus, le débit d'élocution serait plus lent sur les segments initiaux traitant d'un nouveau topique et sur les segments finaux clôturant un topique, que sur le reste des segments de l'énoncé (Brubaker, 1972; Butterworth, 1975; Hirose & Kawanami, 1998; C. Smith, 2005).

5.2 Annotation fonctionnelle et base de données

Afin d'observer les variations de tempo en fonction des changements de topique, nous nous sommes servie de l'annotation fonctionnelle qui les concerne, effectuée à partir d'une annotation en groupes clitiques. Nous ne revenons pas sur le détail de ces annotations. Nous renvoyons le lecteur aux sections 2 et 5.2 du chapitre 3. Pour rappel cependant, nous examinons la structure hiérarchique du discours en termes de structure intentionnelle, pour laquelle nous reconnaissons 3 niveaux d'intentions de discours (*discourse segment purpose*) ou DSP : l'étiquette DSP0 entre deux unités signifie qu'elles partagent les mêmes intentions de communication, i.e. un même topique. L'étiquette DSP1 se trouve à un niveau supérieur de la structure hiérarchique intentionnelle par rapport à DSP0. Elle est positionnée entre deux unités si ces dernières partagent un même topique et si une information nouvelle est apportée par la seconde unité. Enfin, l'étiquette DSP2 se trouve au sommet de la hiérarchie intentionnelle. Elle signifie que deux unités consécutives ne partagent pas un même topique.

La base de données à partir de laquelle nous menons notre analyse est extraite des corpus Aix-Marsec (AM), pour l'étude des changements de topique en anglais, et CID pour celle des changements de topique en français. Nous avons sélectionné les dialogues de 9 locuteurs du corpus AM (2 femmes et 7 hommes) et de 5 locuteurs du CID (3 femmes et 2 hommes). Le nombre de données est ici moins important car l'étude nécessite pour chaque enregistrement une annotation en groupes clitiques, une annotation fonctionnelle (i.e. DSP0, DSP1 et DSP2) et une annotation en phonèmes. La durée totale des enregistrements est de 30 minutes environ.

Les mesures des variations de débit et des durée de pauses sont obtenues par l'algorithme ADoTeVA. Nous rappelons que l'algorithme permet de corréler des annotations fonctionnelles à la détection automatique des variations de tempo. Suite à l'étape 7, nous obtenons deux tables de sortie (e.g. 19), à partir desquelles nous cherchons à observer d'éventuels liens entre, d'un côté, l'annotation en DSP (i.e. intentions de discours) et les variations du débit d'élocution et de l'autre côté, entre l'annotation en DSP et les durées des pauses.

FILENAME	UNITS	START	END	RATE	DRATE	LDSP	RDSP	LDIST	RDIST
J01SP1M	Paddy'not	0.105	0.445	0.934	NA	DSP2	DSP0	114.208	0.025
J01SP1M	long'/after	0.445	1.005	0.959	0.025	DSP0	DSP0	0.025	1.486
J01SP1M	that|	1.005	1.285	2.171	1.212	DSP0	DSP1	1.486	0.209
J01SP1M	the'tears	1.285	2.005	1.962	0.209	DSP1	DSP0	0.209	0.472
J01SP1M	of grief|	2.005	2.745	1.594	0.368	DSP0	DSP0	0.472	2.519
J01SP1M	andfrust|	2.745	3.625	0.93	0.664	DSP0	DSP0	2.519	0.156
J01SP1M	turned	3.684	3.884	1.086	0.156	DSP0	DSP0	0.156	0.357

TABLE 19 – Table de données de sortie suite à l'étape 7 du plugin ADoTeVA. FILENAME indique le locuteur, UNITS, l'unité en question, START le début de l'unité, END la fin de l'unité, RATE le débit d'élocution sur l'unité, DRATE la différence de débit entre une unité et celle qui la précède, LDSP et RDSP les DSP aux frontières gauche et droite de l'unité, et LDIST et RDIST les distances gauche et droite entre les noeuds feuilles de la structure arborescente.

5.3 Analyses statistiques

100

Afin de modéliser la relation entre l'annotation fonctionnelle et les valeurs de débits d'élocution et de durée des pauses, nous avons mené plusieurs analyses de variance. Nous avons en effet étudié l'effet du facteur DSP (intentions du discours) sur les variables quantitatives RATE, DIFFRATE et LEFTDIST, où RATE est le débit d'élocution, DIFFRATE, la différence de débit entre deux unités consécutives et LEFTDIST, la distance à gauche de l'unité. Nous avons aussi étudié l'effet du facteur DSP sur la durée des pauses. Nous présentons séparément les résultats obtenus pour les corpus AM et CID.

100. Les données à partir desquelles ont été effectuées les analyses statistiques sont sur CD ROM - ANNEXES_CHAP4 : Table 2 et Table 3.

5.3.1 AM : Résultats

Au vu des boîtes à moustaches représentées dans le graphique 81, il semble qu'il existe un effet DSP (intentions de discours) sur le débit des locuteurs. Les tests de significativité du facteur corroborent les impressions graphiques (débit : $F(2, 3242)=28.81$, p-val=3.977e-13 ; différence de débit : $F(2,3233)<2.2e-16$). En revanche, l'effet du facteur est significatif seulement pour le niveau DSP2 lorsque l'on regarde son effet sur les distances gauches des noeuds feuilles (LEFTDIST : DSP1 : p-val=0.895 vs. DSP2 : p-val = 2.69e-10). En anglais, la structure hiérarchique du discours exprimée en termes de structure intentionnelle est donc marquée par des variations du débit d'élocution. Plus l'intention du discours (DSP) est haute dans la structure hiérarchique, plus l'unité a un débit lent et plus la différence et la distance entre l'unité en question et celle qui la précède sont importantes.

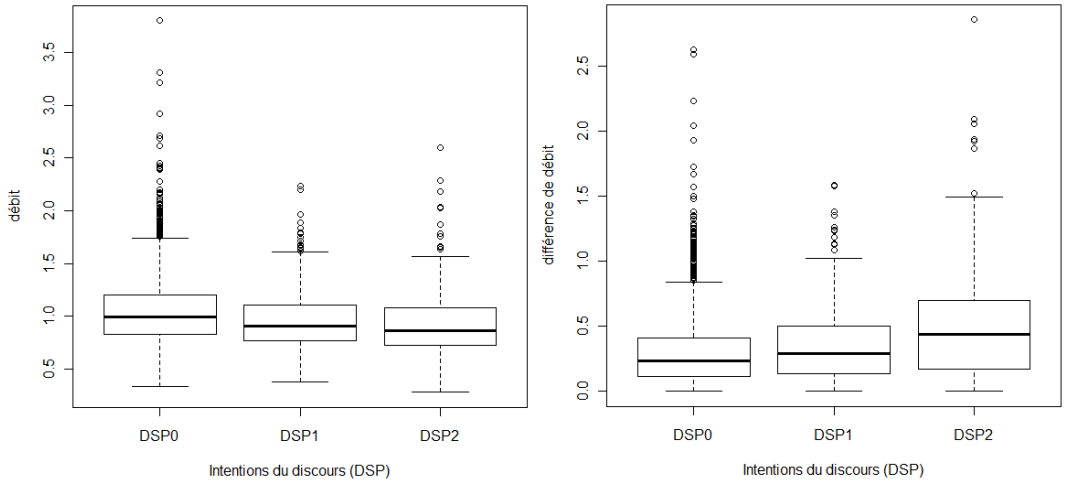


FIGURE 81 – Sur la gauche, boîtes à moustaches du débit d'élocution en fonction de l'intention de discours (DSP) ; sur la droite, boîtes à moustaches de la différence de débit en fonction de DSP. DSP0 indique une continuité de topique, DSP1, un ajout d'information, DSP2 un changement de topique.

Nous nous intéressons à présent à l'effet DSP sur la durée des pauses. Au vu du graphique 82, l'effet DSP semble important. La boîte à moustaches obtenue pour les DSP0 est écrasée sur la médiane à une valeur à 0, ce qui indique que la plupart des DSP0 ne sont pas précédés de pause. La médiane des DSP1 est estimée à 0.3 et celle des DSP2 à 0.4. Il est à noter cependant un certain nombre de « valeurs aberrantes » pour les DSP0. Cela montre que certains DSP0 sont précédés de pauses, de durée sensiblement égale à celles qui précèdent les DSP1 ou les DSP2. Il est intéressant par conséquent de regarder l'histogramme des DSP0 (figure 83). Nous pouvons constater que le nombre d'individus supérieurs à 0 est assez réduit. On peut donc s'attendre

à un effet des intentions de discours (DSP) sur la durée des pauses. L'analyse de variance corrobore l'impression graphique, l'effet du facteur DSP est très significatif (F(2,4282)=1178, p-val<2e-16). On peut donc conclure que plus l'intention du discours (DSP) est haute dans la structure hiérarchique (e.g. DSP2), plus la pause qui la précède est longue.

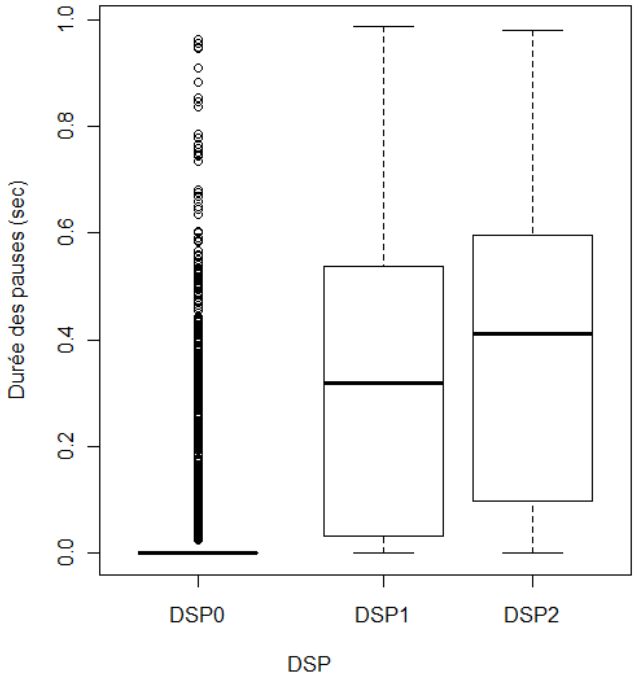


FIGURE 82 – Boîtes à moustaches des durées des pauses en fonction de l'intention de discours (DSP) ; DSP0 indique une continuité de topique, DSP1, un ajout d'information, DSP2 un changement de topique.

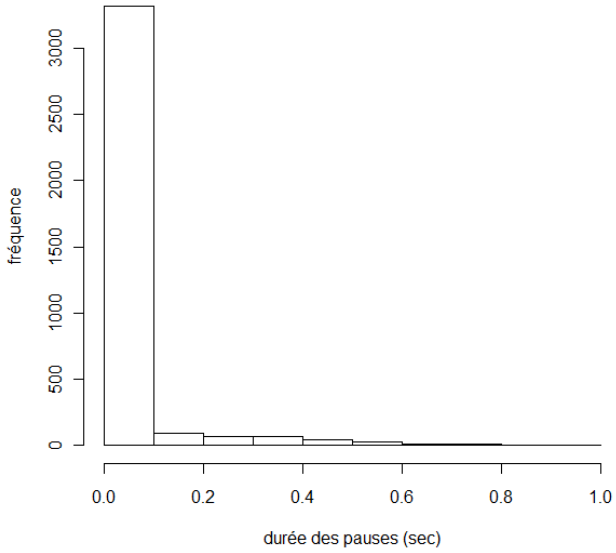


FIGURE 83 – Histogramme des durées des pauses pour le niveau DSP0. Ce niveau indique une continuité de topique.

5.3.2 CID : Résultats

Le graphique 84 représente les débits (fenêtre gauche) et les différences de débit (fenêtre droite) en fonction des intentions du discours (DSP). Au vu du graphique, il ne semble pas y avoir d'effet DSP sur le débit de l'unité (fenêtre gauche). En effet, le niveau DSP1 est peu significatif (p-val=0.0105) et le niveau DSP2 ne l'est pas du tout (p-val=0.0523). La fenêtre droite du graphique suggère un effet DSP sur les différences de débit. L'analyse de variance montre que seul le niveau DSP2 est significatif aussi bien lorsque l'on regarde l'effet DSP sur les différences de débit et les distances gauches des noeuds feuilles (différence : DSP2, p-val=7.75e-07 vs. DSP1, p-val=0.339 ; distance : DSP2, p-val<2.2e-16 vs. DSP1, p-val= 0.684). En français, la structure hiérarchique du discours est donc marquée par des variations de débit uniquement à de hauts niveaux de la hiérarchie, i.e. pour des changements de topique. Les unités qui introduisent un nouveau topique ne sont pas toujours marquées par un débit plus lent ou plus rapide que leur précédente, mais par une différence importante de débit.

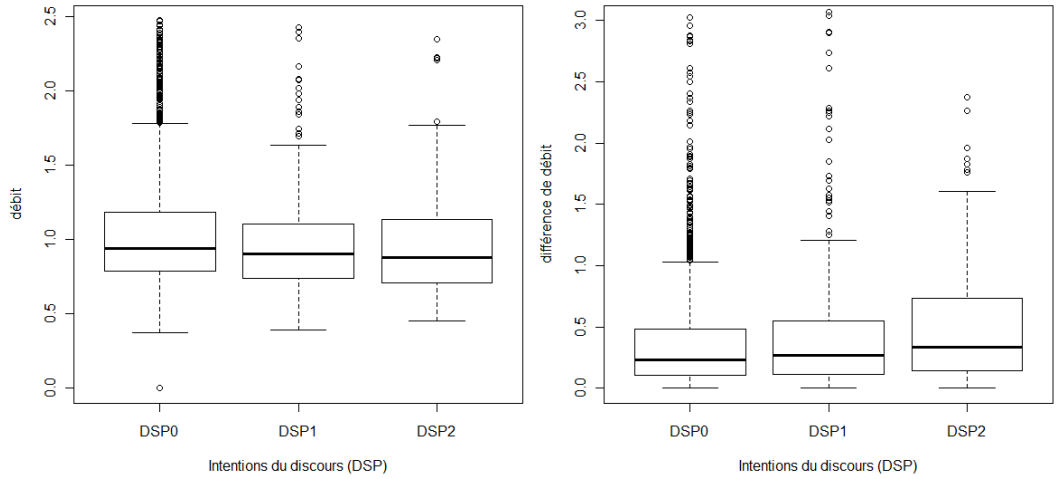


FIGURE 84 – Sur la gauche, boîtes à moustaches du débit d'élocution en fonction de l'intention de discours (DSP) ; sur la droite, boîtes à moustaches de la différence de débit en fonction de DSP. DSP0 indique une continuité de topique, DSP1, un ajout d'information, DSP2 un changement de topique.

Nous nous intéressons à présent à l'effet DSP sur la durée des pauses. Le graphique 85 montre aussi pour ce corpus un effet DSP important. La boîte à moustaches obtenue pour les DSP0 est écrasée sur la médiane à une valeur à 0, ce qui indique que la plupart des DSP0 ne sont pas précédés de pause. La médiane des DSP1 est aussi écrasée à zéro mais ce niveau montre davantage de variance. Enfin, celle des DSP2 est à 0.6. Ici aussi, on observe un certain nombre de « valeurs aberrantes » pour le niveau DSP0, qui indiquent que certains DSP0 sont précédés de pauses. Si on regarde l'histogramme des DSP0 (figure 86), nous constatons que le nombre d'individus supérieurs à 0 est assez réduit. Nous observons également une forte variance au niveau des DSP1. Si on observe les fréquences des durées de pause qui précèdent un DSP1, nous constatons que les DSP1 peuvent ne pas être précédés de pause ou au contraire précédés de pauses assez longues. On peut cependant penser, malgré cette variance, à un effet des intentions de discours (DSP) sur la durée des pauses. L'analyse de variance montre un effet très significatif du facteur DSP ($F(2,3109)=457$, p-val$<$2e-16). En français aussi, plus l'intention du discours (DSP) est haute dans la structure hiérarchique, plus la pause qui la précède est longue.

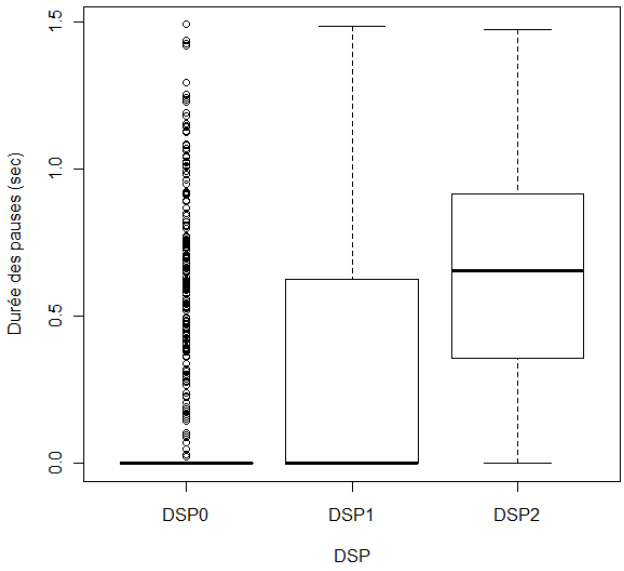


FIGURE 85 – Boîtes à moustaches des durées des pauses en fonction de l'intention de discours (DSP) ; DSP0 indique une continuité de topique, DSP1, un ajout d'information, DSP2 un changement de topique.

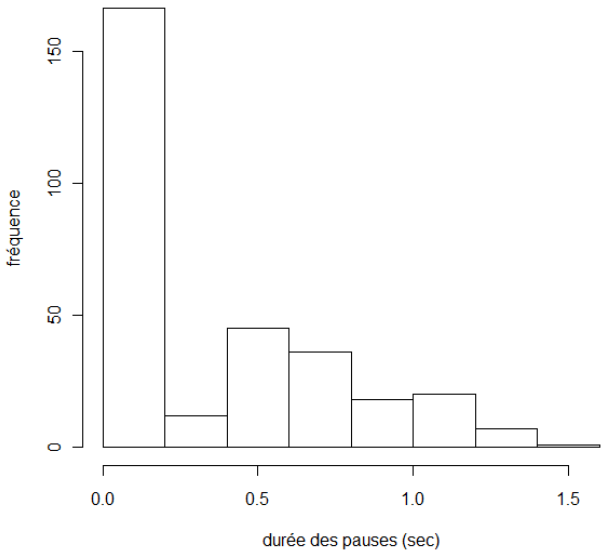


FIGURE 86 – Histogramme des durées des pauses pour le niveau DSP0. Ce niveau indique une continuité de topique.

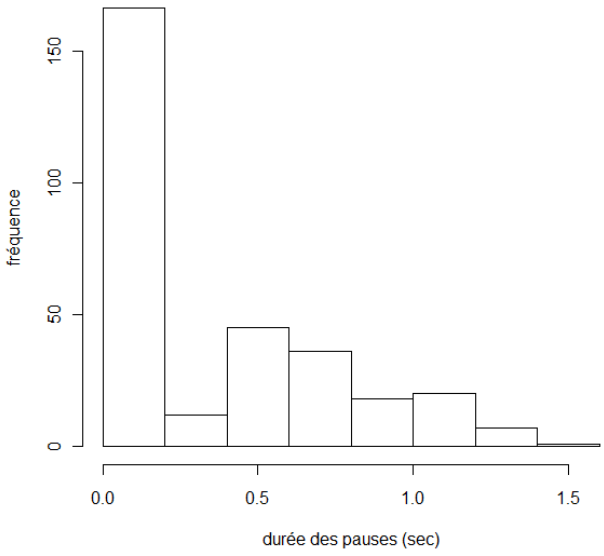


FIGURE 87 – Histogramme des durées des pauses pour le niveau DSP1. Ce niveau indique un ajout d'information.

5.4 Discussion

Au vu des ces analyses, nous pouvons conclure que, en anglais et en français, la structure hiérarchique du discours, exprimée en termes de structure intentionnelle, est indiquée par des variations du débit d'élocution et de la durée des pauses. En anglais, plus l'intention de discours (DSP) est haute dans la structure intentionnelle (e.g. DSP2), plus le débit est lent et la durée des pauses longue. En d'autres mots, l'unité qui introduit un nouveau topique est marquée par un débit plus lent que celui de l'unité qui la précède ; par ailleurs, elle est séparée de cette dernière par une pause relativement longue. En français, les changements de topique sont aussi indiqués par des variations du débit d'élocution. En revanche, l'unité qui introduit un nouveau topique n'est pas toujours marquée par un débit plus lent que celui de l'unité qui la précède. Les changements de topique sont aussi indiqués par des pauses plus longues.

Comme pour les variations de registre, nous observons ici aussi de nombreux résidus au niveau des DSP0 (i.e. continuité de topique), et ce, pour les deux corpus. Ces résidus suggèrent, selon nous, que les variations du débit d'élocution n'indiquent pas uniquement des changements de topique et que ces autres fonctions sont révélées sous l'étiquette DSP0. En effet, nous avons expliqué dans le chapitre 3 que l'étiquette DSP0 peut être considérée comme une étiquette « fourre-tout »qui indique à la fois une continuité de topique mais qui révèle aussi d'autres annotations fonctionnelles possibles, comme celles de l'information nouvelle, du focus ou encore du changement de l'état émotionnel du locuteur. Il en est de même pour les pauses. Certaines pauses longues se trouvent en effet au niveau des DSP0 mais leur présence et leur durée n'indiquent pas une continuité de topique.

Puisque nous avons montré qu'il existait une corrélation entre les variations de débit et de durée des pauses et la structure intentionnelle du discours, nous étudions à présent, comme nous l'avons fait pour le registre, la possibilité d'utiliser cette corrélation afin de prédire la position temporelle des marqueurs de discours à partir des variations de débit et de la durée des pauses. Nous pouvons d'ores et déjà formuler l'hypothèse, comme nous l'avons fait pour le registre, que le taux de prédiction ne sera pas très élevé vu le nombre de résidus observés pour la catégorie DSP0.

5.5 Prédiction des changements de topique à partir de la détection automatique des variations de tempo

La prédiction des changements de topique se fait à partir de classifieurs bayésiens qui permettent, en fonction de la valeur d'un paramètre discret ou continu (e.g. la différence de débit entre deux unités ou encore la durée de la pause), d'attribuer une classe particulière (e.g. présence d'une frontière DSP2). Nous ne revenons pas sur l'explication du classifieur que nous

avons donnée dans la section 5.5 du chapitre 3.

Dans cette étude, nous avons considéré 2 classes pour la prédiction des changements de topique : l'une correspondant aux unités étiquetées DSP0, l'autre aux unités DSP1 et DSP2. Comme nous l'avons expliqué dans le chapitre 3, ce choix a été effectué compte tenu de l'importante assymétrie de la distribution de nos classes. Les DSP0 à eux seuls représentent en effet plus de 90% des données. L'entrainement des classifieurs binaires se fait à partir de nos échantillons de données (i.e. à partir des corpus AM et CID). Nous testons, comme paramètres d'entrée des différents classifieurs, le débit d'élocution de l'unité (RATE), la différence de débit entre deux unités (DRATE), la distance gauche entre deux unités (LDIST), la durée de la pause (DPAUSE) et leur combinaison.

L'évaluation des classifieurs est obtenue pour chaque échantillon par la construction de matrices de confusion (i.e. les classes observées vs. les classes prédites). Les scores des F-mesures associés à la prédiction des DSP1|DSP2 sont donnés dans le tableau 20 [101]. Les scores des classifieurs basés uniquement sur un paramètre montrent que la différence de débit entre deux unités permet l'obtention de scores plus élevés que les paramètres débit et distance gauche (en moyenne 30% vs. 18% et 25%). La combinaison de ces paramètres n'augmentent d'ailleurs pas le score obtenu à partir d'un seul paramètre. La durée des pauses est le paramètre qui permet l'obtention des scores les plus élevés (en moyenne de 57%). Il est à noter que la prédiction est en général meilleure pour le corpus AM que pour le CID.

Corpus	RATE	DRATE	LDIST	COMB	DPAUSE
AM	0.22	0.37	0.28	0.37	**0.62**
CID	0.14	0.23	0.23	0.24	**0.53**

TABLE 20 – Scores des F-mesures pour les classifieurs basés sur les paramètres débit d'unité (RATE), différence de débit entre deux unités (DRATE), distance gauche d'unité (LDIST), sur la combinaison de ces trois paramètres (COMB) et sur le paramètre durée des pauses (DPAUSE).

Si nous rappelons à présent les scores obtenus à partir des variations de registre et que nous les comparons à ceux obtenus avec les paramètres débit d'élocution et durée des pauses (cf. tableau 21), nous constatons que les pauses sont un meilleur prédicteur des changements de topique que les variations de registre et de débit d'élocution.

101. Le détail des matrices est donné sur CD ROM - ANNEXES_CHAP4 : Table4.

Corpus	DKEY	DRATE	DPAUSE
AM	0.38	0.37	0.62
CID	0.27	0.23	0.53

TABLE 21 – Scores des F-mesures pour les classifieurs basés sur les paramètres différence de hauteur du registre (DKEY), différence du débit d'unité (DRATE) et durée des pauses (DPAUSE).

Ce constat est très certainement dû au fait que le registre et le débit d'élocution varient pour de multiples raisons, peut être plus nombreuses que celles des pauses. Il est à noter que le taux d'erreur, à partir d'une prédiction basée sur la durée des pauses uniquement, tombe à 43% en moyenne (vs. 65% basée sur les variations de registre et de débit), ce qui nous laisse à penser qu'il sera possible dans le futur d'annoter automatiquement les changements de topique. La combinaison de ces trois paramètres (registre, débit et pause), accompagnée d'une annotation plus fine, pourrait en effet augmenter la qualité d'une annotation automatique de la structure intentionnelle du discours.

La détection automatique des variations de débit d'élocution, que permet l'outil ADoTeVA, peut en revanche, comme nous l'avons soulevé plus haut, être critiquée. En effet, l'algorithme de détection, en l'état, néglige toute interaction des variations de débit d'élocution avec d'autres effets, tels que l'allongement de frontière. Nous avons pourtant expliqué, lors du premier chapitre, que les deux effets étaient très liés, l'allongement de frontière étant parfois décrit comme une réduction localisée du débit. Nous ne pouvons donc certifier que la détection automatique des variations de débit ne reflète pas aussi celle des effets de frontière.

Une façon de séparer ces deux phénomènes, nous l'avons dit, c'est de déterminer le locus de leur effet. Nous avons vu que les variations de tempo agissaient sur un empan assez global, contrairement aux effets de frontière qui eux, au vu de la littérature, semblent plus localisés, e.g. au niveau de la rime de la dernière syllabe d'un constituant.

Nous terminons donc ce chapitre par l'étude de l'allongement de frontière, notamment nous chercherons à évaluer le locus de son effet. L'étude que nous proposons est, à ce stade, préliminaire. Elle porte uniquement sur l'allongement de frontière au niveau de l'unité intonative, en anglais britannique standard. Elle se situe par ailleurs dans le prolongement des travaux de Hirst, Bouzon et Auran (Bouzon & Hirst, 2002; Auran et al., 2004; Bouzon & Hirst, 2004; Hirst & Bouzon, 2005; Hirst, 2009), que nous rappellerons ci-après. Nous montrerons que cette analyse est très complexe et qu'elle mérite à elle seule une étude séparée.

6 De l'empan temporel de l'allongement de frontière

6.1 Rappel : domaine et locus

Nous avons montré, au cours du premier chapitre, que l'allongement de frontière se situe aux bornes gauche et droite de constituants morpho-syntaxiques et/ ou prosodiques. Au vu de la littérature, il est difficile de déterminer quels sont les domaines de l'allongement de frontière, mais il semblerait que l'unité rythmique étroite, le syntagme accentuel, l'unité intonative ou encore l'énoncé peuvent être des domaines possibles d'allongement de frontière initial et final en anglais (Edwards & Beckman, 1987; Beckman & Edwards, 1990; Fletcher & McVeigh, 1992; Wightman et al., 1992; Fougeron & Keating, 1997; Cummins, 1999; Byrd et al., 2006; Hirst et al., 2007; Yoon et al., 2007; Turk & Shattuck-Hufnagel, 2007). Le locus de l'allongement initial semble se situer sur l'attaque de la première syllabe du constituant. Le locus de l'allongement final, lui, semble opérer au niveau de la rime de la syllabe finale du constituant. Il toucherait aussi l'attaque ou la rime de la syllabe accentuée (*main-stress syllable*) du constituant lorsqu'elle n'est pas en position finale (Lindblom, 1968; T. Crystal & House, 1990; Wightman et al., 1992; Cambier-Langeveld, 2000; White, 2002; Turk & Shattuck-Hufnagel, 2007). Les auteurs suggèrent alors un allongement de frontière final multiple. Par ailleurs, on constate que l'allongement de frontière final est progressif : plus les segments sont proches de la frontière, plus l'effet de l'allongement serait important. Enfin, le degré des allongements de frontière initial et final marquerait le niveau du constituant dans la hiérarchie prosodique. Plus le constituant serait élevé dans la hiérarchie, plus le degré d'allongement de frontière serait important (Ladd & Campbell, 1991; Wightman et al., 1992; Fougeron & Keating, 1997; Cambier-Langeveld, 2000; Gee & Grosjean, 2002; Byrd & Saltzman, 2003; White, 2002; Yoon et al., 2007).

6.2 Premières études menées sur le corpus Aix-Marsec

La plupart des études qui portent sur l'analyse et la modélisation de la durée segmentale se basent sur des corpus de phrases lues, contrôlées. Notre objectif est ici de comprendre l'organisation temporelle de la parole, lorsqu'elle est motivée par les intentions communicatives du locuteur. Plus précisément, nous cherchons à déterminer le locus de l'allongement de frontière au niveau de l'unité intonative en parole authentique. Le corpus Aix-Marsec offre la possibilité d'une telle analyse. Il représente en effet, comme nous l'avons mentionné dans le premier chapitre de cette thèse, 5h de parole authentique, représentative de 11 types de productions différents. Riche d'annotations, il permet l'étude de la durée segmentale au niveau d'unités multiples : le phonème, la syllabe, le pied accentuel, l'unité rythmique (étroite et anacrouse), le mot et l'unité intonative. Il indique également à la fois la position de l'unité et le nombre

d'unités au sein des unités supérieures. Il est riche d'autres informations encore : le facteur accentuel de la syllabe, sa composition (i.e. attaque, noyau, coda), le type de phonème (voyelle courte, voyelle longue, diphtongue, consonne), etc.

Les premières études menées sur le corpus (Hirst & Bouzon, 2005) ont cherché à déterminer quelle unité rythmique est la plus appropriée pour décrire les phénomènes temporels. Les auteurs démontrent les effets de la structure prosodique en effectuant des corrélations linéaires entre le nombre de segments au sein d'une unité et la durée de ces segments au sein de cette dernière. Ils ont tout d'abord observé de fortes corrélations négatives entre la durée des segments et leur nombre dans le pied accentuel, dans l'unité rythmique étroite et dans le mot. Aucune corrélation n'est en revanche démontrée au niveau de la syllabe et de l'anacrouse. De plus, les auteurs mentionnent que le degré de compression est plus important au niveau de l'unité rythmique étroite qu'au niveau du mot et du pied. L'unité rythmique étroite apparaît donc, en anglais, l'unité optimale pour décrire les phénomènes temporels de la parole.

Les auteurs montrent également un allongement final significatif au niveau de l'unité intonative. Le locus de l'allongement est graphiquement déterminé comme les 3 derniers phonèmes de l'unité (ou peut-être les 5 derniers). Dernièrement, Hirst (2009) s'est tout particulièrement intéressé au phénomène d'allongement de frontière, au niveau de l'unité rythmique étroite. Après avoir exclu les trois derniers phonèmes de l'unité intonative, l'auteur rapporte, de l'analyse de variance, un effet significatif du nombre de phonèmes et de la position du phonème dans l'unité rythmique étroite. Lorsque les phonèmes sont codés en fonction de leur position, i.e. (1) initial (pour le premier phonème de l'unité), (2) médian (pour les phonèmes au sein de l'unité qui ne sont ni en position initiale ni en position finale), et (3) final (pour le dernier phonème de l'unité), l'analyse de variance montre à nouveau des différences significatives entre les trois positions ($p<2.2e\text{-}16$). Les valeurs moyennes des durées des phonèmes obtenues (z-score) sont de 0.245 pour les phonèmes en position initiale, -0.118 pour ceux en position médiane et 0.073 pour ceux en position finale. Ces résultats suggèrent donc un allongement initial sur l'attaque de la syllabe initiale et un allongement final sur le dernier phonème de l'unité rythmique étroite, qui est donc le dernier phonème du mot accentué.

L'algorithme de modélisation du rythme de la parole, ProZed, développé par Hirst et al. (2007), peut être dès lors enrichi. L'algorithme, en l'état, calcule en effet, pour chaque énoncé, une valeur de tempo, puis, un poids scalaire est assigné à chaque unité rythmique de l'énoncé traité. Pour décrire l'effet d'allongement, au niveau de l'unité rythmique, un facteur d'allongement, indépendant de la taille de l'unité, lui est incrémenté. Les objections qui peuvent être cependant portées à cet algorithme sont que (1) le tempo est calculé sans avoir, en amont, neutralisé les effets d'allongement de frontière et que (2) le facteur d'allongement est réparti uniformément sur les différents segments de l'unité rythmique. Les derniers résultats obtenus pour l'unité rythmique étroite permettent à présent de déterminer la localisation de l'effet de

frontière au niveau de l'unité rythmique. Reste donc à implémenter les effets de frontière à d'autres niveaux de la structure prosodique et de calculer le tempo des énoncés une fois l'empan temporel de ces effets déterminés. L'étude de ces effets sur d'autres niveaux de constituance est donc nécessaire. Les premiers résultats rapportés pour l'unité intonative suggèrent déjà qu'il est possible de déterminer plus précisément le locus de l'allongement de frontière pour cette unité.

6.3 Etude préliminaire de l'allongement de frontière

102

L'étude menée sur l'allongement de frontière, au niveau de l'unité intonative, est basée sur les 5h de parole du corpus Aix-Marsec. La transcription phonématique a été automatiquement alignée au signal, après avoir défini manuellement les unités intonatives. Quelques erreurs d'alignement ont pu être repérées suite à une inspection manuelle, mais elles peuvent être, malgré la quantité de données, facilement localisées, du fait que la durée du phonème touché par l'erreur d'alignement présente une valeur aberrante (au delà d'1 sec et en deçà de 19ms). Par conséquent, dans notre analyse, nous avons exclu tout phonème supérieur à 500ms et inférieur à 20ms.

Nous nous intéresserons tout particulièrement ici à l'effet de frontière au niveau de l'unité intonative. Nous chercherons en effet à voir si, comme dans la littérature, nous observons un allongement de frontière initial et final sur cette unité. Dans le cas où l'unité intonative est le domaine de l'allongement de frontière, nous chercherons également à déterminer son locus.

6.3.1 Le domaine de l'allongement de frontière : l'unité intonative

Une première représentation graphique (cf. figure 88) de la durée moyenne des phonèmes (en log) en fonction de leur position et de leur nombre dans l'unité intonative nous permet d'asseoir l'hypothèse d'un allongement de frontière final au niveau de cette unité. Au vu du graphique, il semblerait en effet que les deux derniers phonèmes de l'unité intonative soient allongés. Quant à un possible allongement initial, le graphique ne révèle rien en l'état.

102. Nous remercions tout particulièrement Robert Espesser du Laboratoire Parole et Langage pour son aide et ses conseils précieux qui nous ont permis de mener cette étude.

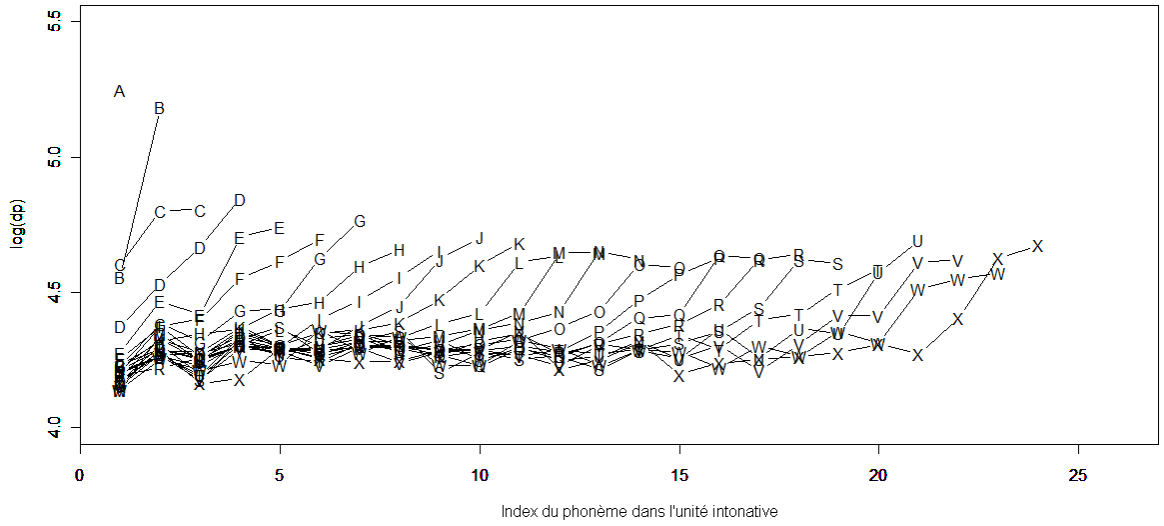


FIGURE 88 – Représentation graphique de la durée moyenne des phonèmes (en log) en fonction de leur position et de leur nombre dans l'unité intonative. A indique que l'unité intonative est constituée d'un seul phonème, B de deux, C de trois, D de quatre, etc.

Une analyse de variance révèle un effet significatif du nombre de phonèmes ($F(1, 157898)=984.3$; p-val< 2.2e-16) dans l'unité intonative et de la position du phonème ($F(71, 157828)=11.86$; p-val<2.2) dans l'unité intonative.

Si nous regardons l'histogramme du nombre de phonèmes dans l'unité intonative (figure 89), nous observons que la plupart des unités intonatives contiennent entre 10 et 20 phonèmes et que les unités constituées de moins de 5 phonèmes et de plus de 30 phonèmes sont moins fréquentes. Le sommaire des données indique plus précisément que l'unité intonative la plus petite contient un seul phonème et que l'unité maximale en contient 74. La moyenne du nombre de phonèmes dans l'unité intonative est de 20 phonèmes, avec le 1^e quartile de la distribution estimé à 13 phonèmes, le 3^{me} quartile estimé à 25 phonèmes. Par conséquent, nous nous intéressons aux effets d'allongement de frontière, uniquement pour les unités intonatives constituées de 5 à 25 phonèmes.

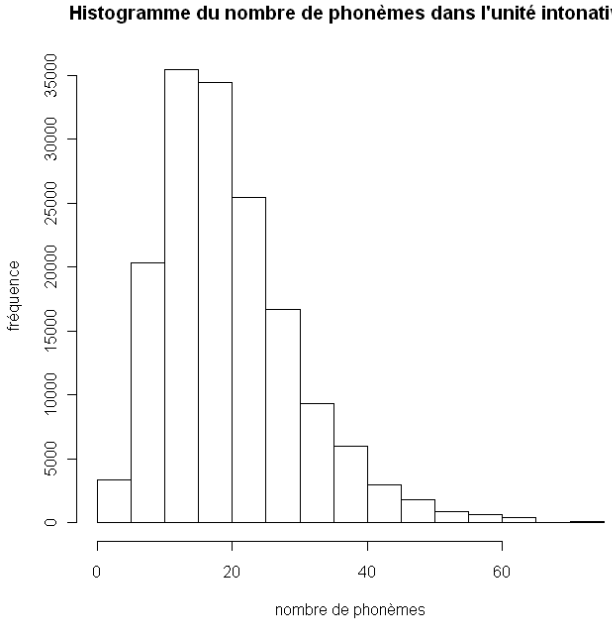


FIGURE 89 – Histogramme du nombre de phonèmes dans l'unité intonative.

Pour toute élaboration de modèle prédictif, il est intéressant, en amont, d'observer la distribution des données à partir desquelles le modèle sera construit. Aussi, avant de traiter plus précisément la question de l'empan temporel de l'allongement de frontière, nous proposons une étude succincte de la distribution de nos données. Cela nous permettra par ailleurs de justifier les choix théoriques portés pour cette analyse.

6.3.2 Analyse de la distribution des données

Si nous regardons l'histogramme des durées des phonèmes de nos données (en ms ; figure 90), nous observons une forte asymétrie à gauche.

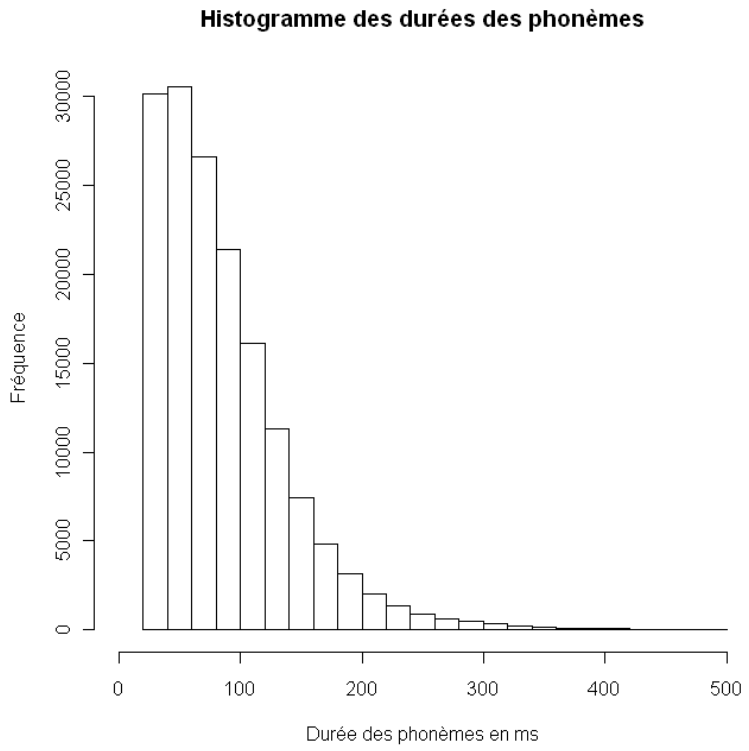


FIGURE 90 – Histogramme des durées des phonèmes (en ms).

Il est par ailleurs intéressant de noter séparément la distribution des durées des consonnes, des voyelles brèves, des voyelles longues et des diphtongues (figure 91). Alors que les histogrammes des durées des consonnes et des voyelles brèves montrent une forte asymétrie à gauche, ceux des durées des voyelles longues et des diphtongues révèlent des distributions un peu plus centrées, bien que toujours portées sur la gauche.

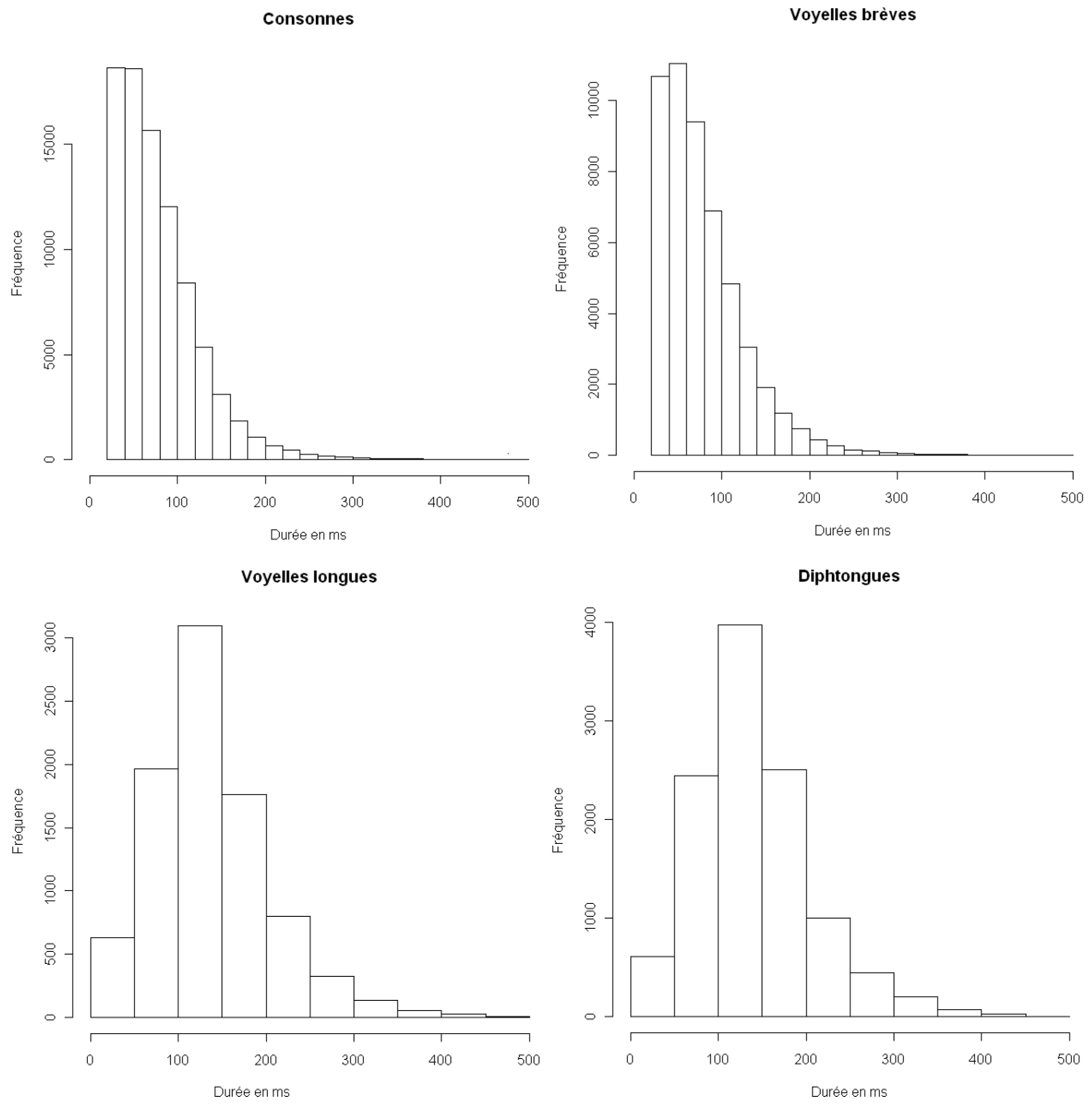


FIGURE 91 – Histogrammes des durées (en ms) des consonnes, des voyelles brèves, des voyelles longues et des diphtongues.

Si nos distributions dans l'ensemble sont similaires aux distributions des durées segmentales observées dans les travaux de Campbell (1992) et de Fletcher et McVeigh (1992), à savoir une forte asymétrie positive, il est à noter (tableau 92) que la distribution des durées des consonnes, à partir du corpus AM, présente une plus forte asymétrie que la distribution des durées des consonnes observée à partir du corpus ATR (Campbell, 1992). Par ailleurs, l'auteur montre une plus forte asymétrie de la distribution des durées des voyelles que celle obtenue pour les durées

des consonnes. Pour notre part, nous avons un effet inverse où les coefficients de dissymétrie obtenus pour les distributions des durées des voyelles sont inférieurs au coefficient obtenus pour la distribution des durées des consonnes. Il est à noter toutefois que les coefficients de dissymétrie obtenus pour la distribution des voyelles brèves et pour celle des consonnes sont très proches.

	Données Aix-Marsec		Données ATR (Campbell, 1992)
Coeff. de dissymétrie (V)	V brèves :	1.739	1.003
	V longues :	1.063	
	Diphtongues :	0.978	
Coeff. de dissymétrie (C)		1.805	0.713

FIGURE 92 – Tableau des valeurs des coefficients de dissymétrie obtenus pour les distributions des durées des voyelles (V) et des consonnes (C), des corpus Aix-Marsec et ATR (tiré de Campbell (1992, 103)).

Afin de réduire l'asymétrie de nos données, nous proposons, comme nous l'avons fait pour les valeurs en Hz dans le premier chapitre de cette thèse, de normaliser les durées des phonèmes brutes (ie. en millisecondes) par une transformation logarithmique. Outre la transformation logarithmique, nous aurions pu opter pour une transformation quadratique ou encore une transformation Gamma. Campbell (1992) montre en effet qu'une transformation Gamma permet une meilleure égalisation des distributions des durées des segments que les transformations logarithmique ou quadratique [103].

Pour notre part, nous nous en tenons à une transformation logarithmique qui apparaît suffisante pour notre base de données. Le tableau 22 montre en effet que les coefficients de dissymétrie et les coefficients d'aplatissement de Pearson [104] permettent l'approximation de distributions centrées, similaires pour les consonnes et les voyelles.

103. pour plus de précisions, cf. Campbell (1992, p98-108).

104. Les définitions des coefficients de dissymétrie et des coefficients d'aplatissement de Pearson ont été données dans le premier chapitre, section 3.4.

	Phonèmes	durées brutes	transfo. log
Coeff. de dissymétrie	V	1.739	0.152
	Vl	1.063	-0.613
	VD	0.978	-0.630
	C	1.805	0.144
Coeff. d'aplatissement	V	8.212	2.479
	Vl	5.006	3.640
	VD	4.566	3.855
	C	8.890	2.485

TABLE 22 – Valeurs des coefficients de dissymétrie (*skewness*) et des coefficients d'aplatissement de Pearson (*kurtosis*), obtenus pour les distributions des durées brutes des voyelles (V : brèves ; Vl : longues ; Vd : diphtongues) et des consonnes et après transformation logarithmique.

6.3.3 Le locus de l'allongement de frontière

Une fois la question de la distribution des données traitée, il est important de prendre en considération l'ensemble des facteurs qui pouvent interférer dans la prédiction segmentale. Lorsque l'on s'intéresse à la question d'allongement de frontière, il faut notamment garder à l'esprit que la durée intrinsèque du phonème ou encore le débit d'élocution du locuteur peuvent influencer les durées observées. Or, ce que l'on cherche à déterminer, c'est un facteur d'allongement, applicable à toutes les données, quel que soit le phonème et quel que soit le débit intrinsèque du locuteur. Pour cela, Campbell et Isard (1991) et Campbell (1992) proposent une transformation z-score, interprétée sous l'hypothèse d'élasticité. La transformation z-score estime la quantité de variation en termes d'écart type autour des moyennes définies pour chaque phonème, par la formule suivante :

$zp = \frac{dp - \mu p}{\sigma p}$,

où zp est le z-score de l'instant donné d'un phonème p, dp la durée brute correspondante, μp et σp la moyenne et l'écart type de la durée de chaque phonème (pour chaque locuteur).

Pour gommer l'effet du phonème (i.e. l'effet de sa durée intrinsèque) et le débit d'élocution du locuteur, on peut aussi utiliser un modèle à effets mixtes (ou modèle mixte ; Pinheiro et Bates (2000)), modèle adapté au traitement de données non équilibrées et ayant une structure de groupe. Les effets aléatoires représentent en effet des déviations par rapport à la moyenne de chaque groupement. Ici, par exemple, on distingue deux niveaux de groupement : les locuteurs et les phonèmes. Ainsi, si l'on compare les critères d'information d'Akaike d'un modèle

linéaire simple, construit à partir des valeurs z-score des durées des phonèmes, et d'un modèle mixte, également construit à partir des valeurs z-score et où le facteur aléatoire est le type de phonème, nous observons que le modèle mixte, du moins le facteur aléatoire du modèle mixte, n'apporte rien à la modélisation (AIC=17145 vs. AIC=17223). La transformation z-score a suffi à neutraliser la contribution spécifique de l'identité du phonème.

Pour notre part, nous utiliserons des modèles à effets mixtes pour mesurer l'empan temporel des allongements de frontière initial et final. Nous pensons en effet qu'ils permettent une analyse plus fine des données que la transformation z-score. Cependant, nous utiliserons au cours de cette étude une telle transformation et montrerons qu'elle est essentielle lorsque l'on cherche à représenter graphiquement les données. Nous proposons, avant de présenter les résultats de cette étude, d'expliciter ce qu'on entend par modèle mixte.

Un modèle mixte distingue des effets fixes et des effets aléatoires. Les effets fixes sont les variables contrôlées par l'expérience, les effets aléatoires sont les variables non contrôlées, dont on ne dispose que d'un échantillonage partiel. Dans les modèles mixtes de notre étude, l'unique effet fixe du modèle est la position du phonème dans l'unité intonative (désormais **ipi**), décrit par un facteur à n niveaux où n est le rang maximal du phonème dans l'unité. Nous y avons introduit deux facteurs aléatoires simples, l'un pour prendre en compte la variabilité inter-locuteur (désormais **loc**), l'autre pour prendre en compte la variabililité inter-phonèmes (désormais **phon**).

La phase exploratoire de cette étude consiste en l'élaboration de 21 modèles. Nous avons expliqué, au vu de la distribution du nombre de phonèmes dans les unités intonatives de notre corpus, que nous nous intéressions uniquement aux unités intonatives constituées de 5 à 25 phonèmes. Nos modèles sont donc construits à partir de sous-ensembles. Chacun d'eux correspond à une « longueur » d'unité intonative. Le premier modèle, par exemple, est élaboré à partir de données qui représentent les unités intonatives composées de 5 phonèmes. Dans ce modèle donc, l'effet fixe est modélisé à l'aide de 4 contrastes. Ainsi, dans nos modèles, à travers une comparaison des différences successives entre chaque niveau, nous obtenons des informations sur la façon dont chaque niveau de facteur (i.e. la position du phonème) influence la variable dépendante (i.e. la durée segmentale).

La construction des modèles nous permet ainsi d'appréhender les effets d'allongement initial et final pour chaque type d'unités, de valider ou non leur effet sur la durée segmentale et de « prédéterminer » la région de leur effet. La détermination des degrés de liberté des modèles mixtes étant toujours sujet de controverse, le module utilisé (package lmer) ne fournit pas de valeurs classiques p pour les t-tests. Nous optons ici pour la valeur *pMCMC* (*Monte Carlo Markov Chain*), calculée par un échantillonage MONTE CARLO, basé sur la construction d'une chaîne de Markov. Dans cette étude, tout effet significatif correspond à un α de *pMCMC* <0.01.

Avant de présenter les résultats de notre analyse statistique, nous cherchons à valider, en premier lieu, la pertinence de l'inclusion des variables aléatoires **phon** et **loc** dans nos modèles à effets mixtes. La significativité de l'inclusion des facteurs à effets aléatoires est évaluée par une série de tests de vraisemblance. Ce type de test s'applique à deux modèles mixtes emboîtés (i.e. dont l'un des deux modèles est le sous-ensemble de l'autre), s'ils incorporent les mêmes effets fixes mais un nombre différent de facteurs à effets aléatoires (Pinheiro & Bates, 2000). A cet effet, nous comparons un modèle mixte dont le facteur aléatoire est **phon** à un modèle dont les facteurs aléatoires sont **phon** et **loc**, afin d'évaluer si l'ajout **loc** est pertinente. Le test de vraissemblance indique la valeur logarithmique de vraissemblance (LogLik) du modèle le plus petit (i.e. avec un seul facteur à effets aléatoires) et la compare à celle obtenue pour le modèle le plus grand (i.e. avec deux facteurs à effets aléatoires) (Baayen, 2008). De la différence entre les deux valeurs de vraissemblance ($(-3739.4) - (-4286.7) = 1094.6$), multipliée par deux, résulte une distribution de chi-carré, dont les degrés de liberté sont calculés comme la différence du nombre de paramètres ($23 - 22 = 2$). La probabilité associée à la distribution ($p<2.2e\text{-}16$) permet ici de justifier l'inclusion du facteur aléatoire **loc**.

Nous comparons également le modèle mixte dont le facteur aléatoire est **loc** au modèle à deux facteurs aléatoires (i.e. **phon** et **loc**), afin d'évaluer la pertinence de l'inclusion du facteur à effets aléatoires **phon**. La probabilité associée à la distribution de chi-carré est également ici très petite ($p<2.2e\text{-}16$) ce qui nous permet de valider l'inclusion du facteur aléatoire **phon**.

Les tests de vraissemblance, effectués pour la comparaison d'un modèle linéaire simple (i.e. sans facteur à effets aléatoires) à un modèle mixte à deux facteurs, révèlent tout naturellement que le modèle mixte est préférable à un modèle simple ($p=0$). Les critères d'information d'Akaike (AIC), qui complètent cette analyse, sont affichés dans le tableau 23. L'AIC le plus petit est en effet celui du modèle mixte à deux facteurs aléatoires.

Type de modèle	AIC
Simple (lm)	9936
Mixte (f.a. **loc**)	8773
Mixte (f.a. **phon**)	8720
Mixte (f.a. **loc** + **phon**)	7625

TABLE 23 – Tableau des critères d'information d'Akaike (AIC) pour chacun des modèles : Simple (i.e. sans facteur aléatoire), Mixte avec le facteur aléatoire (f.a.) **loc**, Mixte avec le facteur aléatoire **loc** et Mixte avec les deux facteurs aléatoires **loc** + **phon**.

Les intercepts estimés par le modèle mixte pour le facteur aléatoire (**phon**) sont représentés dans le graphique 93. On peut ainsi observer que les valeurs les plus importantes correspondent

aux voyelles longues et aux diphtongues alors que les valeurs les plus petites correspondent aux consonnes voisées, reflétant ainsi la durée intrinsèque de chaque phonème. Il est à noter que de nombreux intercepts sont éloignés de la valeur 0, ce qui confirme la pertinence de l'introduction de **phon** dans le modèle mixte.

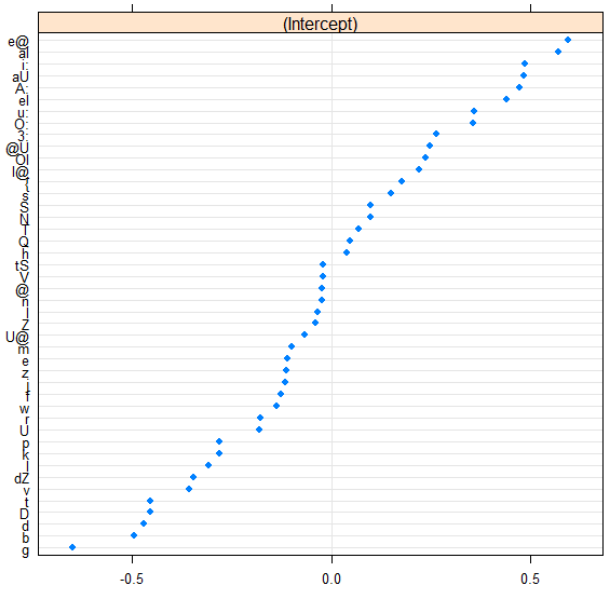


FIGURE 93 – Graphique des valeurs du facteur à effets aléatoires **phon**.

Les intercepts produits par le modèle mixte pour le facteur aléatoire (**loc**) (cf. figure 94) révèlent aussi des différences de débit entre locuteurs et justifient le choix de cette variable aléatoire.

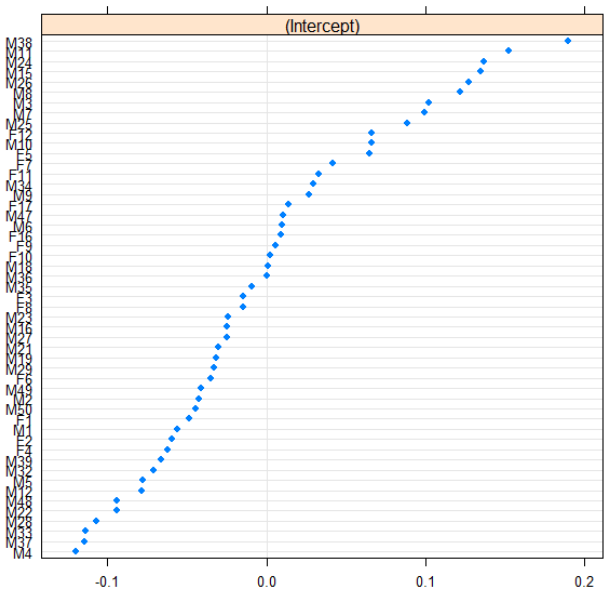


FIGURE 94 – Graphique des valeurs du facteur à effets aléatoires **loc**.

L'utilisation d'un modèle mixte à deux facteurs aléatoires à présent justifiée, nous proposons de présenter les résultats obtenus pour chacun de nos modèles. Dans le tableau 24, sont présentés les niveaux de facteurs (i.e. la position du phonème) qui ont un effet significatif ($pMCMC$ $<.01$) sur la durée segmentale.

Type de modèles	Contrastes Diff.	Estimation	pMCMC
Modèle 5	F-1	0.1350	0.0012
	F	0.1916	0.0001
Modèle 6	F	0.1824	0.0001
Modèle 7	F-1	0.1062	0.0016
	F	0.1820	0.0001
Modèle 8	F	0.1805	0.0001
Modèle 9	F	0.1749	0.0001
Modèle 10	F-2	0.0932	0.0004
	F-1	0.1135	0.0002
	F	0.1276	0.0001
Modèle 11	F-1	0.0878	0.0028
	F	0.1433	0.0001
Modèle 12	F-1	0.1227	0.0001
	F	0.0987	0.0016
Modèle 13	I	0.1147	0.0004
	F-1	0.1607	0.0001
	F	0.0832	0.0092
Modèle 14	F-1	0.1815	0.0001
Modèle 15	F-1	0.1359	0.0001
Modèle 16	F-1	0.0913	0.0038
	F	0.1082	0.0014
Modèle 17	F-1	0.1985	0.0001
Modèle 18	I+1	0.1232	0.0008
	AV	0.1723	0.0001
Modèle 19	F-1	0.1391	0.0002
Modèle 20	F-1	0.1056	0.0088
Modèle 21	F-1	0.1643	0.0001
	F	0.1498	0.0008
Modèle 22	F-1	0.2009	0.0001
Modèle 23	F-2	0.1363	0.0032
Modèle 24	F-1	0.1941	0.0002
Modèle 25	F-1	0.1883	0.0001

TABLE 24 – Tableau des paramètres estimés pour chaque niveau de facteur qui ont un effet sur la durée segmentale. *Modèle x* indique que, dans ce modèle, seules les unités constituées de x phonèmes sont considérées. I, I+1, F-2, F-1 et F indiquent la position du phonème dans l'unité intonative : I indique la position initiale, I+1 la seconde position, F-2, la position antépénultième, F-1, l'avant-dernière position et F la position finale. La colonne pMCMC donne la significativité du niveau du facteur.

Au vu du tableau, nous ne trouvons aucun effet significatif du facteur position pour les premiers niveaux de facteurs (i.e. position initiale). Bien que deux modèles présentent un effet significatif de la position initiale sur la durée segmentale, l'effet ne peut être généralisable à l'ensemble des données. L'allongement initial que l'on trouve pour les modèles 13 et 18 résulte, selon nous, d'un artefact, que seule une analyse approfondie permettrait d'expliquer. Ces résultats, en revanche, suggèrent que l'effet de frontière initial, rapporté par (Hirst et al., 2007), au sein de l'unité rythmique, n'est pas, lui, un artefact ; il ne peut être expliqué par un effet d'allongement initial au niveau de l'unité intonative puisque cet effet est inexistant. L'allongement initial au niveau de l'unité rythmique résulterait plutôt du fait que tout phonème en position initiale d'unité rythmique se trouve dans une syllabe accentuée (*stressed*). L'allongement initial au niveau de l'unité rythmique serait donc plutôt un allongement de proéminence qu'un allongement de frontière.

En revanche, on peut conclure à un allongement de frontière final au niveau de l'unité intonative. Il est cependant difficile, à ce stade de l'analyse, de déterminer son locus. L'empan temporel varie en effet entre :

- les 3 derniers phonèmes de l'unité intonative (fréquence : 5%)

- les 2 derniers phonèmes de l'unité intonative (fréquence : 33%)

- le dernier phonème de l'unité intonative (fréquence : 14%)

- l'avant dernier phonème de l'unité intonative (fréquence : 43%)

- l'antépénultième phonème de l'unité intonative (fréquence : 5%).

Il serait donc intéressant d'étudier l'effet de frontière par la prise en compte de la position du phonème dans la syllabe finale (i.e. la composition de la syllabe : attaque, noyau, coda) de l'unité intonative. En effet, au vu de ces résultats, il semblerait que seule la dernière syllabe de l'unité intonative soit touchée par l'effet de frontière, et le fait que, dans certains modèles, seul l'avant dernier phonème soit significativement allongé, peut nous mener à penser que le noyau de la syllabe, plus que l'attaque et le coda, subit l'effet d'allongement de frontière. C'est ce que nous chercherons à étudier dans la suite de cette analyse.

Nous proposons à présent l'analyse de la durée des phonèmes en fonction de la position du phonème dans l'unité intonative et de la structuration de la syllabe (i.e. attaque, noyau, coda). Au vu du graphique 95, qui représente la durée du phonème (en log) en fonction de sa position dans l'unité intonative et en fonction de sa position dans la syllabe (i.e. attaque, noyau, coda), il apparaît que les 4 derniers phonèmes de l'unité intonative sont allongés lorsqu'ils occupent la position de noyau. Les phonèmes qui occupent les positions d'attaque et de coda, pourraient,

eux, répondre d'un phénomène général de ralentissement, bien que les deux derniers phonèmes ne semblent pas touchés.

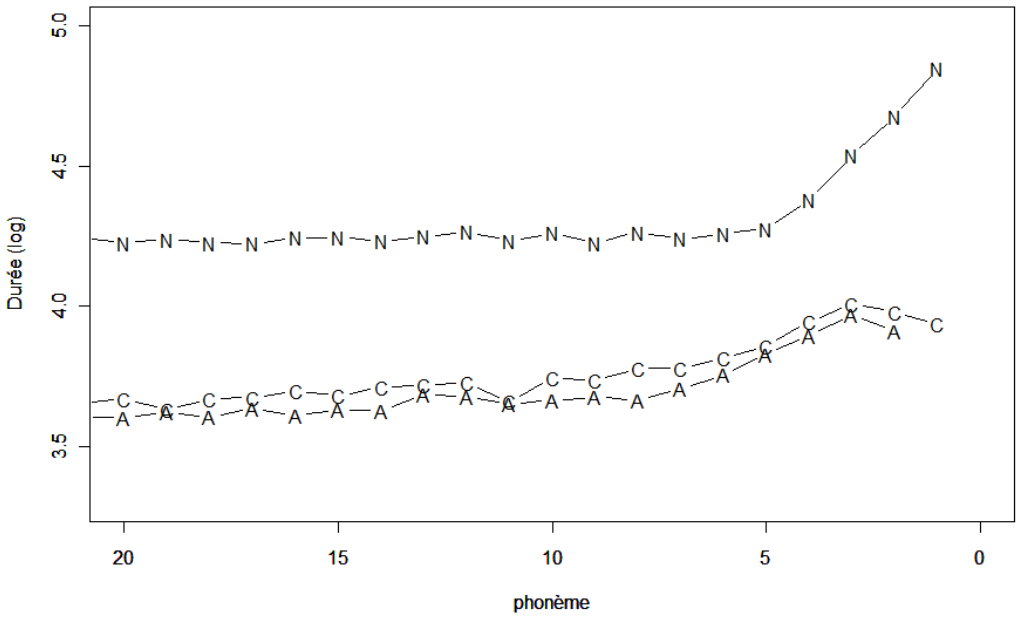


FIGURE 95 – Graphique des durées des phonèmes en fonction de leur position dans l'unité intonative et en fonction de leur position dans la syllabe. A représente l'attaque, N le noyau et C le coda. L'axe des abscisses représente la position du phonème à partir de la fin de l'unité intonative (position finale codée 1 ; avant-dernière codée 2, etc.), l'axe des ordonnées représente la durée du phonème en log.

Si l'on regarde à présent la durée des phonèmes en fonction de la position de la syllabe dans l'unité intonative et en fonction de la composition syllabique (figure 96), on voit que le noyau des deux dernières syllabes ainsi que les codas et onsets des 3 dernières syllabes subissent l'effet de l'allongement de frontière, où les consonnes révèlent une augmentation progressive de l'allongement, jusqu'à la fin de l'unité. L'effet de frontière semble donc plus clair lorsque déterminé en termes de position de la syllabe plutôt qu'en termes de position du phonème. On notera, par ailleurs, que l'allongement est plus important pour le noyau de la dernière syllabe que pour le noyau de l'avant dernière syllabe. On peut se demander ici si l'allongement au niveau de l'avant-dernière syllabe n'est pas le reflet d'un allongement de proéminence. Il semble donc intéressant d'étudier l'éventuelle corrélation entre le caractère accentuel de la syllabe et l'effet de frontière sur la durée segmentale.

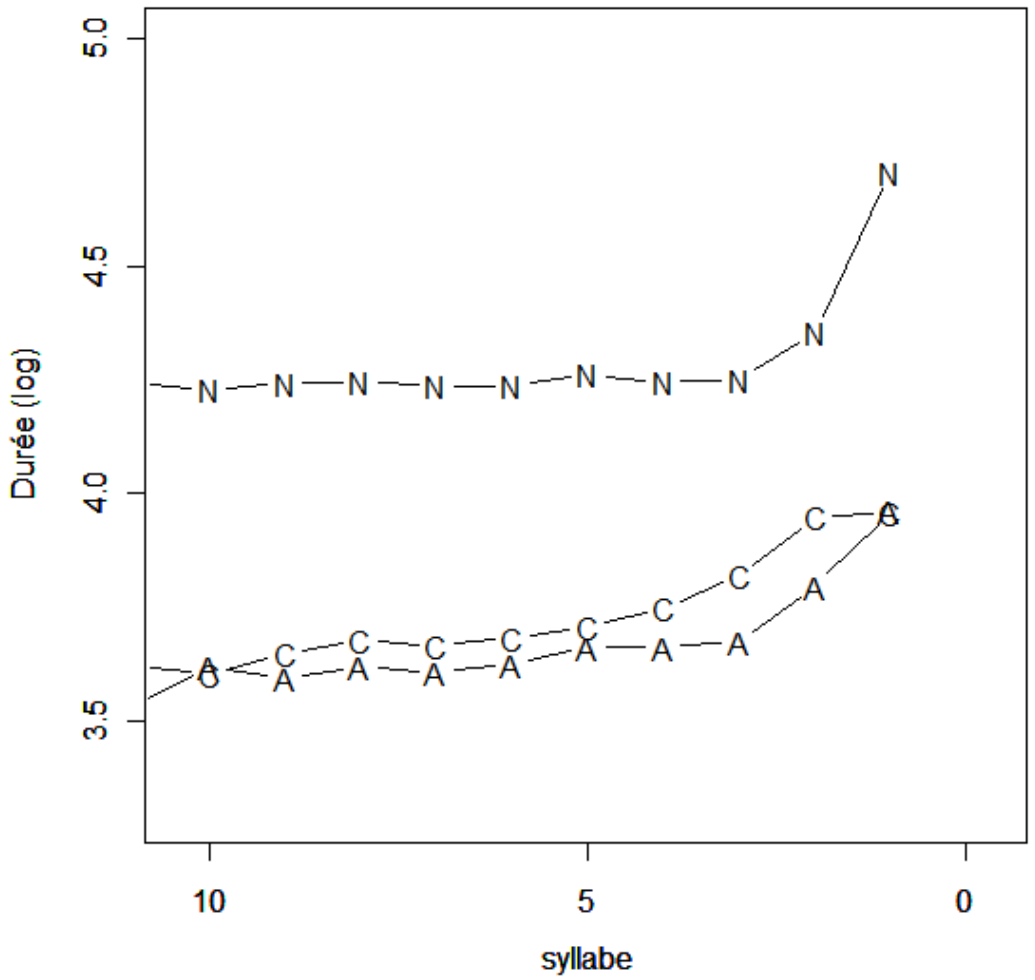


FIGURE 96 – Graphique des durées des phonèmes en fonction de la position de la syllabe dans l'unité intonative et en fonction de la composition syllabique. A représente l'attaque, N le noyau et C le coda. L'axe des abscisses représente la position de la syllabe dans l'unité intonative (plus la position est petite, plus la syllabe est proche de la fin de l'unité), l'axe des ordonnées représente la durée du phonème en log.

Le graphique 97 montre l'allongement du phonème en position d'attaque, de noyau et de coda, en fin d'unité intonative en fonction du caractère accentuel de la syllabe. Ici aussi, nous observons un plus fort allongement au niveau des phonèmes en position de noyau qu'en position d'attaque ou de coda. Une telle différence pourrait s'expliquer par la nature intrinsèque des phonèmes, les voyelles des noyaux étant intrinsèquement plus longues que les consonnes des

attaques et des codas. Il semble aussi, au vu du graphique, que les phonèmes accentués en position de noyau et d'attaque (N et O) sont plus allongés que les phonèmes inaccentués, ce qui se voit d'ailleurs très clairement au niveau des attaques. La nature intrinsèque du phonème expliquerait le fait que les noyaux accentués sont plus allongés que les noyaux inaccentués, qui sont, eux, porteurs de voyelles brèves ou réduites. En revanche, la différence entre les codas accentués et inaccentués est moins flagrante. Quant au locus de l'allongement final, il semble se situer au niveau des 4 derniers phonèmes de l'unité intonative, quel que soit le caractère accentuel de la syllabe et quelle que soit leur position dans la syllabe. On note cependant que l'allongement est plus important au niveau des phonèmes en position de noyau pour les syllabes accentuées et inaccentuées.

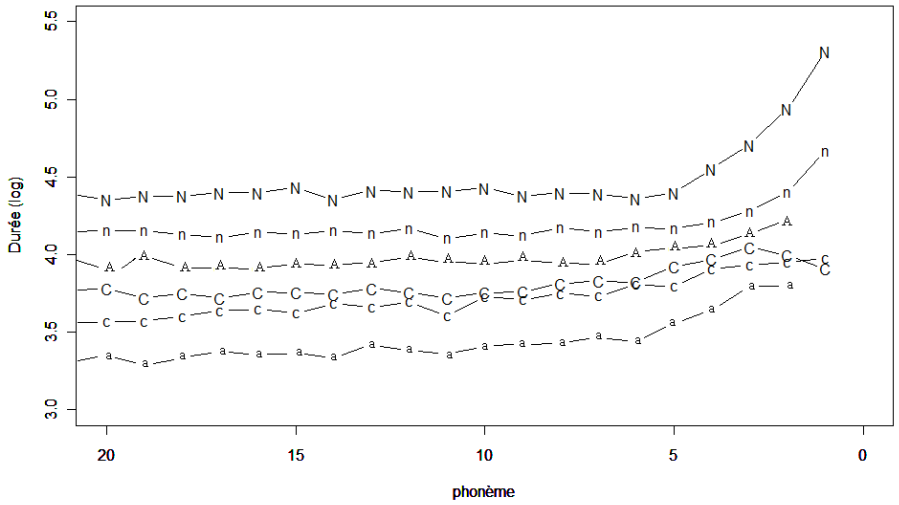


FIGURE 97 – Graphique des durées des phonèmes en position d'attaque, de noyau et de coda au niveau de l'unité intonative, en fonction du caractère accentuel de la syllabe. A, N et C représentent les attaque, noyau et coda des syllabes accentuées ; a, n et c, les attaque, noyau et coda des syllabes inaccentuées. L'axe des abscisses représente la position du phonème dans l'unité intonative (plus la position est petite, plus le phonème est proche de la frontière droite de l'unité), l'axe des ordonnées représente la durée du phonème en log.

Afin de comprendre à quel niveau de la structure prosodique se situe l'allongement des 4 phonèmes finaux, nous représentons graphiquement la durée du phonème en fonction de sa position dans la syllabe (accentuée et inaccentuée) et en fonction de la position de la syllabe dans l'unité (cf. figure 98). Le graphique révèle un allongement des noyaux des deux dernières syllabes accentuées et inaccentuées, bien que l'allongement soit plus marqué au niveau de la syllabe accentuée ; les attaques des deux dernières syllabes accentuées et inaccentuées sont

aussi touchées par l'effet de frontière. Par ailleurs, les attaques des syllabes inaccentuées pourraient répondre d'un phénomène général de ralentissement. La représentation graphique des codas suggère aussi un allongement progressif des codas des syllabes jusqu'à la fin de l'unité intonative.

Au vu de ce graphique, il semblerait donc que les noyaux des syllabes accentuées et inaccentuées et les attaques des syllabes accentuées se comportent de façon identique sous l'effet de frontière, i.e. avec un allongement sur les deux dernières syllabes de l'unité intonative. En revanche, les codas des syllabes accentuées et inaccentuées et les attaques des syllabes inaccentuées semblent plutôt reléter un phénomène global de ralentissement. Le locus de l'allongement de frontière, pour les noyaux des syllabes accentuées et des syllabes inaccentuées, et pour les attaques des syllabes accentuées, serait donc concentré sur les deux dernières syllabes de l'unité intonative, alors que, pour les attaques des syllabes inaccentuées et les codas des syllabes accentuées et inaccentuées, il s'étendrait sur les 7 dernières syllabes de l'unité, où l'effet de frontière serait progressif tout en étant plus marqué sur les deux dernières syllabes.

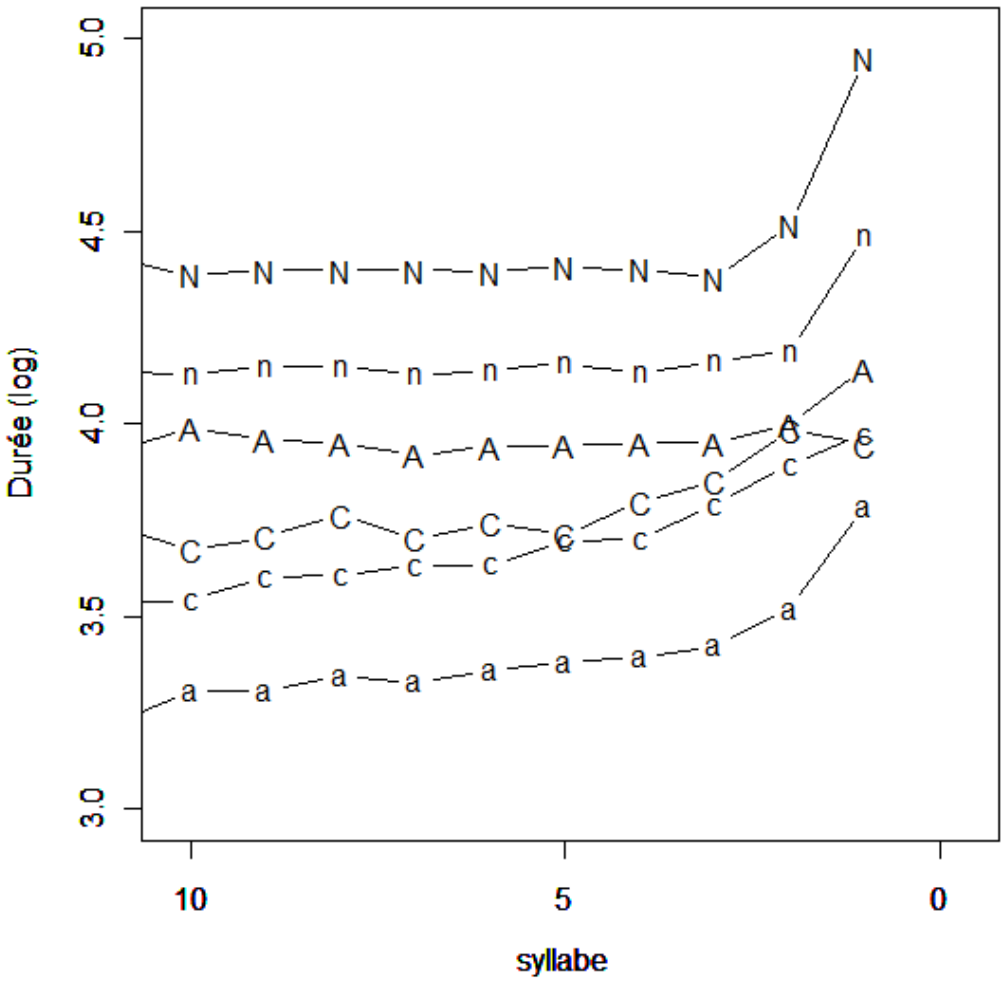


FIGURE 98 – Graphique des durées des phonèmes en fonction de leur position dans la syllabe (accentuée et inaccentuée) et en fonction de la position de la syllabe dans l'unité intonative. A, N et C représentent les attaque, noyau et coda des syllabes accentuées ; a, n et c, les attaque, noyau et coda des syllabes inaccentuées. L'axe des abscisses représente la position de la syllabe dans l'unité intonative (plus la position est petite, plus la syllabe est proche de la frontière droite de l'unité), l'axe des ordonnées représente la durée du phonème en log.

Il est toutefois important de concevoir que les différences qu'on trouve, par exemple entre les noyaux accentués et les noyaux inaccentués, ou encore entre les noyaux accentués et les codas accentués, peuvent résulter de la durée intrinsèque des phonèmes. Il serait donc également intéressant de représenter graphiquement les durées des phonèmes, en fonction de leur

position dans l'unité intonative et en fonction de leur position dans les syllabes accentuées et inaccentuées, après avoir été normalisées par une transformation z-score, pour gommer tout effet de la nature intrinsèque du phonème.

Le graphique 99 représente la durée des phonèmes en fonction de leur position dans la syllabe et en fonction de la position de la syllabe dans l'unité intonative. La transformation z-score révèle que la forte différence qu'on avait notée entre les noyaux d'un côté, et les codas et les attaques de l'autre côté, est estompée, dès lors que l'on prend en considération la durée intrinsèque du phonème. Au vu de ce graphique, les durées des attaques et des noyaux sont similaires et sont allongées sur les deux dernières syllabes de l'unité intonative. Les segments des codas sont plus courts et répondent d'un allongement progressif, où l'effet est plus important sur les deux dernières syllabes de l'unité.

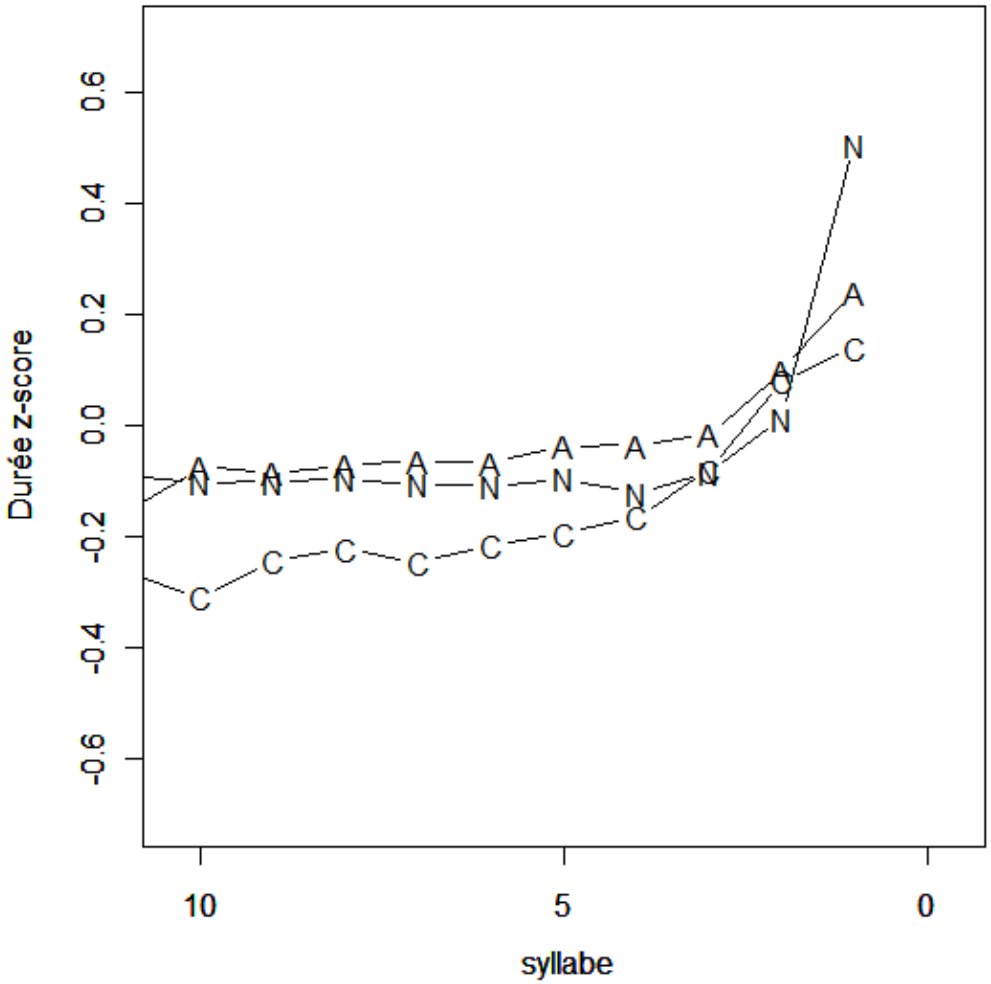


FIGURE 99 – Graphique des durées des phonèmes en fonction de leur position dans la syllabe et en fonction de la position de la syllabe dans l'unité intonative. A, N et C représentent les attaque, noyau et coda des syllabes. L'axe des abscisses représente la position de la syllabe dans l'unité intonative (plus la position est petite, plus la syllabe est proche de la frontière droite de l'unité), l'axe des ordonnées, la durée du phonème en z-score.

Si nous représentons, maintenant, les syllabes accentuées et les syllabes inaccentuées séparément [105] (cf. Figure 100), nous observons que certains effets sont des artefacts. Notamment,

105. Notons que nous « zoomons » la représentation graphique sur les sept dernières syllabes de l'unité intonative.

nous comprenons à présent que la similitude que nous avons pu observer entre les durées des noyaux et les durées des attaques est en fait artificielle et que l'impression graphique des attaques résultait de la moyennisation des durées des attaques des syllabes accentuées et inaccentuées. En réalité, on voit clairement à travers le graphique que les durées des attaques accentuées sont supérieures aux noyaux accentués et inaccentués et que ces derniers sont à leur tour supérieurs aux attaques des syllabes inaccentuées. Cette représentation graphique révèle ainsi que, dès lors que la durée intrinsèque du phonème est gommée, les noyaux des syllabes accentuées et inaccentuées sont très proches, que les attaques des syllabes accentuées sont les plus allongées, et que les codas accentués sont plus longs que les codas des syllabes inaccentuées. Quant au locus de l'allongement de frontière final, il semble toujours se confiner sur les deux dernières syllabes de l'unité intonative. Que la syllabe finale soit accentuée ou inaccentuée, l'allongement est plus important sur les noyau et attaque de la syllabe que sur le coda. Il est à noter que l'allongement du noyau de la syllabe accentuée est plus important que celui de la syllabe inaccentuée. Lorsque l'avant-dernière syllabe est accentuée, ce sont le noyau et le coda qui sont les plus allongés ; lorsqu'elle est inaccentuée, l'allongement porte aussi sur l'attaque.

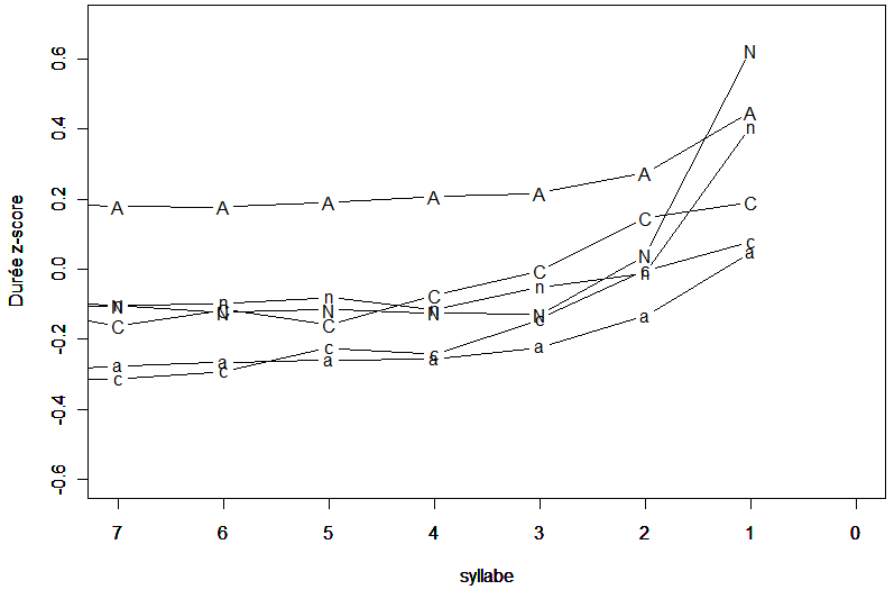


FIGURE 100 – Graphique des durées des phonèmes en fonction de leur position dans la syllabe (accentuée et inaccentuée) et en fonction de la position de la syllabe dans l'unité intonative. A, N et C représentent les attaque, noyau et coda des syllabes accentuées ; a, n et c, les attaque, noyau et coda des syllabes inaccentuées. L'axe des abscisses représente la position de la syllabe dans l'unité intonative (plus la position est petite, plus la syllabe est proche de la frontière droite de l'unité), l'axe des ordonnées, la durée du phonème en z-score.

Les analyses statistiques viennent préciser l'impression graphique du locus de frontière : l'ANOVA du modèle de régression à effets mixtes révèle que l'effet de frontière est significatif sur les deux dernières syllabes de l'unité intonative, lorsque la composition syllabique et le caractère accentuel ne sont pas spécifiés (pMCMC$<$0.01). L'allongement est plus important sur la dernière syllabe (0.18 vs. 0.04).

Lorsque la composition syllabique est prise en compte, l'ANOVA montre que l'effet de frontière est significatif sur les noyaux et les attaques des deux dernières syllabes, et sur les codas des trois dernières syllabes (pMCMC$<$0.01). L'allongement des noyaux de la dernière syllabe est plus important que ceux des attaques et des codas de la dernière syllabe (noyau=0.27, vs. attaque=0.07, vs. coda=0.1137). On note que l'allongement est plus important pour les codas que pour les attaques. Par ailleurs, plus la syllabe est proche de la frontière, plus l'allongement est important, aussi bien pour les noyaux, que pour les attaques et les codas (noyaux : 0.04, 0.27 ; attaques : 0.04, 0.07 ; codas : 0.07, 0.10, 0.11).

Lorsque on considère la composition syllabique et le caractère accentuel de la syllabe, l'ANOVA indique que les noyaux et les codas des deux dernières syllabes, lorsqu'elles sont accentuées, et les attaques des deux dernières syllabes, lorsqu'elles sont inaccentuées, sont significativement allongés (pMCMC<0.01). L'allongement des noyaux et des attaques est par ailleurs plus important dans la dernière syllabe que dans l'avant-dernière syllabe (N=0.34 vs. 0.07 ; a=0.09 vs. 0.04). Il est en revanche identique pour les codas des deux dernières syllabes lorsqu'elles sont accentuées(C=0.12). Ces résultats corroborent d'ailleurs ceux de White (2002) qui montre également un allongement du coda sur l'avant-dernière syllabe lorsqu'elle est accentuée, en anglais britannique. En revanche, le coda des syllabes accentuées en antépénultième position n'est pas touché par l'effet de frontière, dans notre corpus.

Le noyau des syllabes inaccentuées et l'attaque des syllabes accentuées sont affectés uniquement lorsque les syllabes sont en position finale d'unité (pMCMC<0.01). Dans ce cas l'allongement est plus important pour le noyau de la syllabe inaccentuée que pour l'attaque de la syllabe accentuée (0.22 vs. 0.08). Notre étude ne corrobore donc pas les observations d'Oller (1973) qui rapporte que seul le noyau et le coda des syllabes accentuées en fin de constituants sont affectés par l'effet de frontière.

Enfin, les codas des syllabes inaccentués sont allongés lorsque les syllabes se trouvent en position finale, avant-dernière et antépénultième (pMCMC<0.01). L'allongement du coda est plus important sur la dernière syllabe que sur l'avant-dernière ou l'antépénultième (0.12 vs. 0.09). Comme le pensait Oller (1973), lorsque la dernière syllabe est inaccentuée, ce sont l'attaque, le noyau est le coda qui sont affectés par l'allongement de frontière.

Autrement dit, le locus de l'allongement de frontière est de 3 syllabes uniquement pour les codas inaccentués, de 2 syllabes pour les noyaux et codas accentués et les attaques inaccentuées, et d'1 syllabe pour les noyaux inaccentués et les attaques accentuées. De plus, les attaque, noyau et coda de la dernière syllabe sont toujours affectés par l'effet de frontière, quel que soit le caractère accentuel de la syllabe (i.e. accentuée vs. inaccentuée).

En outre, l'allongement des noyaux des syllabes accentuées et inaccentuées se fait de façon quasi identique, avec un allongement très important sur la dernière syllabe (N = 0.34 et n = 0.22) et un allongement peu marqué ou inexistant sur l'avant dernière syllabe (N = 0.07). L'effet semble être un effet miroir, mais seulement là où les syllabes accentuées sont plus marquées par l'effet d'allongement que les syllabes inaccentuées. En revanche, l'allongement des attaques et des codas sur la dernière syllabe, qu'elle soit accentuée ou non, est similaire (A = 0.08, a = 0.09 ; C = 0.12, c = 0.12). De plus, l'allongement des codas des avant-dernières syllabes inaccentuées est à peine plus important que celui des codas des avant-dernières syllabes accentuées (c = 0.11 et C = 0.09).

6.4 Discussion

Nous avons pu démontrer, à partir d'un corpus de parole authentique, un effet de frontière finale au niveau de l'unité intonative en anglais britannique. Le locus de l'allongement au sein de ce domaine, en revanche, est difficile à déterminer, du fait de son interaction avec d'autres facteurs et de l'effet cumulé de ces différents facteurs. Nous avons pu voir, par exemple, que le locus est déterminé sur les deux dernières syllabes de l'unité intonative lorsque la composition et la position syllabique ne sont pas prises en compte, alors qu'il varie entre la dernière syllabe et les trois dernières syllabes de l'unité, dès lors qu'on les considère. Ces résultats nous mènent cependant à formuler l'hypothèse que le locus de l'allongement se trouve au niveau de la dernière unité rythmique de l'unité intonative, ou encore, se niche au sein d'un pied bissyllabique final. Il ne semble donc pas, au vu de ces premiers résultats, que l'effet de frontière est confiné au niveau de la rime de la dernière syllabe (Lindblom, 1968; T. Crystal & House, 1990; Wightman et al., 1992; White, 2002). Il semblerait plutôt qu'il ait un empan temporel plus large ou encore qu'il relève d'un double mécanisme (Turk & Shattuck-Hufnagel, 2007). Par ailleurs, ces résultats révèlent la possibilité d'un effet de rallentando global sur les codas des dernières syllabes, mais il faudrait voir si ce phénomène ne résulte pas de l'influence d'autres facteurs.

Nous avons pu aussi noter un allongement progressif des durées segmentales en fonction de la position de la syllabe dans l'unité intonative. Les attaques, les noyaux et les codas des dernières syllabes sont en effet plus allongés que ceux des avant-dernières syllabes, ce qui vient corroborer les observations de la littérature. En outre, comme Campbell et Isard (1991) et Wightman et al. (1992), nous trouvons pour la dernière syllabe un allongement plus important pour la rime que pour l'attaque. Cependant, on remarque que les parallèles effectués entre nos observations et celles de la littérature sont quelque peu artificiels, dans le sens où ces observations résultent du domaine observé (i.e. du niveau de constituance prosodique) et des facteurs interagissant considérés.

Il est clair que cette étude, en l'état, ne permet pas de déterminer avec certitude le locus de l'allongement de frontière final au sein de l'unité intonative. De nombreux facteurs, nous l'avons dit, sont à considérer, dans l'étude de ce phénomène, notamment le caractère accentuel des syllabes qui se succèdent. Pour preuve, si l'on étudie l'effet de frontière sur les trois dernières syllabes de l'unité intonative, 8 combinaisons sont possibles et doivent donc être étudiées séparément, dans le but de déterminer plus précise du locus de l'allongement de frontière : (1) S s S, (2) s s S, (3) s S s et (4) S s s, (5) s s s, (6) S S S, (7) S S s et (8) s S S où « S » représente une syllabe accentuée et « s » une syllabe inaccentuée. Ou encore, lorsque l'on prend en compte la composition syllabique, il serait intéressant d'y intégrer le nombre de phonèmes au sein des attaques et des codas.

Par ailleurs, nous avons pu montrer l'importance de prendre en compte la durée intrinsèque des phonèmes dans ces analyses, par une comparaison des représentations graphiques des durées normalisées en log et celles des durées normalisées en z-score. Nous avons pu notamment présenter « des cas d'école », où les impressions « à première vue » s'avéraient faussées par l'interaction et l'effet cumulé de divers facteurs. Nous avons montré que, dès lors que l'on neutralisait l'effet intrinsèque du phonème par une transformation z-score, les importantes différences de durée que l'on avait pu observer entre les noyaux accentués et les noyaux inaccentués, se révélaient en fait très petites. Ou encore, alors qu'à première vue les durées des noyaux accentués semblaient supérieures aux durées des attaques accentuées, la normalisation z-score a pu montrer qu'il n'en était rien et qu'en réalité, les durées des attaques accentuées étaient supérieures aux durées des noyaux accentués. Par ailleurs, la prise en compte du caractère accentuel de la syllabe a aussi montré la nécessité d'étudier les syllabes accentuées et inaccentuées séparément. Le fait de les avoir séparé a pu révéler que la similitude que nous avons observée entre les durées des noyaux et des attaques était en fait artificielle et qu'elle résultait tout simplement de la moyennisation des durées des attaques des syllabes accentuées et inaccentuées.

Cette étude préliminaire a pu souligner la complexité de l'étude de l'allongement de frontière et la nécessité d'un travail consacré à cet aspect. Elle nous permet cependant de formuler l'hypothèse que la détection automatique des variations de tempo pourrait être affinée par la neutralisation, en amont, des effets d'allongements de frontière au niveau des deux dernières syllabes de l'unité intonative.

6.5 Conclusion : une synthèse

Ce chapitre se penche sur la difficulté qui se pose à la modélisation de l'organisation temporelle. Cette modélisation est en effet complexe parce qu'elle résulte de l'influence de nombreux facteurs et de leur interaction. Dans la modélisation segmentale, l'un des facteurs à prendre en considération est le tempo. Ce chapitre est donc consacré à son étude.

Tout d'abord, nous avons montré, à travers la détection automatique des variations de tempo, que la détermination de leur empan temporel est complexe. En effet, nous avons pu observer, à partir d'une structure arborescente, que les variations du débit d'élocution opèrent sur différents domaines, par ailleurs emboîtés. Nous avons conclu que l'étude de l'empan temporel des variations de tempo ne peut s'effectuer sans considérer l'emboîtement de ces effets.

Une étude succincte des composantes du tempo, à savoir le débit d'élocution et les pauses, a montré que ces deux éléments ne sont pas toujours interdépendants. En effet, nous avons montré que la corrélation entre le débit d'élocution et la durée des pauses ainsi que celle entre

la durée des pauses et leur nombre sont faibles. D'autre part, la corrélation entre le débit d'élocution et le nombre de pauses n'est pas significative. La tendance pour varier le tempo ne résulterait donc pas d'une seule et même combinaison des composantes.

Notre étude porte aussi sur les fonctions extra-linguistiques et linguistiques qu'endossent les variations de tempo. Nous nous sommes notamment intéressée à la façon dont les variations de tempo traduisent le sexe du locuteur et le style de parole qu'il adopte. Notre étude montre qu'il n'existe pas de différences significatives de tempo entre hommes et femmes. En revanche, l'étude des différents types de production étudiés et extraits du corpus Aix-Marsec montrent qu'ils sont marqués par des tempos différents.

Par ailleurs, nous nous sommes penchée sur la façon dont les variations de tempo indiquent des changements de topique. Nous avons constaté, qu'en anglais et en français, la structure hiérarchique du discours, exprimée en termes de structure intentionnelle, est soulignée par des variations du débit d'élocution et de la durée des pauses. Nous avons tout particulièrement montré que l'unité qui introduit un nouveau topique est marquée par un changement de débit et est précédée d'une pause relativement longue.

A partir de ces résultats, nous avons envisagé la prédiction des changements de topiques en fonction des variations du débit d'élocution et de la durée des pauses. La prédiction des changements de topique, à partir de ces variations, semble difficile du fait que ces mêmes variations revêtent d'autres fonctions. Nous avons donc montré que, bien que la prédiction des topiques, en l'état, n'est pas satisfaisante, elle n'en demeure pas moins réalisable.

Conclusion

Cette thèse s'est donné pour objectif d'appréhender la problématique de l'empan temporel des variations prosodiques à long terme, i.e. les variations prosodiques de registre et de tempo, et d'examiner les fonctions qu'elles revêtent, notamment la façon dont elles renseignent sur l'identité du locuteur ou encore la façon dont elles indiquent la structure intentionnelle du discours. Nous avons en effet soulevé que l'une des problématiques majeures de l'étude des variations prosodiques était celle de leur délimitation, à plus ou moins long terme, et par là, de leur interaction et chevauchement, qui rend leur analyse séparée difficile.

L'étude de l'empan temporel des variations de registre et de tempo et des fonctions qu'elles revêtent soulève un certain nombre de problématiques auxquelles nous avons cherché à répondre, comme celle de la définition des termes d'empan temporel, de registre ou encore celle de la mesure du registre, du tempo et de leurs variations. De telles problématiques ont d'ailleurs nécessité l'élaboration d'outils automatiques, et ont débouché sur des apports méthodologiques dans l'étude des variations prosodiques. Par ailleurs, ce travail s'est basé sur des corpus de parole authentique (dont le but est de « communiquer du sens »), ce qui nécessite une approche complexe des variations prosodiques. En outre, l'étude repose sur un nombre important de données (quatre corpus : deux corpus en anglais et deux en français), dont la richesse a permis des études comparatives (de l'anglais et du français ; de la parole lue, semi-spontanée ou authentique).

Nous tenons à présenter, dans une première partie, les avancées majeures de ce travail et les apports méthodologiques sous-jacents à ces avancées, ce qui nous amène à envisager, dans une seconde partie, d'autres perspectives de recherche.

1 Avancées et apports méthodologiques

1.1 A propos des définitions

La toute première question à laquelle nous nous sommes attelée est celle de la définition du terme d'empan temporel. Inspirée des travaux de White (2002), cette définition met en exergue deux termes que nous avons pensé nécessaire de distinguer : celui de l'*empan temporel* (ou *domaine*) que nous définissons comme un intervalle de temps (ou une unité de temps), à partir duquel opère un phénomène prosodique et celui de *locus* que l'on définit comme la localisation exacte des effets de ce phénomène au sein du domaine. Nous avons pensé en effet qu'une telle distinction est nécessaire pour décrire au mieux les faits prosodiques. Prenons l'exemple du registre et de ses variations. Nous avons trouvé dans la littérature de nombreux termes qui se réfèrent à cette notion : déclinaison, downstep, downdrift, remise à niveau, ou encore variations d'étendue et de hauteur globale. On distingue souvent ces termes par les termes de local et global. Or, ces termes se réfèrent-ils au terme de domaine ou de locus ? Lorsqu'on parle du downstep en tant que changement local, cela n'indique pas que le domaine à partir duquel il opère peut être assez large (notamment dès lors que l'on le décrit en termes de downstep emboité). En outre, lorsque l'on traite de remise à niveau de façon très locale (de l'ordre de la syllabe), cela n'indique pas que cette marque permet la délimitation d'un domaine plus large. Utiliser les termes de locus et de domaine devient donc nécessaire. Dans le cas de remise à niveau par exemple, le locus des variations de registre serait la syllabe et le domaine pourrait être l'énoncé. On parlerait alors de locus étroit. Inversement, si un locuteur élève sa voix sur l'ensemble d'un énoncé (pour se faire entendre de loin par exemple), le domaine du registre serait l'énoncé et comme l'ensemble des cibles tonales seraient affectées par cette élévation de la voix, on parlerait de locus large, où le locus serait finalement de même taille que le domaine. Cette distinction est aussi intéressante, à notre avis, à travers celle de la distinction d'un empan de réalisation et d'un empan d'interprétation. La remise à niveau serait réalisée sur une syllabe mais son interprétation permettrait la délimitation d'un domaine plus large.

Une autre difficulté à laquelle nous nous sommes attachée est celle de la définition du terme de registre. Comme nous l'avons soulevé plus haut, le terme de registre englobe un certain nombre de phénomènes décrits sous différents termes. Nous avons montré que les auteurs distinguent généralement les effets de déclinaison, de downdrift et de downstep des effets de variations de hauteur et d'étendue du registre. Or, ils décrivent ces effets de façon identique, i.e. comme la modification de l'espace tonal dans lequel s'échelonnent les cibles tonales. Nous avons donc proposé de regouper l'ensemble de ces phénomènes sous le terme de « registre » en considérant toutefois la distinction que l'on fait de ces phénomènes par l'utilisation des termes de « variations déclinantes » et de « variations verticales ». Cette définition permet d'ôter l'idée de considérer que les variations verticales n'opèrent que sur des domaines très

larges : on peut élever sa voix sur l'ensemble d'un énoncé comme sur un seul mot.

1.2 A propos des mesures de registre et de tempo

Un autre point de difficulté se pose aux variations de registre et de tempo (et par ailleurs à l'allongement de frontière), c'est celui de leur mesure.

La mesure du registre repose en effet sur une détection de la fréquence fondamentale parfois erronée. Nous avons expliqué que si les algorithmes de détection de la fréquence fondamentale sont aujourd'hui robustes, ils ne présentent pas moins des erreurs de détection et de calcul. Nous avons donc proposé, pour améliorer l'analyse de la f_0, de passer par un ajustement de ses paramètres. Notre étude est appliquée au logiciel Praat que de nombreux linguistes utilisent à des fins diverses. Nous avons montré, par exemple, que dans ce logiciel, la fréquence fondamentale était aussi sujet d'erreurs de détection. Mais nous avons prouvé que l'ajustement des seuils plancher et plafond, comme d'ailleurs le recommandent les auteurs du logiciel, permet une analyse automatique plus fiable de la f_0, où les erreurs aberrantes aux extrêmes de la courbe sont écartées grâce à ces ajustements. Nous avons conclu qu'ajuster les seuils plancher et plafond aux produits des percentiles 15 et 65 et de leurs coefficients respectifs 0.83 et 1.92 rend fiable l'estimation des extrema de la f_0. Il n'est donc plus nécessaire d'ajuster les seuils manuellement. Par ailleurs, la forte corrélation que nous avons trouvée entre la médiane et la moyenne des cibles basses et entre la médiane et la moyenne des cibles hautes, nous a amené à proposer d'autres ajustements des seuils : à l'octave supérieure par rapport à la médiane pour le seuil plafond et à la demi-octave inférieure par rapport à la médiane pour le seuil plancher.

C'est cette forte corrélation qui reste d'ailleurs un des résultats majeurs de ce travail. Elle implique de nombreuses conclusions :

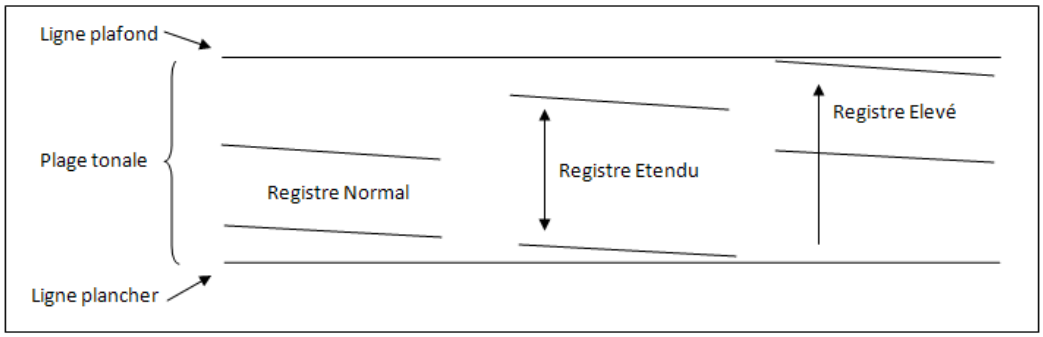
(1) Elle prouve la co-variance entre la hauteur du registre et son étendue. En effet, nous avons pu conclure que plus le registre est haut, plus il est étendu, et vice-versa, plus le registre est bas, plus il est étroit. Cette étude expérimentale vient donc confirmer ce que Ladd (1996, p260) dit dans sa définition du registre, i.e. que la difficulté d'admettre deux dimensions au registre vient du fait que ces deux dimensions co-varient.

(2) Cette corrélation nous a aussi permis de proposer une mesure de la hauteur et de l'étendue du registre. Nous avons ainsi déterminé que la médiane est une bonne mesure de la hauteur du registre, notamment du fait qu'elle est non paramétrique, et que, par conséquent, elle est identique quelle que soit l'échelle de mesure utilisée. Nous avons aussi proposé une mesure de l'étendue du registre à partir de la médiane, par la formule suivante : $E = 0.161 \times log_2(MED)$.

(3) De plus, cette corrélation nous mène à envisager une nouvelle échelle de mesure. Si, comme

le montre notre expérimentation, le registre d'un locuteur est délimité par l'octave supérieure et la demi-octave inférieure par rapport à la hauteur globale de son registre, nous pouvons envisager de proposer une échelle de mesure naturelle des variations de la f_0, où la courbe serait centrée sur une ligne correspondant à la médiane et ses variations délimitées par une ligne plancher correspondante à la demi-octave inférieure par rapport à la médiane et une ligne plafond correspondante à l'octave supérieure par rapport à la médiane. Les variations de registre seraient ainsi plus clairement appréhendées au sein de cette plage tonale.

(4) Enfin, elle justifie aussi notre représentation graphique du registre. Nous avions proposé dans l'introduction de ce travail une schématisation des modifications de registre, que nous confirmons à nouveau ici.



Nous avions proposé en effet que les lignes plancher et plafond, délimitant la plage tonale, restent constantes et servent de limites aux variations verticales et que les lignes du registre, elles, délimitent l'espace tonal à partir duquel s'échelonne les cibles tonales. Contrairement à Ladd (1992), les lignes plancher et plafond délimitant la plage tonale (*range* chez l'auteur), dans notre conception, sont constantes, elles ne s'élèvent pas ou ne s'abaissent pas à des fins para-linguistiques mais ce sont les lignes du registre qui sont utilisées dans ce but, puisque les variations déclinantes et verticales sont toutes deux des modifications de l'espace tonal. La proposition de limites constantes est validée expérimentalement dans cette étude, par laquelle on prouve que les variations de registre se trouvent limitées à l'octave supérieure et la demi-octave inférieure par rapport à la hauteur de voix du locuteur.

Par ailleurs, nous avions souligné la dichotomie entre mesures acoustiques et mesures linguistiques avancée par Patterson (2000). L'auteur suggère en effet dans sa thèse que les mesures linguistiques (exprimées en termes de mesures extraites de points d'inflexion de la f_0) sont préférables aux mesures acoustiques (exprimées en termes de mesures statistiques effectuées à partir de la f_0) dans l'étude du registre pour plusieurs raisons. L'auteur avance notamment la difficulté d'une mesure fiable du registre lorsqu'elle est effectuée automatiquement : une mesure automatique à partir de la f_0 prendrait en compte des valeurs aberrantes, causées

elles-mêmes par une détection et un calcul erronés de la f_0. Or, nous avons pu montrer qu'une mesure *contrôlée* du registre, notamment par l'ajustement des paramètres de la f_0, est aussi efficace qu'une mesure dite linguistique. L'auteur avance aussi comme argument que mesurer la hauteur du registre (à partir de la distribution de la f_0 autour de la moyenne) et mesurer l'étendue du registre (comme la distance entre la valeur maximale et minimale) ne peut être efficace du fait que la distribution de la f_0 ne suit pas une loi normale. Nous avons donc suggéré qu'il suffit d'utiliser la médiane, plus stable que la moyenne aux valeurs aberrantes, et qui permet la division d'un ensemble de données en deux sous-ensembles égaux. Quant à l'étendue, nous avons proposé une transformation des données en base de log 2, qui permet, comme nous avons pu le montrer, de rendre la distribution plus proche d'une distribution normale centrée. Enfin, l'auteur suggère que les mesures linguistiques sont plus corrélées à la perception de l'auditeur que ne le sont les mesures acoustiques. Pour notre part, nous n'avons pas mené d'étude perceptive mais nous avons montré que les résultats obtenus à partir de ces deux types de mesures sont très similaires. Nous pensons donc qu'une étude perceptive du registre par la comparaison de mesures acoustiques contrôlées et de mesures linguistiques devrait confirmer le fait que ces mesures s'équivalent. Cette dichotomie semble donc artificielle.

En revanche, nous avons montré que la mesure du tempo posait moins de difficultés que celle du registre, certainement du fait que sa définition est plus établie, et que les auteurs s'entendent assez bien aujourd'hui sur le type de mesure à utiliser (e.g. le phonème). Nous avons cependant montré que les composantes du tempo ne sont pas systématiquement corrélées comme peut le laisser penser la littérature. Nous avons en effet observé la relation entre le débit d'élocution et la durée des pauses et avons montré que, même si elle paraît linéaire, elle n'est que très faible. Par ailleurs, la relation linéaire entre la durée des pauses et le nombre de pauses, même si elle existe, semble aussi très faible. En revanche, le nombre de pauses et le débit ne sont pas corrélés. Nous avons donc conclu que la tendance, pour ralentir le tempo, serait de ralentir le débit d'élocution et d'augmenter la durée des pauses ou encore d'augmenter la durée des pauses et de réduire le nombre de pauses ; inversement, pour accélérer le tempo, le débit serait accéléré, la durée des pauses diminuée et le nombre de pauses courtes plus nombreuses. Mais cette tendance ne fait pas loi, puisque les stratégies divergent en fonction des intentions du locuteur ou encore du style de parole qu'il adopte. Nous avons donc conclu que les composantes du tempo ne sont pas interdépendantes. Cela permettrait d'expliquer les divergences qu'on trouve dans la littérature à propos des stratégies utilisées par les locuteurs pour varier leur tempo.

Nous avons aussi soulevé un point important dans l'étude de l'allongement de frontière, celui de sa mesure. Il y a quelques années, Campbell (1992) proposait une transformation z-score afin de réduire l'influence de la durée intrinsèque des phonèmes ou du débit du locuteur. Aujourd'hui, nous disposons d'outils qui permettent de mener des analyses statistiques poussées

et de prendre en compte ces effets de façon plus fiable et plus précise. Nous avons notamment proposé l'utilisation de modèles mixtes, où les facteurs aléatoires permettent de gommer l'effet du phonème et l'effet du débit. Ainsi, la transformation z-score semble assez drastique et, pour notre part, bien que nous pensions qu'elle est essentielle (notamment dans la représentation graphique des données, afin d'appréhender le comportement d'un phénomène tel que l'allongement de frontière), nous pensons que l'utilisation de modèles mixtes rend l'analyse plus fine.

1.3 A propos de l'empan temporel

Si nous avons pu apporté des solutions aux problématiques de mesure du registre et du tempo, en revanche, la problématique de l'empan temporel de leurs variations est loin d'être résolue. Nous avons expliqué que la difficulté d'une telle problématique reposait sur le fait qu'il est difficile de présupposer ces variations du fait qu'elles dépendent de la structure hiérarchique, organisationnelle et informationnelle du discours et des intentions du locuteur. Il est en effet difficile de prédire *le* domaine pour lequel variera le registre. Il est donc difficile d'établir le domaine à partir duquel elles opèrent ainsi que le locus de leur effet. Cette thèse n'aura donc pas donné la possibilité d'asseoir un domaine pour lequel le registre et le tempo varient. Nous ne pourrons donc trancher entre les divers domaines proposés dans la littérature, tels que le syntagme, la proposition, la phrase, le groupe de souffle, l'unité intonative, l'énoncé ou encore le paraton, qu'ils soient définis selon des critères syntaxiques, prosodiques ou phonologiques. L'analyse de la parole authentique ou semi-authentique est en effet très complexe, à cause de l'interaction de nombreux facteurs.

Toutefois, nous pensons que l'analyse de la parole authentique peut s'avérer plus pertinente si nous ne considérons pas qu'il existe *un* domaine pour lequel le registre et le débit varient mais *des* domaines. En effet, du fait des nombreuses fonctions que revêtent les variations de registre et de débit, nous ne pouvons envisager qu'elles opèrent sur un seul domaine constant. Nous avons par ailleurs montré une amélioration du codage des cibles tonales dans le système d'intonation INTSINT par l'intégration des variations de registre définies en fonction de seuils. Ces seuils ne représentent pas un domaine en particulier, mais délimitent des domaines de tailles variables.

Par ailleurs, nous défendons l'idée que ces domaines sont emboités, nous l'avons d'ailleurs montré à partir de structures arborescentes. Par exemple, nous avons pu observer un abaissement du registre sur plusieurs niveaux d'emboitement. Nous soutenons ainsi la notion de downstep emboité ou de *wheels within wheels model of downstep* proposée par Ladd (1988) et Van Den Berg et al. (1992). Nous pensons par ailleurs qu'une telle conception, outre l'enjeu théorique qu'elle soulève, devrait être aussi appliquée dans la parole de synthèse. Nous pensons

en effet qu'un module d'implémentation des variations de registre et de débit d'élocution à plusieurs niveaux permettrait une parole synthétisée plus naturelle.

Nous avons également soulevé un point important dans la détermination de l'empan temporel des variations de débit d'élocution. Elle ne peut être envisagée sans la prise en compte de l'empan temporel de l'allongement de frontière. Nous avons en effet expliqué que les deux effets étaient très liés, l'allongement de frontière étant parfois décrit comme une réduction localisée du débit d'élocution. Nous ne pouvons donc certifier que la détection automatique des variations de débit d'élocution que nous avons proposée, en l'état, n'est pas aussi celle des effets de frontière. Nous avons proposé une méthode pour séparer ces deux phénomènes, à savoir de déterminer le domaine à partir duquel ils opèrent et le locus de leur effet. Notre étude qui porte sur le locus de l'allongement au niveau de l'unité intonative en anglais britannique est d'ailleurs prometteuse. Nos résultats, issus d'un corpus de parole authentique, montrent que le locus de l'allongement de frontière se trouve au niveau des deux dernière syllabes de l'unité intonative. En revanche, il est difficile d'assurer que le locus de l'allongement se fait uniquement sur les deux dernières syllabes du fait de son interaction avec d'autres facteurs et de leur effet cumulé. Nous avons pu voir, par exemple, que le locus est déterminé sur les deux dernières syllabes de l'unité intonative lorsque la composition et la position syllabique sont ignorées, alors qu'il varie entre la dernière syllabe et les trois dernières syllabes de l'unité, dès lors qu'on les prend en compte. Nos résultats amènent à penser que l'effet de frontière n'est pas uniquement confiné au niveau de la rime de la dernière syllabe (Lindblom, 1968; T. Crystal & House, 1990; Wightman et al., 1992; White, 2002) et que son locus pourrait être plus large ou encore relever d'un double mécanisme (Turk & Shattuck-Hufnagel, 2007). De plus, ces résultats soulignent la possibilité d'un effet de rallentando global sur les codas des dernières syllabes, mais il faudrait toutefois s'assurer que ce phénomène ne résulte pas de l'influence d'autres facteurs. Cette étude montre enfin un allongement progressif des durées segmentales en fonction de la position de la syllabe dans l'unité intonative. Cette étude préliminaire a ainsi diagnostiqué la complexité de l'étude de l'allongement de frontière et la nécessité d'un travail consacré à cet aspect. Elle nous permet cependant de formuler l'hypothèse que la détection automatique des variations de tempo pourrait être affinée par la neutralisation, en amont, des effets d'allongements de frontière, au niveau des deux dernières syllabes de l'unité intonative.

1.4 A propos des fonctions prosodiques

Outre la problématique de l'empan temporel, nous nous sommes aussi intéressée dans cette étude aux fonctions extra-linguistiques et linguistiques que revêtent les variations de registre et de tempo, notamment la façon dont elles révèlent d'identité d'un locuteur, à savoir son sexe, son origine géographique et le style de parole qu'il adopte ou encore la structure intentionnelle

du discours. Nous avons pu montrer que le registre n'est pas influencé par la langue parlée ou encore le type de parole pratiqué. Les différences de registre semblent donc être uniquement dues aux différences biologiques entre individus. L'étude du tempo ne révèle elle non plus aucun effet du sexe ou de la langue. Il semblerait en revanche, au vu des premières analyses, que les différences de tempo permettent de distinguer des styles de parole et de déterminer trois grands groupes de production (au moins à partir de ce corpus) : (1) les fictions et les émissions religieuses, (2) les paroles publiques et les propagandes, (3) les dialogues, les reportages, les commentaires et les informations. Il est en revanche nécessaire d'approfondir ces premiers résultats par des analyses plus poussées et des données plus contrôlées.

Nous nous sommes également intéressée à la façon dont les variations de registre et de tempo marquent la structure hiérarchique du discours, en termes de structure intentionnelle. Nous avons pu montrer que les changements de topique, en français et en anglais, en lecture oralisée et en parole authentique et conversationnelle, sont indiqués par des variations de registre : une unité introduisant un nouveau topique est en effet caractérisée par un registre plus haut et plus étendu que celle qui la précède, une différence qui s'avère être importante. Par contre, lorsque deux unités consécutives partagent un même topique, elles ont un registre équivalent. Les changements de topique sont aussi marqués par des variations de tempo. Notre étude, menée ici uniquement sur les corpus de parole authentique, révèle que, en anglais comme en français, l'unité qui introduit un nouveau topique est marquée par un changement de débit d'élocution par rapport à l'unité qui la précède et par des pauses plus longues. En revanche, pour l'anglais, l'unité qui introduit un nouveau topique est toujours marquée par un débit d'élocution plus lent, ce qui n'est pas le cas en français. Cette étude montre donc, qu'en anglais et en français, la structure intentionnelle du discours, exprimée dans cette étude à partir de 3 niveaux, est transcrite par des variations du registre, du débit d'élocution et de la durée des pauses. Plus on se situe haut dans la structure hiérarchique, plus les variations sont importantes. Nous avons aussi montré qu'à partir de ces variations, la prédiction et en conséquence l'annotation automatique des changements de topique pouvaient être envisagées, même si, en l'état, les résultats ne sont pas encore pleinement satisfaisants. Ils révèlent en tous cas et encore une fois, la complexité de l'étude de la parole authentique.

1.5 Elaboration d'outils de détection automatique des variations de registre et de tempo

Ce travail a nécessité l'élaboration d'outils qui permettent la détection automatique des variations de registre et de tempo. Les algorithmes de regroupement hiérarchique ADoReVA et ADoTeVA créés à cet effet, permettent ainsi la détection des variations de registre et de tempo et la représentation de ces variations sous forme de structure arborescente. Leur intérêt réside

dans le fait qu'ils permettent, en amont de toute analyse, d'appréhender ces variations graphiquement. De surcroît, ils permettent d'analyser les fonctions des variations de registre et de tempo en donnant la possibilité à l'utilisateur d'intégrer ses annotations et de les regrouper aux détections des variations sous forme de table de données. Enfin, ils ont été conçus pour être implémentés dans Praat, sous forme de plugin. L'avantage de ces algorithmes repose aussi sur le fait qu'ils sont facilement accessibles (i.e. téléchargeables gratuitement).

Nous pouvons dès à présent affirmer que ces algorithmes pourront être améliorés de diverses façons et ainsi envisager les perspectives de ce travail.

2 Perspectives

En abordant la problématique de l'empan temporel des variations prosodiques, nous avons touché directement à la problématique très complexe de la constituance prosodique et des niveaux qu'on lui reconnaît. En effet, les notions d'emboîtement de domaines et la représentation de structures emboîtées touchent directement à cette étude. Pour autant, nous n'avons pas étudié les variations de registre et de tempo à travers l'étude des constituants phonologiques ou prosodiques d'une structure prosodique hiérarchique en particulier. Il serait cependant intéressant de prendre en compte les théories phonologiques sous-jacentes à cette problématique. En effet, l'interaction entre les niveaux phonétiques et phonologiques, par l'apport de critères phonologiques d'une part et par la détection automatique des paramètres acoustiques d'autre part, permettrait d'analyser plus finement les variations de registre et de tempo (e.g. régies par des règles phonologiques), mais aussi de mieux appréhender et définir, par la détection objective de ces variations, la structure prosodique ou phonologique sous-jacente. A l'avenir, on pourrait ainsi mieux analyser les variations prosodiques à court et à long terme malgré leur chevauchement et leur interaction. Il serait d'ailleurs intéressant d'intégrer dans cette analyse un troisième paramètre acoustique : l'intensité. En effet, à notre connaissance, peu d'études ont porté sur les variations de l'intensité et leur influence sur les faits prosodiques à plus court terme, ou encore sur la façon dont ces variations peuvent indiquer la structure hiérarchique d'un énoncé.

Outre l'interaction des niveaux phonétiques et phonologiques, ce travail mériterait d'être complété par une étude perceptive. Par exemple, nous avons proposé une mesure de l'étendue du registre basée sur la différence entre la moyenne des cibles basses et la moyenne des cibles hautes. Nous avons par conséquent mis de côté l'éventuelle mesure à partir d'un point maximal et d'un point minimal. Or, il serait opportun de vérifier si un auditeur juge de l'étendue du registre à partir d'une moyennisation des cibles hautes et basses ou à partir de deux points extrêmes. D'autre part, il serait bien de valider perceptivement l'utilisation de mesures

acoustiques, cette fois contrôlées, afin de voir si, comme les mesures linguistiques, elles sont corrélées au jugement perceptif de l'auditeur. Enfin, une étude perceptive devrait être appliquée à l'étude de la structure arborescente et de l'emboîtement de ces niveaux. En effet, il serait intéressant de vérifier si le seuil que nous avons proposé pour la détection des variations de registre, et qui s'est montré efficace dans le codage des cibles tonales, est aussi un « seuil perceptif ».

L'étude que nous avons menée sur l'allongement de frontière nous permet d'envisager d'autres perspectives de travail. Nous pensons qu'outre la composition syllabique et le caractère accentuel de la syllabe, il serait intéressant d'intégrer d'autres facteurs qui interfèrent dans l'analyse de l'allongement de frontière. Il serait notamment intéressant de prendre en considération la structure accentuelle des syllabes qui avoisinent la dernière ou l'avant-dernière syllabe. En effet, le caractère accentuel des syllabes qui précèdent les syllabes allongées interagit très certainement avec le degré d'allongement de la syllabe et peut ainsi participer à déplacer le locus de l'effet de frontière sous certaines conditions. Si, par exemple, on étudie l'effet de frontière sur les trois dernières syllabes de l'unité intonative, 4 structures accentuelles sont possibles : (1) S s S, (2) s s S, (3) s S s et (4) S s s, où « S » représente une syllabe accentuée et « s » une syllabe inaccentuée. Il semble donc tout à fait envisageable que le locus de l'effet soit quelque peu modifié selon la structure. Par ailleurs, la prise en compte des éléments au sein de l'attaque et du coda pourrait aussi permettre une analyse plus fine de cet effet et de son locus. Enfin, nous pourrions pousser l'étude à d'autres niveaux prosodiques pour lesquels on constate aussi des effets d'allongement de frontière.

Enfin, les analyses que nous avons effectuées à partir de la fréquence fondamentale et les résultats obtenus ouvrent la voie à un projet de recherche. En effet, il serait intéressant d'étudier plus en profondeur la distribution de la fréquence fondamentale et ainsi pousser plus loin l'analyse du registre. Outre une mesure plus fine du registre, une analyse de la fréquence fondamentale et de ses variations plus poussée pourrait nous permettre d'établir avec certitude les premiers résultats de ce travail, à savoir que le registre est délimité par l'octave supérieure et la demi-octave inférieure par rapport à la médiane. Cette avancée nous mène à proposer une échelle de mesure naturelle de la voix, où les variations de la fréquence fondamentale, centrées sur la médiane seraient délimitées par l'octave supérieure et la demi-octave inférieure. Si on obtient confirmation de ces limites, on pourrait envisager ce phénomène comme universel et mécanique et on pourrait penser que l'auditeur devine les limites du registre d'un locuteur à partir d'une seule valeur médiane. Cela pourrait expliquer le fait que malgré des patrons intonatifs descendants, les auditeurs sont capables de juger du caractère terminal ou non de ce patron.

Liste des tableaux

Table des figures

Références

Abercrombie, D. (1964). Syllable quantity and enclitics in English. *In Honour of Daniel Jones*, 216–222.

Abercrombie, D. (1967). *Elements of General Phonetics.* Aldine Pub. Co.

Adams, S., Weismer, G., & Kent, R. (1993). Speaking rate and speech movement velocity profiles. *Journal of Speech, Language and Hearing Research*, *36*(1), 41-54.

Aguilar, C., Bonafonte, & Escudero. (2009). Determining intonational boundaries from the acoustic signal. *Interspeech 2009*.

Apple, W., Streeter, L., & Krauss, R. (1979). Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology*, *37*(5), 715–727.

Arons, B. (1994). Pitch-based emphasis detection for segmenting speech recordings. *Third International Conference on Spoken Language Processing*, 1931–1934.

Arvaniti, A. (1999). Effects of speaking rate on the timing of single and geminate sonorants. *Proceedings of the 14th International Congress of Phonetic Sciences*, *1*, 599–602.

Arvaniti, A., & Garding, G. (2007). Dialectal variation in the rising accents of American English. *Laboratory phonology*, *9*, 547–576.

Auran, C., & Bouzon, C. (2003). Phonotactique prédictive et alignement automatique : apports et perspectives pour le traitement de grands corpus oraux. *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA), Aix-en-Provence, France*, 33–63.

Auran, C., Bouzon, C., & Hirst, D. (2004). The Aix-MARSEC Project : An Evolutive Database of Spoken British English. *Speech Prosody 2004, International Conference, March 23-26*

2004, Nara..

Ayers, G. (1994). Discourse functions of pitch range in spontaneous and read speech. *Working papers in linguistics(Columbus, Ohio)*, *44*, 1–49.

Baayen, R. (2008). *Analyzing linguistic data : A practical introduction to statistics using R*. Cambridge Univ Pr.

Banse, R., & Scherer, K. (1996). Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, *70*, 614–636.

Banziger, T., & Scherer, K. (2005). The role of intonation in emotional expressions. *Speech communication*, *46*(3-4), 252–267.

Barry, W. (1991). Phonetics and phonology of speaking styles. *Proc. 13th Intern Conf Phonetic Sciences*, *2*, 4–10.

Bartkova, K. (1985). Nouvelle approche dans le modèle de prédiction de la durée segmentale. *14ème JEP*.

Bartkova, K. (1991). Speaking rate in French application to speech synthesis. *XIIème Congrès International des Sciences Phonétiques*, 482–485.

Batliner, A., Kießling, A., Kompe, R., Niemann, H., & Nöth, E. (1997). Tempo and its change in spontaneous speech. *Fifth European Conference on Speech Communication and Technology*.

Beckman, M., & Edwards, J. (1990). Lengthenings and shortenings and the nature of prosodic constituency. *Between the grammar and physics of speech : Papers in laboratory phonology I*, 152–178.

Beckman, M., & Pierrehumbert, J. (1986). Intonational structure in Japanese and English. *Phonology yearbook*, 255–309.

Beinum, F. Koopmans-van, & Donzel, M. van. (1996). Relationship between discourse structure and dynamic speech rate. *Proceedings of the Fourth International Conference on Spoken Language - ICSLP 96*, *3*.

Beller, G., Schwarz, D., Hueber, T., & Rodet, X. (2006). Speech rates in french expressive speech. *Speech Prosody*.

Berg, R. van den, Gussenhoven, C., & Rietveld, T. (1992). Downstep in Dutch : Implications for a model. *Papers in laboratory phonology II : Gesture, segment, prosody*, 335–359.

Berkovits, R. (1991). The effect of speaking rate on evidence for utterance-final lengthening. *Phonetica*, *48*, 57–66.

Berkovits, R. (1993). Utterance-final lengthening and the duration of final-stop closures. *Journal of Phonetics*, *21*(4), 479–490.

Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., et al. (2008). Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle. *TAL Vol. 49 (3) Phonétique et Phonologie*.

Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., et al. (2007). Le CID-Corpus of Interactional Data- : protocoles, conventions, annotations.

Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix en Provence (TIPA), *25*, 25–55.

Bezooijen, R. van. (1984). *Characteristics and recognizability of vocal expressions of emotion.* Dordrecht : Foris. 164p.

Bhatt, P., & Léon, P. (1991). Melodic Patterns in Three Types of Radio Discourse. *Phonetics and Phonology of Speaking Styles*, 11.1–11.5.

Biersack, S., & Kempe, V. (2005). Tracing vocal emotion expression through the speech chain : do listeners perceive what speakers feel. *Proc. ISCA Workshop Plast. in Speech Perception*, 211–214.

Binnenpoorte, D., Bael, C., Os, E., & Boves, L. (2005). Gender in everyday speech and language : A corpus-based study. *Ninth European Conference on Speech Communication and Technology*.

Blaauw, E. (1991). Phonetic characteristics of spontaneous and read-aloud speech. *Phonetics and Phonology of Speaking Styles*.

Boersma, P., & Weenink, D. (2009). Praat : doing phonetics by computer [computer program]. *Retrieved from http ://www. praat. org*.

Bolinger, D. (1951). Intonation : Levels vs Configurations. *Word*, *7*, 199–210.

Bolinger, D. (1972). Around the edge of language : Intonation. *Harvard Educational Review, 34 (1964). Reprinted in Bolinger, Intonation : Selected Readings*, 282–296.

Bonnet, C. (1986). *Manuel pratique de psychophysique - Chapitre 4.* A. Colin.

Bouzon, C., & Hirst, D. (2002). The influence of prosodic factors on the duration of words in British English. *n Bel, B. & Marlien, I. (eds) : Proceedings from Speech Prosody 2002, Aix-en-Provence*, 191–194.

Bouzon, C., & Hirst, D. (2004). Isochrony and prosodic structure in British English. *Speech Prosody 2004, March 23-26 2004, Nara*.

Braun, A., & Kunzel, H. (2003). The effect of Alcohol on Speech Prosody. *Proceedings of the International Congress of Phonetic Sciences, Barcelona*, 2645–2648.

Brazil, D. (1980). Discourse Intonation and Language Teaching. , 205p.

Breitenstein, C., Van Lancker, D., & Daum, I. (2001). The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample. *Cognition and Emotion*, *15*(1), 57–79.

Brøndsted, T. (1997). Intonation Contours distorted by Tone Patterns of Stress Groups and Word Accent. *Intonation : Theory, Models and Applications, Athens (Athanasopoulos)*, 55–58.

Browman, C., & Goldstein, L. (1992). Articulatory phonology. *Phonetica*, *49*, 155–180.

Brown, B., Strong, W., & Rencher, A. (1973). Perceptions of personality from speech : Effects of manipulations of acoustical parameters. *The Journal of the Acoustical Society of America*, *54*, 29–35.

Brown, B., Strong, W., & Rencher, A. (1974). Fifty-four voices from two : the effects of

simultaneous manipulations of rate, mean fundamental frequency, and variance of fundamental frequency on ratings of personality from speech. *The Journal of the Acoustical Society of America*, *55–61*, 313.

Brown, G. (1983). Prosodic structure and the given/new distinction. *Prosody : Models and measurements*, 67–77.

Brown, G., Currie, K., & Kenworthy, J. (1980). *Questions of intonation.* Coll. Croom Helm Linguistics Series. Londres, Grande Bretagne : Croom Helm Ltd. 206 p.

Brubaker, R. (1972). Rate and pause characteristics of oral reading. *Journal of Psycholinguistic Research*, *1*(2), 141–147.

Bruce, G. (1977). *Swedish Word Accents in Sentence Perspective.* Travaux de l'institut de linguistique de Lund. 155 p.

Bruce, G. (1982). Textual aspects of prosody in Swedish. *Phonetica Basel*, *39*(4-5), 274–287.

Bruce, G. (1995). Modelling Swedish intonation for read and spontaneous speech. *Proceedings of International Congress on Phonetic Sciences*, *2*, 28–35.

Bruce, G., & Gårding, E. (1978). A prosodic typology for Swedish dialects. *Nordic prosody*, 219–228.

Burkhardt, F., Audibert, N., Malatesta, L., Turk, O., Arslan, L., & Auberge, V. (2006). Emotional Prosody-Does Culture Make A Difference ? *Speech Prosody 2006*, 2–5.

Butcher, A. (1981). *Aspects of the Speech Pause : Phonetic Correlates and Communicative Functions.* Institut für Phonetik der Universität Kiel.

Butterworth, B. (1975). Hesitation and semantic planning in speech. *Journal of Psycholinguistic Research*, *4*(1), 75–87.

Butterworth, B. (1980). Evidence from pauses in speech. *Language production*, *1*, 155–176.

Byrd, D. (1992). Sex, dialects, and reduction. *Second International Conference on Spoken Language Processing*, 827–830.

Byrd, D. (2000). Articulatory vowel lengthening and coordination at phrasal junctures. *Phonetica*, *57*(1), 3–16.

Byrd, D., Krivokapić, J., & Lee, S. (2006). How far, how long : On the temporal scope of prosodic boundary effects. *The Journal of the Acoustical Society of America*, *120*, 1589–1599.

Byrd, D., Lee, S., Riggs, D., & Adams, J. (2005). Interacting effects of syllable and phrase position on consonant articulation. *The Journal of the Acoustical Society of America*, *118*, 3860-3873.

Byrd, D., & Saltzman, E. (2003). The elastic phrase : modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, *31*(2), 149–180.

Caelen-Haumont, G. (2005). Les états émotionnels et la Prosodie : paradigmes, modèles, paramètres. *Phonologie et phonétique : forme et substance, NGuyen, N. (ed.)*, 397–424.

Cambier-Langeveld, T. (1997). The domain of final lengthening in the production of Dutch. *Linguistics in the Netherlands*, *14*, 13–24.

Cambier-Langeveld, T. (2000). Temporal marking of accents and boundaries, Chapter 2 : The domain of final lengthening in Dutch. *Leiden : Holland Institute of Generative Linguistics*, 25–62.

Campbell, W. (1988). Extracting speech-rate values from a real-speech database. *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, 683–686.

Campbell, W. (1992). Multi-level Timing in Speech. *PhD Thesis - University of Sussex, 300p*.

Campbell, W., & Isard, S. (1991). Segment durations in a syllable frame. *Journal of Phonetics*, *19*(1), 37–47.

Campbell, W., & Sagisaka, Y. (1992). Automatic annotation of speech corpora. *Proceedings of the SST92 Queensland, Australia*, 686–691.

Cao, J. (2004). Restudy of segmental lengthening in Mandarin Chinese. *Speech Prosody 2004, International Conference*, 231–234.

Carlson, R. (1991). Duration models in use. *Proceedings of the XIIth Meeting*.

Carlson, R., Elenius, K., & Swerts, M. (2004). Perceptual judgments of pitch range. *Speech Prosody 2004, International Conference*.

Cedergren, H., & Perreault, H. (1994). Speech rate and syllable timing in spontaneous speech. *Third International Conference on Spoken Language Processing*, 1087–1090.

Chafe, W. (1974). Language and consciousness. *Language*, 111–133.

Chen, Y., & Gussenhoven, C. (2008). Emphasis and tonal implementation in Standard Chinese. *Journal of Phonetics*, *36*(4), 724–746.

Cho, H. (2009). *Etude des propriétés acoustiques de la structure prosodique du coréen.* Thèse de doctorat. Université de Provence.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.* Coll. Studies in Language. New York, NY, USA : Harper & Row, 1968, 470 p.

Clark, R. (1999). Using prosodic structure to improve pitch range variation in text to speech synthesis. *XIVth International Congress of Phonetic Sciences*, *1* , 69–72.

Clements, G. (1979). The description of terraced-level tone languages. *Language*, 536–558.

Clements, G. (1990). The status of register in intonation theory : comments on the papers by Ladd and by Inkelas and Leben. *Kingston J, Beckman ME (Hg.), Papers in Laboratory Phonology I : Between the Grammar and Physics of Speech*, 58–71.

Clements, G., & Ford, K. (1979). Kikuyu Tone Shift and its Synchronic Consequences. *Linguistic Inquiry Amherst, Mass.*, *10*(2), 179–210.

Cohen, A., & Hart, J. t. (1967). On the anatomy of intonation. *Lingua*, *19*, 177–192.

Colletta, J., Pellenq, C., & Rousset, I. (2005). Evolution du débit de parole chez l'enfant francophone dans des tâches narrative et conversationnelle. *w3.u-grenoble3.fr*.

Collier, R. (1985). F0 declination : setting and resetting of the baseline. *Ann.Bull.RILP*, *19*, 111–132.

Collier, R. (1987). FO declination : the control of its setting, resetting and slope. , 403–421.

Collier, R., & Cohen, A. (1990). *A perceptual study of intonation : an experimental-phonetic approach to speech melody. Chapter 5 : Declination.* Cambridge University Press.

Connell, B. (2002). Downdrift, downstep, and declination. *Typology of African Prosodic Systems*, 3–12.

Connell, B., & Ladd, D. (1990). Aspects of pitch realisation in Yoruba. *Phonology*, 1–29.

Cooney, O., McGuigan, K., Murphy, P., & Conroy, R. (1998). Acoustic analysis of the effects of alcohol on the human voice. *The Journal of the Acoustical Society of America*, *103*, 2895–2895.

Cooper, W., & Danly, M. (1981). Segmental and Temporal Aspects of Utterance-Final Lengthening in Temporal Aspects of Speech Production and Perception. *Phonetica*, *38*(1-3), 106–115.

Cooper, W., Eady, S., & Mueller, P. (1985). Acoustical aspects of contrastive stress in question–answer contexts. *The Journal of the Acoustical Society of America*, *77*, 2142–2156.

Cosmides, L. (1983). Invariances in the acoustic expression of emotion during speech. *Journal of experimental psychology. Human perception and performance*, *9*(6), 864–881.

Cowie, R., & Cornelius, R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, *40*(1-2), 5–32.

Cowie, R., Douglas-Cowie, E., & Romano, A. (1999). Changing emotional tone in dialogue and its prosodic correlates. *ESCA Tutorial and Research Workshop (ETRW) on Dialogue and Prosody*.

Cruttenden. (2001). Gimson's Pronunciation of English. revised by Cruttenden. *London : Edward Arnold*.

Cruttenden, A. (1997). *Intonation.* Cambridge textbooks in linguistics. Cambridge University Press, 214 p.

Crystal, D. (1969). *Prosodic systems and intonation in English.* Cambridge University Press, 381 p.

Crystal, D. (1975). *The English Tone of Voice : Essays in Intonation, Prosody and Paralanguage.* Hodder and Stoughton Educational.

Crystal, D., & Davy, D. (1969). *Investigating English Style.* Coll. English Language Series. Londres, Grande Bretagne : Longman Group Ltd., 3ème édition, 264 p.

Crystal, T., & House, A. (1982). Segmental durations in connected speech signals : Preliminary results. *The Journal of the Acoustical Society of America*, *72*, 705–716.

Crystal, T., & House, A. (1988). Segmental durations in connected-speech signals : Syllabic stress. *The Journal of the Acoustical Society of America*, *83*, 1574–1585.

Crystal, T., & House, A. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *The Journal of the Acoustical Society of America*, *88*, 101–112.

Cummins, F. (1999). Some lengthening factors in English speech combine additively at most rates. *The Journal of the Acoustical Society of America*, *105*, 476–480.

Cutler, A., & Butterfield, S. (1990). Durational cues to word boundaries in clear speech. *Speech Communication*, *9*(5-6), 485–495.

Cutugno, F., & Savy, R. (1999). Correlation between segmental duration and prosodic feactures in spontaneous speech : the role of tempo. *Proceedings of the XII ICPhS*, 471–474.

Daly, N., & Zue, V. (1992). Statistical and linguistic analyses of F0 in read and spontaneous speech. *Second International Conference on Spoken Language Processing*, 763–766.

Dankovicova, J. (1997). The domain of articulation rate variation in Czech. *Journal of Phonetics*, *25*.

Dankovicova, J. (1999). Articulation rate variation within the intonation phrase in Czech and English. *Proceedings of the 14th International Congress of Phonetic Sciences*, 269–272.

Deinse, J. van. (1981). Registers. *Folia Phoniatr (Basel)*, *33*(1), 37–50.

Delais-Roussarie, E., & Durand, J. (2003). *Corpus et variation en phonologie du français : méthodes et analyses.* Presses universitaires du Mirail, 368p.

Delais-Roussarie, E., Rialland, A., Doetjes, J., & Marandin, J. (2002). The prosody of post-focus sequences in French. *Speech Prosody 2002, International Conference.*

Delattre, P. (1966). Les dix intonations de base du français. *The French Review*, *40*(1), 1–14.

Delgado-Martins, M., & Freitas, M. (1991). Temporal Structures of Speech : Reading News on TV. *Phonetics and Phonology of Speaking Styles.*

Dellwo, V. (2006). Rhythm and Speech rate : A variation coefficient for σ C. *Language and Language-Processing*, 231–241.

Dellwo, V., & Wagner, P. (2002). Relations between language rhythm and speech rate. *XVth International Congress of Phonetic Sciences*, 471–474.

Dellwo, V., & Wagner, P. (2003). Relations between language rhythm and speech rate. *Proceedings of the 15th international congress of phonetics sciences*, 471–474.

De Looze, C., & Hirst, D. (2008). Detecting changes in key and range for the automatic modelling and coding of intonation. *Speech prosody, Campinas, Brazil.*, 135–138.

De Looze, C., & Rauzy, S. (2009). Automatic Detection and Prediction of Topic Changes Through Automatic Detection of Register Variations and Pause Duration. *Proceedings of the 10th Annual Conference of the International Speech Communication Association - Interspeech 2009. Brighton, 6-10 September 2009.*

Demers, M. (2000). Le registre en voix parlée : un indicateur social pour homme seulement ? *XXIIIèmes Journées d'Etude sur la Parole, Aussois, 19-23 juin 2000.*

Den Ouden, H., Noordman, L., & Terken, J. (2008). Prosodic realizations of global and local structure and rhetorical relations in read aloud news report. *Speech Communication, 51*, 116–129.

Di Cristo, A. (1985). *De la microprosodie à l'intonosyntaxe.* Publications, Université de Provence, 854p.

Di Cristo, A. (1998). Intonation in French. *Intonation systems : a survey of twenty languages*, 195–218.

Di Cristo, A. (2000). Interpréter la prosodie. *Actes des XXIIIèmes Journées d'Etude sur la Parole*, 13–29.

Di Cristo, A. (2004). La prosodie au carrefour de la phonétique, de la phonologie et de lárticulation formes-fonctions. *Travaux interdisciplinaires du Laboratoire parole et langage dÁix-en-Provence*(23), 67–211.

Di Cristo, A. (2005). Elements de prosodie. *Phonologie et phonétique (Traité IC2, série Cognition et Traitement de l'Information)*, 117–157.

Di Cristo, A., Auran, C., Bertrand, R., Chanet, C., Portes, C., & Régnier, A. (2004). Outils prosodiques et analyse du discours. *Cahiers de l Institut de Linguistique de Louvain*, *30*(1), 27–84.

Di Cristo, A., & Hirst, D. (1986). Modelling French micromelody : analysis and synthesis. *Phonetica*, *43*(1-3), 11–30.

Di Cristo, A., & Hirst, D. (1993). Rythme syllabique, rythme mélodique et représentatin hiérarchique de la prosodie du français. *Travaux de l'Institut de Phonétique d'Aix*, *15*, 9–24.

Di Cristo, A., & Hirst, D. (1996). Vers une typologie des unités intonatives du français. *Actes des XXIème JEP*, 223–226.

D'Imperio, M., & A., M. (2009). Mapping syntax onto prosodic structure : evidence for the intermediate phrase in French. *PAPI Conference*.

D'Imperio, M., Espesser, R., Loevenbruck, H., Menezes, C., Nguyen, N., & Welby, P. (2007). Are tones aligned with articulatory events ? Evidence from Italian and French. *Papers in Laboratory Phonology 9, Cole, J. (Ed.)*, 57-608.

Dioubina, O. (2004). Prosody of Dialogues : Influence of Recognition Failure on Local Speech Rate. *Speech Prosody 2004, International Conference*, 275–278.

Doherty, R., & Lee, A. (2009). Speech rates of Irish English-speaking adults. *Paper to be presented at the RCSLT Scientific Conference, London, UK*.

Dommelen, W. van, & Moxness, B. (1995). Acoustic parameters in speaker height and weight identification : sex-specific behaviour. *Lang Speech*, *38*(Pt 3), 267–87.

Dubois, J. (1994). *Dictionnaire de linguistique et des sciences du langage*. Coll. Trésors du Français. Paris, France : Larousse.

Duez, D. (1982). Silent and non-silent pauses in three speech styles. *Language and Speech Teddington*, *25*(1), 11–28.

Duez, D. (1997). Acoustic markers of political power. *Journal of Psycholinguistic Research*, *26*(6), 641–654.

Dziubinski, M., & Kostek, B. (2004). High accuracy and octave error immune pitch detection algorithms. *Archives of Acoustics*, *29*(1), 3–24.

Eady, S., & Cooper, W. (1986). Speech intonation and focus location in matched statements

and questions. *The Journal of the Acoustical Society of America*, *80*, 402–415.

Earle, M. (1975). An acoustic phonetic study of Northern Vietnamese tones. *Santa Barbara, Speech Communications Research Laboratory*, 200–211.

Edwards, J., & Beckman, M. (1987). Perception of Final Lengthening. *Annual Meeting of the Linguistic Society of America (San Francisco, CA, December 27-30)*, 1–13.

Eefting, W. (1988). Temporal Variation in Natural Speech : Some Explorations. *Proc. of Speech*, *7*, 503–507.

Eefting, W. (1991). The effect of « information value »and « accentuation »on the duration of Dutch words, syllables, and segments. *The Journal of the Acoustical Society of America*, *89*, 412–424.

Eefting, W., & Rietveld, A. (1989). Just noticeable differences of articulation rate at sentence level. *Speech Communication*, *8*(4), 355–361.

Erickson, M. (2000). Simultaneous effects on vowel duration in American English : A covariance structure modeling approach. *The Journal of the Acoustical Society of America*, *108*, 2980–2995.

Ericson, D., & Lehiste, I. (1995). Contrastive emphasis in elicited dialogue : durational compensation. *Proceedings ICPhS*, *95*, 352–355.

Eskenazi, M. (1993). Trends in speaking styles research. *Third European Conference on Speech Communication and Technology*, 501–509.

Evans, S., Neave, N., & Wakelin, D. (2006). Relationships between vocal characteristics and body size and shape in human males : An evolutionary explanation for a deep male voice. *Biological psychology*, *72*(2), 160–163.

Fackrell, J., Vereecken, H., Buhmann, J., Martens, J., & Coile, B. (2000). Prosodic variation with text type. *Sixth International Conference on Spoken Language Processing*, *3*, 231–234.

Fagyal, Z. (2002). Tonal Template for Background Information : the Scaling of Pitch in Utterance-Medial Parentheticals in French. *Speech Prosody 2002, International Conference*.

Fant, G., Kruckenberg, A., & Nord, L. (1991). Durational correlates of stress in Swedish, French, and English. *Journal of phonetics*, *19*(3-4), 351–365.

Fant, K. A., G., & Nord, L. (1990). Prosodic and segmental speaker variations. *Proceedings of the tutorial and research workshop on speaker characterization in speech technology Edinburgh 26-8 June*, 106–120.

Féry, C., & Truckenbrodt, H. (2005). Sisterhood and tonal scaling*. *Studia Linguistica*, *59*(2-3), 223–243.

Fitzsimons, M., Sheahan, N., & Staunton, H. (2001). Gender and the integration of acoustic dimensions of prosody : Implications for clinical studies. *Brain and Language*, *78*(1), 94–108.

Fletcher, J., & McVeigh, A. (1992). Towards a model of segment and syllable duration in

Australian English. *Proceedings of the fourth Australian International Conference on Speech Science and Technology. Brisbane : ASSTA*, 28–33.

Fon, J. (1999). Speech rate as a reflection of variance and invariance in conceptual planning in storytelling. *Proceedings of 14th International Congress of Phonetics Sciences. San Francisco, USA.*, 663–666.

Fon, J. (2002). *A cross linguistic study on syntactic and discourse boundary cues in spontaneous speech.* Doctorial dissertation, The Ohio State University.

Fónagy, I., & Magdics, K. (1960). Speed of utterance in phrases of different lengths. *Language and Speech*, *3*, 179–192.

Fougeron, C. (2001). Articulatory properties of initial segments in several prosodic constituents in French. *Journal of Phonetics*, *29*(2), 109–135.

Fougeron, C., & Jun, S. (1998). Rate effects on French intonation : prosodic organization and phonetic realization. *Journal of Phonetics*, *26*(1), 45–69.

Fougeron, C., & Keating, P. (1997). Articulatory strengthening at edges of prosodic domains. *The journal of the acoustical society of America*, *101*, 3728–3740.

Fowler, C., & Housum, J. (1987). Talkers' signaling of new and old. words in speech and listeners' perception and use of the distinction. *Journal of memory and language(Print)*, *26*(5), 489–504.

Fox, A. (2000). *Prosodic Features and Prosodic Structure : The Phonology of Suprasegmentals.* Oxford University Press, USA ; 375p.

Frick, R. (1985). Communicating emotion : The role of prosodic features. *Psychological Bulletin*, *97*(3), 412–429.

Fujisaki, H. (1991). Modeling the generation process of F0 contours as manifestation of linguistic and paralinguistic information. *Proceedings of the XIIth International Congress of Phonetic Sciences*, 1–10.

Gandour, J., Tumtavitikul, A., & Satthamnuwong, N. (2000). Effects of speaking rate on Thai tones. *Phonetica*, *56*(3-4), 123–134.

Gårding, E. (1983). A generative model of intonation. *Prosody : Models and measurements*, 11–25.

Gee, J., & Grosjean, F. (2002). Performance structures. *Psycholinguistics : critical concepts in psychology*, *15*, 411–458.

Ghio, A. (2007). L'évaluation acoustique. *Les dysarthries, Auzou P. ; Rolland V. ; Pinto S. ; Ozsancak C. (eds.) - ISBN 978-2-35327-021-7. Marseille : Solal. 2007*, 236–247.

Gibbon, D. (1998). Intonation in German. *Intonation systems : A survey of twenty languages*, 78–95.

Goldman-Eisler, F. (1954). On the variability of the speed of talking and on its relation to the length of utterances in conversations. *Br J Med Psychol*, *45*(2), 94–107.

Goldman-Eisler, F. (1956). The determinants of the rate of speech output and their mutual relations. *J Psychosom Res*, *1*(2), 137–43.

Goldman-eisler, F. (1958). Speech production and the predictability of words in context. *The Quarterly Journal of Experimental Psychology*, *10*(2), 96–106.

Goldman-Eisler, F. (1961). The significance of changes in the rate of articulation. *Language and Speech*, *4*(3), 171–174.

Goldman-Eisler, F. (1968). *Psycholinguistics : Experiments in spontaneous speech.* Academic Press New York, 169 p.

Goldsmith, J. (1976). An overview of autosegmental phonology. *Linguistic Analysis*, *2*(1), 23–68.

Goldsmith, J. (1990). *Autosegmental and metrical phonology.* B. Blackwell Cambridge, Mass., USA.

Goldsmith, J. (1999). *Phonological theory : the essential readings.* Blackwell Malden, MA, 428p.

Graddol, D. (1986). Discourse specific pitch behaviour. *Intonation in discourse*, 221–238.

Graham, C., Hamblin, A., & Feldstein, S. (2001). Recognition of emotion in English voices by speakers of Japanese, Spanish and English. *IRAL-International Review of Applied Linguistics in Language Teaching*, *39*(1), 19–37.

Grawunder, S., Bose, I., Hertha, B., Trauselt, F., & Anders, L. (2006). Perceptive and acoustic measurement of average speaking pitch of female and male speakers in German radio news. *Ninth International Conference on Spoken Language Processing*.

Gros, J., Miheli, F., & Pave, N. (1999). Slovenian speech timing at different speaking rates. *ICPhS99, San Francisco.*, 261–264.

Grosjean, F. (1977). The perception of rate in spoken and sign languages. *Perception and Psychophysics*, *22*, 408–413.

Grosjean, F., & Collins, M. (1979). Breathing, Pausing and Reading. *Phonetica Basel*, *36*(2), 98–114.

Grosjean, F., & Deschamps, A. (1972a). Analyse des variables temporelles du francais spontane (Analysis of Time Variables in Spontaneous French). *Phonetica 26(3)*, 129–56.

Grosjean, F., & Deschamps, A. (1972b). Analyse des variables temporelles du francais spontané (II. Comparaison du français oral dans la description avec l'anglais (description) et avec le français (interview radiophonique). *Phonetica 28(3).*

Grosjean, F., & Deschamps, A. (1975). Analyse contrastive des variables temporelles de l'anglais et du français : vitesse de parole et variables composantes, phénomènes d'hésitation. *Phonetica*, *31*(4), 144–184.

Grosjean, F., & Dommergue, J. (1983). Les structures de performance en psycholinguistique. *L'année Psychologique*, *83*(2), 513–536.

Grosz, B., & Hirschberg, J. (1992). Some intonational characteristics of discourse structure. *In Proceedings of the International Conference on Spoken Language Processing*, 1–5.

Grosz, B., & Sidner, C. (1986). Attention, intentions, and the structure of discourse. *Computational linguistics*, *12*(3), 175–204.

Gussenhoven, C. (2002). Phonology of intonation. *Glot International*, *6*(9/10), 271–284.

Gussenhoven, C. (2004). *The Phonology of Tone and Intonation.* Cambridge University Press.

Gussenhoven, C. (2005). *Understanding Phonology.* London, UK : Hodder Arnold, 2005, 2nd edition, 284 p.

Gussenhoven, C. (2007). *Phonology of Intonation.* In Paul de Lacy (ed.) The Cambridge Handbook of Phonology. Cambridge University Press.

Gussenhoven, C., & Rietveld, A. (1988). Fundamental frequency declination in Dutch : Testing three hypotheses. *Journal of phonetics*, *16*, 355–369.

Gussenhoven, C., & Rietveld, T. (2000). The behavior of H* and L* under variations in pitch range in Dutch rising contours. *Language and Speech*, *43*(2), 183–203.

Gut, U. (2007). Foreign accent. *Lecture notes in computer science*, *4343*, 1611–3349.

Halliday, M. (1967). *Intonation and grammar in British English.* Mouton, 62p.

Halliday, M. (1970). Author A course in spoken English : intonation. *London : Oxford University Press, 134p.*

Halliday, M. (1985). Dimensions of discourse analysis : grammar. *Handbook of discourse analysis*, *2*, 29–56.

Hanson, H. (2009). Effects of obstruent consonants on fundamental frequency at vowel onset in English. *The Journal of the Acoustical Society of America*, *125*, 425-441.

Hansson, P. (2003). *Prosodic phrasing in spontaneous Swedish.* Citeseer.

Hart, C. R. 't, J., & Cohen, A. (1990). *A perceptual study of intonation : an experimental-phonetic approach to speech melody.* Cambridge University Press.

Hartmann, R., & Stork, F. (1972). *Dictionary of Language and Linguistics.* John Wiley and Sons, Halsted Press Division, 605 Third Avenue, New York, NY 10016.

Heerden, C. van, & Barnard, E. (2006). Speech rate normalization used to improve speaker verification. *Proceedings of the Symposium of the Pattern Recognition Association of South Africa*, 2–7.

Helfrich, H. (1979). Age markers in speech. *Social markers in speech*, 63–107.

Henton, C. (1995). Pitch dynamism in female and male speech. *Language and Communication*, *15*(1), 43–61.

Hermes, D., & Van Gestel, J. (1991). The frequency scale of speech intonation. *The Journal of the Acoustical Society of America*, *90*, 97–102.

Hess, W., et al. (1983). *Pitch determination of speech signals : algorithms and devices.* Springer-Verlag, 698p.

Hirose, K., & Kawanami, H. (1998). On the relationship of speech rates with prosodic units in dialogue speech. *Fifth International Conference on Spoken Language Processing.*

Hirose, K., & Kawanami, H. (2002). Temporal rate change of dialogue speech in prosodic units as compared to read speech. *Speech Communication*, *36*(1-2), 97–111.

Hirschberg, J. (2000). A corpus-based approach to the study of speaking style. *Prosody : theory and experiment. Studies presented to Gosta Bruce*, 335–350.

Hirschberg, J., & Grosz, B. (1992a). Intonational features of local and global discourse structure. *Proceedings of the Speech and Natural Language Workshop*, 441–446.

Hirschberg, J., & Grosz, B. (1992b). Intonation features of local and global discourse structure. *Proceeding of the DARPA Workshop on Spoken Language Systems*.

Hirschberg, J., & Nakatani, C. (1996). A prosodic analysis of discourse segments in direction-giving monologues. *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, 286–293.

Hirschberg, J., & Pierrehumbert, J. (1986). The intonational structuring of discourse. *Proceedings of the 24th annual meeting on Association for Computational Linguistics*, 136–144.

Hirschberg, J., & Ward, G. (1992). The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in English. *Journal of Phonetics*, *20*(2), 241–251.

Hirst, D. (1977). *Intonative features : a syntactic approach to English intonation.* Mouton, 135p.

Hirst, D. (1987). La représentation linguistique des systèmes prosodiques : une approche cognitive. *Thèse de Doctorat d'Etat en Phonétique*, 521p.

Hirst, D. (1998). Intonation in British English. *Intonation Systems. A Survey of Twenty Languages*, 56–77.

Hirst, D. (2005). Form and function in the representation of speech prosody. *Speech Communication*, *46*(3-4), 334–347.

Hirst, D. (2006). Prosodic aspects of speech and language. *Elsevier Ltd.*.

Hirst, D. (2007). A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation. *Proc. Int. Conf. Phonetic Sci. XVI, Saarbrucken*, 1233–1236.

Hirst, D. (2009). The Rhythm of Text and the Rhythm of Utterances : From Metrics to Models. *Proceedings of the INTERSPEECH conference, Brighton.*, 1519–1523.

Hirst, D., & Bouzon, C. (2005). The effect of stress and boundaries on segmental duration in a corpus of authentic speech (British English). *Ninth European Conference on Speech Communication and Technology*.

Hirst, D., Bouzon, C., & Auran, C. (2007). Analysis by synthesis of British English speech rhythm : from data to models. *In Gunnar Fant; Hiroya Fujisaki; Jiaxuan Shen (eds.) : Festschrift for Professor Wu Zongji's 100th birthday. Beijing, People's Republic of China : Commercial Press*.

Hirst, D., & Di Cristo, A. (1984). French intonation : a parametric approach. *Die Neueren Sprachen*, *83*(5), 554–569.

Hirst, D., & Di Cristo, A. (1998). *Intonation Systems : A Survey of Twenty Languages.* Cambridge University Press, 487 p.

Hirst, D., Di Cristo, A., & Espesser, R. (2000). Levels of representation and levels of analysis for the description of intonation systems. *Prosody : Theory and Experiment. Studies*

Presented to Gösta Bruce, 51–87.

Hirst, D., & Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic sline function. *Travaux de l'Institut de Phonétique d'Aix*, *15*, 71–85.

Hockey, B., & Fagyal, Z. (1998). Pre-Boundary Lengthening : Universal or Language-Specific ? The Case of Hungarian. *University of Pennsylvania Working Popen in Linguistics*, *5*, 71–82.

Hoequist, C., & Kohler, K. (1986). Summary of speech rate perception research at Kiel. *Arbeitsberichte- Institut fur Phonetik*(22), 5–28.

Hollien, H. (1972). Three major vocal registers : a proposal. *Proceedings of the Seventh International Congress of Phonetic Sciences. The Hague : Mouton*, 320–331.

Hollien, H., Dejong, G., Martin, C., Schwartz, R., & Liljegren, K. (2001). Effects of ethanol intoxication on speech suprasegmentals. *The Journal of the Acoustical Society of America*, *110*, 3198–3206.

Hollien, H., Dew, D., & Philips, P. (1971). Phonational frequency ranges of adults. *Journal of Speech, Language and Hearing Research*, *14*(4), 755–760.

Hollien, H., Hollien, P., & Jong, G. de. (1997). Effects of three parameters on speaking fundamental frequency. *The Journal of the Acoustical Society of America*, *102*, 2984–2992.

Hollien, H., & Shipp, T. (1972). Speaking fundamental frequency and chronologic age in males. *Journal of Speech, Language and Hearing Research*, *15*(1), 155–159.

Hombert, J. (1974). Universals of downdrift : their phonetic basis and significance for a theory of tone. *Studies in African Linguistics*, *5*, 169–183.

Honda, K. (2004). Physiological factors causing tonal characteristics of speech : from global to local prosody. *Speech Prosody 2004, International Conference*, 739–744.

Horne, M., Strangert, E., & Heldner, M. (1995). Prosodic boundary strength in Swedish : Final lengthening and silent interval duration. *Proceedings ICPhS*, *95*, 170–173.

House, A. (1961). On vowel duration in English. *The Journal of the Acoustical Society of America*, *33*, 1174–1178.

House, A., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *The Journal of the Acoustical Society of America*, *25*, 105–113.

Huber, D. (1989). Voice Characteristics of Female Speech and Their Representation in Computer Speech Synthesis and Recognition. *First European Conference on Speech Communication and Technology*, 2477–2480.

Hudson, A., & Holbrook, A. (1982). Fundamental frequency characteristics of young black adults : Spontaneous speaking and oral reading. *Journal of Speech and Hearing Research*, *25*(1), 25–28.

Ibrakhim, I. (2004). Universal and Linguistic Features of Expressing Emotional Information : Differentiation in the Perception Level. *Speech Prosody 2004, International Conference*,

659–662.

Igarashi, Y. (2004). Segmental Anchoring of F0 Under Changes in Speech Rate : Evidence from Russian. *Speech Prosody 2004, International Conference*, 25–28.

Iivonen, A., Niemi, T., & Paananen, M. (1995). Comparison of prosodic characteristics in English, Finnish and German radio and TV newscasts. *ICPhS*, *95*, 382–385.

Inkelas, S., & Leben, W. (1990). Where phonology and phonetics intersect : the case of Hausa intonation. *Between the Grammar and Physics of Speech*, 17–34.

Jacewicz, E., Fox, R., O'Neill, C., & Salmons, J. (2009). Articulation rate across dialect, age, and gender. *Language Variation and Change*, *21*(02), 233–256.

Jakobson, R., Fant, G., & Halle, M. (1969). *Preliminaries to speech analysis.* MIT press Cambridge, 64 p.

Janse, E., Sennema, A., & Slis, A. (2000). Fast speech timing in Dutch : durational correlates of lexical stress and pitch accent. *Sixth International Conference on Spoken Language Processing*.

Jassem, W. (1952). *Intonation of Conversational English :(educated Southern British).* Nakl. Wroclawskiego Tow. Naukowego ; skl. gl. : Dom Ksia¸zki.

Jassem, W. (1971). Pitch and compass of the speaking voice. *Journal of Phonetics*, *1*, 59–68.

Johns-Lewis, C. (1986). *Intonation in discourse.* College Hill Press, 302 p.

Johnson, K., Pisoni, D., & Bernacki, R. (1990). Do voice recordings reveal whether a person is intoxicated ? A case study. *Phonetica*, *47*(3-4), 215–237.

Jun, S., & Fougeron, C. (2000). A phonological model of French intonation. *Intonation : Analysis, modelling and technology*, 209–242.

Keating, P., Cho, T., Fougeron, C., & Hsu, C. (1999). Domain-initial articulatory strengthening in four languages. *Hsu - linguistics.ucla.edu*.

Keller, E., & Zellner, B. (1996). A timing model for fast French. *York Papers in Linguistics*, *17*, 53–75.

Kessinger, R., & Blumstein, S. (1997). Effects of speaking rate on voice-onset time in Thai, French, and English. *Journal of Phonetics*, *25*(2), 143–168.

Kiessling, A., Kompe, R., Niemann, H., Nöth, E., & Batliner, A. (1995). Voice source state as a source of information in speech recognition : Detection of laryngealizations. *nato asi series of computer and systems sciences*, *147*, 329–332.

Kirkham, S. (2002). Tempo Modulations in English : selected pilot study results. *ICSLP*, 1765–1768.

Klatt, D. (1973). Interaction between two factors that influence vowel duration. *The Journal of the Acoustical Society of America*, *54*, 1102–1104.

Klatt, D. (1975). Vowel Lengthening is Syntactically Determined in a Connected Discourse. *Journal of Phonetics*, *3*(3), 129–140.

Klatt, D. (1976). Linguistic uses of segmental duration in English : Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, *59*, 1208–1221.

Klatt, D. (1987). Review of text-to-speech conversion for English. *The Journal of the Acoustical Society of America*, *82*, 737–793.

Klingholz, F., Penning, R., & Liebhardt, E. (1988). Recognition of low-level alcohol intoxication from speech signal. *The Journal of the Acoustical Society of America*, *84*, 929–935.

Knowles, G., Wichmann, A., & Alderson, P. (1996). *Working with speech.* Addison Wesley Longman New York, 236 p.

Künzel, H. (1989). How well does average fundamental frequency correlate with speaker height and weight ? *Phonetica*, *46*(1-3), 117–25.

Kohler, K. (1982). F0 in the production of lenis and fortis plosives. *Phonetica*, *39*(4-5), 199–218.

Kohler, K. (1983). Prosodic boundary signals in German. *Phonetica*, *40*(2), 89–134.

Kohler, K. (1986). Parameters of speech rate perception in German words and sentences : Duration, F0 movement, and F0 level. *Language and Speech*, *29*(2), 115–139.

Koiso, H., Shimojima, A., & Katagiri, Y. (1998). Collaborative signaling of informational structures by dynamic speech rate. *Language and Speech*, *41*(3-4), 323–350.

Kong, E. (2004). The role of pitch range variation in the discourse structure and intonation structure of Korean. *Eighth International Conference on Spoken Language Processing*.

Koopmans-Van Beinum, F. (1992). The role of focus words in natural and in synthetic continuous speech : Acoustic aspects. *Speech Communication*, *11*(4-5), 439–452.

Koreman, J. (2003). The perception of articulation rate. *Proc. 15th ICPhS, Barcelona*, 337–342.

Koreman, J. (2006). Perceived speech rate : The effects of articulation rate and speaking style in spontaneous speech. *The Journal of the Acoustical Society of America*, *119*, 582–596.

Kowal, S., O'Connell, D., & Sabin, E. (1975). Development of temporal patterning and vocal hesitations in spontaneous narratives. *Journal of Psycholinguistic Research*, *4*(3), 195–207.

Kutik, E., Cooper, W., & Boyce, S. (1983). Declination of fundamental frequency in speakers' production of parenthetical and main clauses. *The Journal of the Acoustical Society of America*, *73*, 1731–1738.

Kuwabara, H. (1996). Acoustic properties of phonemes in continuous speech for different speaking rate. *Fourth International Conference on Spoken Language Processing*.

Lacheret-Dujour, F., A. et Beaugendre. (1999). La prosodie du français. *Coll. CNRS Langage. Paris, France : CNRS éditions*, 354 p.

Ladd, D. (1983). Phonological features of intonational peaks. *Language*, 721–759.

Ladd, D. (1984). Declination : a review and some hypotheses. *Phonology yearbook*, 53–74.

Ladd, D. (1988). Declination « reset »and the hierarchical organization of utterances. *The Journal of the Acoustical Society of America*, *84*, 530–544.

Ladd, D. (1990). Metrical representation of pitch register. *Papers in laboratory phonology I : Between the grammar and physics of speech*, 35–57.

Ladd, D. (1992). An introduction to intonational phonology. *Docherty, Ladd, Papers in Laboratory Phonology. II : Gesture, segment, prosody*, 321–334.

Ladd, D. (1993). On the theoretical status of the baselin' in modelling intonation. *Language and Speech*, *36*(4), 435–451.

Ladd, D. (1994). Constraints on the gradient variability of pitch range, or, Pitch level 4 lives ! *Phonological Structure and Phonetic Form*, 43–63.

Ladd, D. (1996). *Intonational Phonology.* Cambridge University Press, 334 p.

Ladd, D., & Campbell, N. (1991). Theories of prosodic structure : evidence from syllable duration. *Proceedings of the 12th International Congress of Phonetic Sciences*, *2*, 290–293.

Ladd, D., Faulkner, D., Faulkner, H., & Schepman, A. (1999). Constant segmental anchoring of F movements under changes in speech rate. *The Journal of the Acoustical Society of America*, *106*, 1543–1554.

Ladd, D., & Morton, R. (1997). The perception of intonational emphasis : continuous or categorical ? *Journal of Phonetics*, *25*(3), 313–342.

Ladd, D., Scherer, K., & Silverman, K. (1986). An integrated approach to studying intonation and attitude. *Intonation in discourse. London/Sidney : Crom Helm*, 125–138.

Ladd, D., Silverman, K., Tolkmitt, F., Bergmann, G., & Scherer, K. (1985). Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. *The Journal of the Acoustical Society of America*, *78*, 435–444.

Ladd, D., & Terken, J. (1995). Modelling intra-and inter-speaker pitch range variation. *Proceedings of the International Conference of Phonetic Sciences*, *2*, 386–389.

Ladefoged, P. (2001). A course in phonetics. New York. *NY : Harcourt College Publishers.*

Lambrecht, K. (1996). *Information structure and sentence form.* Cambridge Univ. Press, 388p.

Lane, H., & Grosjean, F. (1973). Perception of reading rate by speakers and listeners. *J Exp Psychol*, *97*(2), 141–7.

Laniran, Y., & Clements, G. (2003). Downstep and high raising : interacting factors in Yoruba tone production. *Journal of Phonetics*, *31*(2), 203–250.

Lass, N., & Clegg, J. (1974). Comparative study of Temporal characteristics of Picture-Elicited and Topic-elicited Speech. *Speech and Hearing Science*, 995–8.

Lass, N., & Davis, M. (1976). An investigation of speaker height and weight identification. *The Journal of the Acoustical Society of America*, *60*, 700–703.

Lass, N., & Deem, J. (1972). Temporal patterns of speech rate alternations. *Proceedings of the 7th International Congress on Phonetic Sciences.*, 922–927.

Lass, N., & Sandusky, J. (1971). A study of the relationship of diadochokinetig rate, speaking rate and reading rate. *Communication Quarterly*, *19*(3), 49–54.

Laver, J. (1991). *The Gift of Speech.* Edinburgh University Press, 400 p.

Laver, J., & Hanson, R. (1981). Describing the normal voice. *Speech evaluation in psychiatry.*

New York : Grune and Stratton.

Laver, J., & Trudgill, P. (1979). Phonetic and linguistic markers in speech. *Social markers in speech*, 1–32.

Leben, W. (1973). *Suprasegmental phonology.* Massachusetts Institute of Technology, 194 p.

Lee, S., & Oh, Y. (1999). Tree-based modeling of prosodic phrasing and segmental duration for Korean TTS systems. *Speech Communication*, *28*(4), 283–300.

Lehiste, I. (n.d.). Perception of Sentence and Paragraph Boundaries. *Frontiers of Speech Communication Research.*

Lehiste, I. (1965). The function of quantity in Finnish and Estonian. *Language*, *41*(3), 447–456.

Lehiste, I. (1970). *Suprasegmentals.* Cambridge, Mass. : MIT Press.

Lehiste, I. (1975). The phonetic structure of paragraphs. *1975*, 195–203.

Lehiste, I. (1976). Suprasegmental features of speech. *Contemporary issues in experimental phonetics*, 225–239.

Lehiste, I., Olive, J., & Streeter, L. (1976). Role of duration in disambiguating syntactically ambiguous sentences. *The Journal of the Acoustical Society of America*, *60*, 1199–1202.

Lehtonen, J. (1978). On factors affecting the pitch level of speech. *Nordic Prosody papers from a symposium*, 55-63.

Lennes, M., & Anttila, H. (2002). Prosodic features associated with the distribution of turns in finnish informal dialogues. *The Phonetics Symposium*, 149–158.

Levelt, W. (1989). *Speaking : From intention to articulation - Chapter 10.* Cambridge, MA : MIT Press.

Liberman, M. (1975). *The intonational system of English.* Massachusetts Institute of Technology.

Liberman, M., & Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. *Language Sound Structure*, 157–233.

Liberman, M., & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, *8*(2), 249–336.

Lieberman, P., Katz, W., Jongman, A., Zimmerman, R., & Miller, M. (1985). Measures of the sentence intonation of read and spontaneous speech in American English. *The Journal of the Acoustical Society of America*, *77*, 649–657.

Lienard, J. (1989). Variabilité, contraintes et spécification de la parole : un cadre théorique. *Actes du séminaire Variabilité et spécificité du locuteur : Etudes et applications SFA*, 1–10.

Lin, H.-Y., & Fon, J. (2009). Perception of Temporal Cues at Discourse Boundaries. *Proceedings INTERSPEECH09*, 808–811.

Lindblom, B. (1968). Temporal organization of syllable production. *Quarterly Progress and Status Report*, *2*(3), 1–5.

Linville, S. (1987). Maximum phonational frequency range capabilities of women's voices with

advancing age. *Folia Phoniatr (Basel)*, *39*(6), 297–301.

Lobacz, P. (1976). Objective and subjective speech tempo in Polish. *Speech Analysis and Synthesis*, *4*, 173–186.

Loveday, L. (1981). Pitch, Politeness and Sexual Role : An Exploratory Investigation into the Pitch Correlates of English and Japanese Politeness Formulae. *Language and Speech*, *24*(1), 71–89.

Maeda, S. (1974). Characterization of fundamental-frequency contours of speech. *The Journal of the Acoustical Society of America*, *56*, S33.

Maeda, S. (1976). *A characterization of American English intonation*. Doctoral dissertation, Massachussets Institute of Technology, 332p.

Magen, H., & Blumstein, S. (1991). Effects of speaking rate on the vowel length distinction in Korean. *The Journal of the Acoustical Society of America*, *89*, 1918–1918.

Mairesse, F., Walker, M., Mehl, M., & Moore, R. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research, 350 p.*, *30*, 457–500.

Malecot, A., Johnston, R., & Kizziar, P. (1972). Syllabic rate and utterance length in French. *Phonetica*, *26*(4), 235–251.

Man, V. (2002). Focus effects on Cantonese tones : An acoustic study. *Speech Prosody 2002, International Conference*.

Martin, P. (1980). Une théorie syntaxique de l'accentuation en français. *Studia Phonetica Montréal*, *15*, 1–12.

Mayer, J., Jasinskaja, E., & Kölsch, U. (2006). Pitch range and pause duration as markers of discourse hierarchy : Perception experiments. *Ninth International Conference on Spoken Language Processing*.

McConnell-Ginet, S. (1978). Intonation in a man's world. *Signs*, 541–559.

Menn, L., & Boyce, S. (1982). Fundamental frequency and discourse structure. *Language and Speech Teddington*, *25*(4), 341–383.

Mennen, I., Schaeffler, F., & Docherty, G. (2007). Pitching it differently : a comparison of the pitch ranges of German and English speakers. *Proceedings of the 16th International Congress of Phonetic Sciences*, 1769–1772.

Mennen, I., Schaeffler, F., & Docherty, G. (2008). A methodological study into the linguistic dimensions of pitch range differences between German and English. *Speech Prosodu 2008. Campinas*.

Mennen, S. F., I., & Docherty, G. (2008). An investigation of cross-language differences in pitch range for speakers of English and German. *Laboratory Phonology (LabPhon) 11 conference - abstracts (edited by Paul Warren)*.

Mertens, P., Auchlin, A., Goldman, J., & Grobet, A. (2001). L'intonation du discours : une implémentation par balises ; motifs et premiers résultats. *Journées Prosodie 2001*.

Meynadier, Y. (2003). Interaction entre prosodie et (co) articulation linguopalatale en fran-

çais. *Unpublished Thèse de Doctorat, Université Aix-Marseille I–Université de Provence, 200p.*

Michelas, A., & D'Imperio, M. (soumis). Durational cues and prosodic phrasing in French : evidence for the intermediate phrase. *Speech Prosody 2010*.

Miller, J., & Baer, T. (1983). Some effects of speaking rate on the production of/b/and/w. *Journal of the Acoustical Society of America*, *73*(5), 1751–1755.

Miller, J., Grosjean, F., & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech : a reanalysis and some implications. *Phonetica*, *41*(4), 215–25.

Möhler, G., & Mayer, J. (1999). A Method for the Analysis of Prosodic Registers. *Sixth European Conference on Speech Communication and Technology*.

Monaghan, A. (2001). An Auditory Analysis of the Prosody of Fast and Slow Speech Styles in English, Dutch and German. *Improvements in speech synthesis : COST 258 : the naturalness of synthetic speech*, 204–217.

Montero, J., Gutiérrez-Arriola, J., Colás, J., Enríquez, E., & Pardo, J. (1999). Analysis and modelling of emotional speech in Spanish. *Proc. of ICPhS*, *2*, 957–960.

Moore, B. (1989). *An introduction to the psychology of hearing. Third Edition.* Academic press.

Mozziconacci, S. (1998). *Speech variability and emotion : Production and perception.* Technische Universiteit Eindhoven.

Mozziconacci, S. (2000). The expression of emotion considered in the framework of an intonation model. *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

Mozziconacci, S. (2002). Prosody and Emotions. *Speech Prosody 2002, International Conference*.

Mozziconacci, S., & Hermes, D. (1999). Role of intonation patterns in conveying emotion in speech. *Proceedings of ICPhS*, 2001–2004.

Mozziconacci, S., & Hermes, D. (2000). Variations temporelles communiquant l'émotion dans la parole. *XXIII journées d'étude sur la parole*.

Murray, I., & Arnott, J. (1993). Toward the simulation of emotion in synthetic speech : A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, *93*, 1097–1108.

Nakajima, S., & Allen, J. (1992). Prosody as a cue for discourse structure. *Second International Conference on Spoken Language Processing*, 425–428.

Nakajima, S., & Allen, J. (1993). A study on prosody and discourse structure in cooperative dialogues. , 197–210.

Nakatani, L., O'Connor, K., & Aston, C. (1981). Prosodic aspects of American English speech rhythm. *Phonetica*, *38*, 84–106.

Nespor, M., & Vogel, I. (1983). Prosodic structure above the word. *Prosody : Models and measurements*, 123–140.

Nicolas, P., & Hirst, D. (1995). Symbolic Coding of Higher-Level Characteristics of Funda-

mental Frequency Curves. *Fourth European Conference on Speech Communication and Technology*, 989–992.

Niemann, H., Denzler, J., Kahles, B., Kompe, R., Kiessling, A., Noth, E., et al. (1994). Pitch determination considering laryngealization effects in spokendialogs. *1994 IEEE International Conference on Neural Networks, 1994. IEEE World Congress on Computational Intelligence*, *7*.

Nolan, F. (2003). Intonational equivalence : an experimental evaluation of pitch scales. *Proceedings of the 15th International Congress of Phonetic Sciences*, 771–774.

Noll, A. (1968). Clipstrum pitch determination. *The Journal of the Acoustical Society of America*, *44*, 1585–1591.

Nooteboom, S. (1985). A functional view of prosodic timing in speech. *Time, Mind, and Behavior. Springer, Berlin*, 242–252.

O'Connor, J., & Arnold, G. (1961). *Intonation of colloquial English : a practical handbook.* Longmans.

Ohala, J. (1983). Cross-language use of pitch : an ethological view. *Phonetica*, *40*(1), 1–18.

Ohala, J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica*, *41*(1), 1.

Ohala, J., Dunn, A., & Sprouse, R. (2004). Prosody and phonology. *Speech Prosody 2004, International Conference*, 161–163.

Ohno, S., Fukumiya, M., & Fujisaki, H. (1996). Quantitative analysis of the local speech rate and its application to speech synthesis. *Fourth International Conference on Spoken Language Processing*.

Oliveira, M. (n.d.). Pitch reset as a cue for narrative segmentation. *Proccedings of IP2003, Prosodic Interfaces*, 73–78.

Oller, D. (1973). The effect of position in utterance on speech segment duration in English. *The Journal of the Acoustical Society of America*, *54*, 1235–1247.

O'Neill, C. (2008). *Dialect Variation in Speaking Rate.* The Ohio State University.

Os, E. den. (1985). Perception of speech rate of Dutch and Italian utterances. *Phonetica*, *42*(2-3), 124–134.

O'Shaughnessy, D. (1981). A study of French vowel and consonant durations. *Journal of Phonetics*, *9*(406), 355.

Osser, H., & Peng, F. (1964). A cross-cultural study of speech rate. *Language and Speech*, *7*, 120–125.

Ostendorf, M. (1993). Prosody. *In Proceedings of Human Language Technology. Plainsboro, San Mateo*, 315–316.

Ostendorf, M., & Veilleux, N. (1994). A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Computational Linguistics*, *20*(1), 27–54.

Ouden, H., Noordman, L., & Terken, J. (2009). Prosodic realizations of global and local structure and rhetorical relations in read aloud news reports. *Speech Communication*,

51 (2), 116–129.

Paeschke, A., Kienast, M., & Sendlmeier, W. (1999). F0-contours in emotional speech. *Proc. 14th Int. Congress of Phonetic Sciences*, *2*.

Paeschke, A., & Sendlmeier, W. (2000). Prosodic characteristics of emotional speech : Measurements of fundamental frequency movements. *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

Pasdeloup, V. (1990). Modèle de règles rythmiques du français appliqué à la synthèse de parole. *Thèse de Doctorat, Université de Provence, 386p*.

Pasdeloup, V. (2004). Le rythme n'est pas élastique : étude préliminaire de l'influence du débit de parole sur la structuration temporelle. *Actes des Journées d'Etudes sur la Parole, Fés*.

Pasdeloup, V., Espesser, R., & Faraj, M. (2006). Sensibilité au débit et marquage accentuel des phonèmes en français. *26èmes Journées d'Etudes sur la Parole*, 251–254.

Patil, U., Kentner, G., Gollrad, A., Kügler, F., Fery, C., & Vasishth, S. (2008). Focus, word order, and intonation in Hindi. *Journal of South Asian Linguistics*, *1* (1), 55–72.

Patterson, D. (2000). A Linguistic Approach to Pitch Range Modelling. *Doctoral dissertation, University of Edinburgh*.

Patterson, D., & Ladd, D. (1999). Pitch range modelling : linguistic dimensions of variation. *Proceedings of ICPhS*, *99*, 1169–1172.

Pellegrino, F., Farinas, J., & Rouas, J. (2004). Automatic estimation of speaking rate in multilingual spontaneous speech. *Speech Prosody 2004, International Conference*, 517–520.

Peterson, G., & Lehiste, I. (1960). Duration of syllable nuclei in English. *Journal of the Acoustical Society of America*, *32*(6), 693–703.

Petrone, C. (2008). *La définition des cibles tonales dans l'italien de Naples*. Thèse de doctorat.

Pfau, T., Faltlhauser, R., & Ruske, G. (2000). A combination of speaker normalization and speech rate normalization for automatic speech recognition. *Sixth International Conference on Spoken Language Processing*.

Pfitzinger, H. (1998). Local speech rate as a combination of syllable and phone rate. *Fifth International Conference on Spoken Language Processing*.

Pfitzinger, H. (2002). Intrinsic phone durations are speaker-specific. *Seventh International Conference on Spoken Language Processing*, 1113–1116.

Pfitzinger, H., Burger, S., & Heid, S. (1996). Syllable detection in read and spontaneous speech. *Proceedings of the Fourth International Conference on Spoken Language, ICSLP 96*, *2*, 1261–1264.

Pfitzinger, H., & Tamashima, M. (2006). Comparing Perceptual Local Speech Rate of German and Japanese Speech. *Proc. of the 3rd Int. Conf. on Speech Prosody*, *1*, 105–108.

Pickering, B., Williams, B., & Knowles, G. (1996). Analysis of transcriber differences in SEC. *Working with speech, London : Longman*, 61–86.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation.* Massachusetts Institute of Technology, 401 p.

Pierrehumbert, J., & Beckman, M. (1988). Japanese Tone Structure. *Linguistic Inquiry Monographs*(15), 1–282.

Pierrehumbert, J., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. *Intentions in Communication. MIT Press, Cambridge MA.*

Pierrehumbert, J., & Talkin, D. (1992). Lenition of/h/and glottal stop. *Papers in laboratory phonology II : Gesture, segment, prosody*, 90–117.

Pijper, J., & Sanderman, A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *The Journal of the Acoustical Society of America*, *96*, 2037–2047.

Pijper, J. de. (1983). *Modelling British English Intonation - Chapter III.* Walter de Gruyter.

Pike, K. (1945). *The Intonation of American English.* University of Michigan Press, Ann Arbor, Mich. 48106.

Pike, K. (1948). A Technique for Determining the Number and Type of Pitch Contrasts in a Language, with Studies in Tonemic Substitution and Fusion. *Ann Arbor : The University of Michigan Press*, p187.

Pinheiro, J., & Bates, D. (2000). *Mixed-effects models in S and S-PLUS.* Springer Verlag.

Portes, C., & Di Cristo, A. (2003). Pitch Range in spontaneous speech : semi-automatic approach versus subjective judgement. *15th International Conference on Phonetic Sciences, Barcelone, Spain.*

Post, B. (2000). *Tonal and phrasal structures in French intonation.* Doctoral dissertation.

Prince, A. (1983). Relating to the Grid. *Linguistic inquiry*, *14*(1), 19–100.

Prince, E. (1981). Toward a taxonomy of given-new information. *Radical pragmatics*, 223–255.

Rabiner, L., Cheng, M., Rosenberg, A., & McGonegal, C. (1976). A comparative performance study of several pitch detection algorithms. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *24*(5), 399–418.

Rakerd, B., Sennett, W., & Fowler, C. (1987). Domain-final lengthening and foot-level shortening in spoken English. *Phonetica*, *44*(3), 147–155.

Reissland, N., & Snow, D. (1996). Maternal pitch height in ordinary and play situations. *J Child Lang*, *23*(2), 269–78.

Rietveld, A., & Gussenhoven, C. (1985). On the relation between pitch excursion size and prominence. *Proceedings. Department of Language and Speech, Phonetics Section, University of Nijmegen*(9), 40–47.

Rietveld, A., & Gussenhoven, C. (1987). Perceived speech rate and intonation. *Journal of Phonetics*, *15*, 273–285.

Rietveld, T., & Vermillion, P. (2003). Cues for Perceived Pitch Register. *Phonetica*, *60*(4), 261–272.

Roach, P. (1998). Some languages are spoken more quickly than others. *Language myths*,

150–158.

Robb, M., Maclagan, M., & Chen, Y. (2004). Speaking rates of American and New Zealand varieties of English. *Clinical linguistics and phonetics*, *18*(1), 1–15.

Rose, P. (1987). Considerations on the normalization of the fundamental frequency of linguistic tone. *Speech Communication*, *6*(4), 343–351.

Rossi, M. (1999). *L'intonation, le système du français. Description et modélisation.* Coll. L'Essentiel. Paris, France : Editions Ophrys.

Rouas, J., Farinas, J., & Pellegrino, F. (2004). Evaluation automatique du débit de la parole sur des données multilingues spontanées. *XXVe Journees d'Etude sur la Parole (JEP 2004), Fes, Maroc*, 437–440.

Russell, A., Penny, L., & Pemberton, C. (1995). Speaking fundamental frequency changes over time in women : a longitudinal study. *Journal of speech and hearing research*, *38*(1), 101.

Saint-Bonnet, M., & Boe, J. (1977). Les pauses et les groupes rythmiques : leur durée et disribution en fonction de la vitesse d'élocution. *8èmes Journées d'Etude sur la Parole*, 337–343.

Savino, M., & Grice, M. (2007). The role of pitch range in realising pragmatic contrasts - the case of two question types in italian. *in ICPhS XVI Saarbrucken, 6-10 August.*, 1037–1040.

Scherer, K. (1979). *Personality markers in speech.* Cambridge Univ. Press, 395 p.

Scherer, K. (2003). Vocal communication of emotion : A review of research paradigms. *Speech Communication*, *40*(1-2), 227–256.

Scherer, K., Ladd, D., & Silverman, K. (1984). Vocal cues to speaker affect : Testing two models. *The Journal of the Acoustical Society of America*, *76*, 1346–1356.

Scherer, K., & Wallbott, H. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, *66*(2), 310–328.

Schroder, M. (2003). Speech and Emotion Research : An overview of research frameworks and a dimensional approach to emotional speech synthesis.

Selkirk, E. (1980). The Role of Prosodic Categories in English Word Stress. *Linguistic inquiry*, *11*(3), 563–605.

Selkirk, E. (1984). *Phonology and syntax : The relation between sound and structure.* Current Studies in Linguistics I., 476 p.

Shattuck-Hufnagel, S., & Turk, A. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of psycholinguistic research*, *25*(2), 193–247.

Shriberg, E., Ladd, D., Terken, J., Int, S., & Park, M. (1996). Modeling intra-speaker pitch range variation : predicting F0 targets when « speaking up ». *Proceedings of the Fourth International Conference on Spoken Language, ICSLP 96*, *2*.

Silverman, K. (1986). Fo segmental cues depend on intonation : The case of the rise after

voiced stops. *Phonetica*, *43*(1-3), 76–91.

Silverman, K. (1987). *The structure and processing of fundamental frequency contours.* University of Cambridge.

Silverman, K., Blaauw, E., Spitz, J., & Pitrelli, J. (1992). Towards using prosody in speech recognition/understanding systems : Differences between read and spontaneous speech. *Proceedings, Fifth DARPA Workshop on Speech and Natural Language*, 435–440.

Sityaev, D., Webster, G., Braunschweiler, N., Buchholz, S., & Knill, K. (2007). Some aspects of prosody of friendly formal and friendly informal speaking styles. *XVI International Conference on Phonetics Sciences, Saarbrücken, 6-10 August 2007*.

Sluijter, A., & Terken, J. (1992). The development and perceptive evaluation of a model for paragraph intonation in Dutch. *Second International Conference on Spoken Language Processing*, 353–356.

Sluijter, A., & Terken, J. (1993). Beyond sentence prosody : paragraph intonation in Dutch. *Phonetica*, *50*(3), 180–188.

Smith, B., Brown, B., Strong, W., & Rencher, A. (1975). Effects of speech rate on personality perception. *Language and Speech*, *18*(2), 145–152.

Smith, C. (2005). Durational prosody and topic organization : differences between English and French. *IDP Symposium on Discourse - Prosody Interface, Aix-En-Provence.*.

Smith, P. (1979). Sex markers in speech. *Social markers in speech*, 109–146.

Snider, K., & Hulst, H. van der. (1992). Issues in the representation of tonal register. *The phonology of tone : The representation of tonal register*, 1–27.

Sobell, L., & Sobell, M. (1972). Effects of alcohol on the speech of alcoholics. *Journal of Speech, Language and Hearing Research*, *15*(4), 861–868.

Son, R., & Pols, L. (1989). Comparing formant movements in fast and normal rate speech. *First European Conference on Speech Communication and Technology*, 2665–2668.

Sorensen, J., & Cooper, W. (1980). Syntactic coding of fundamental frequency in speech production. *Perception and production of fluent speech*, 399–440.

Stewart, J. (1993). Dschang and Ebrié as Akan-type total downstep languages. *The Phonology of Tone–the representation of tonal register*, 185–244.

Stewart, J., Schachter, P., & Welmers, W. (1964). *The typology of the Twi tone system.* Institute of African Studies, University of Ghana.

Stoicheff, M. (1981). Speaking fundamental frequency characteristics of nonsmoking female adults. *Journal of Speech and Hearing Research*, *24*(3), 437–441.

Strangert, E. (2003). Emphasis by pausing. *Proc. 15th ICPhS, Barcelona*.

Swerts, M. (1997). Prosodic features at discourse boundaries of different strength. *The Journal of the Acoustical Society of America*, *101*, 514–521.

Swerts, M., Bouwhuis, D., & Collier, R. (1994). Melodic cues to the perceived « finality » of utterances. *The Journal of the Acoustical Society of America*, *96*, 2064–2075.

Swerts, M., & Geluykens, R. (1993). The prosody of information units in spontaneous mono-

logue. *Phonetica*, *50*(3), 189–196.

Swerts, M., & Geluykens, R. (1994). Prosody as a marker of information flow in spoken discourse. *Language and speech*, *37*(1), 21–43.

Swerts, M., Strangert, E., & Heldner, M. (1996). F0 declination in read-aloud and spontaneous speech. *Proceedings of the Fourth International Conference on Spoken Language, 1996. ICSLP 96*, *3*, 20–24.

Takamaru, K., Hiroshige, M., Araki, K., & Tochinai, K. (2000). A Proposal of a Model to Extract Japanese Voluntary Speech Rate Control. *Sixth International Conference on Spoken Language Processing*.

Takefuta, Y., Jancosek, E., & Brunt, M. (1972). A statistical analysis of melody curves in the intonation of American English. *Proceedings of the Seventh International Congress of Phonetic Sciences, 1971*.

Terken, J., & Collier, R. (1989). Automatic Synthesis of Natural-Sounding Intonation for Text-to-Speech Conversion in Dutch. *First European Conference on Speech Communication and Technology*, 1357–1359.

Thorsen, N. (1985). Intonation and text in Standard Danish. *Journal of the Acoustical Society of America*, *77*(3), 1205–1216.

Tickle, A. (2000). English and Japanese speakers' emotion vocalisation and recognition : A comparison highlighting vowel quality. *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

Trager, G., & Smith, J. (1957). *An outline of English structure*. Washington : American Council of Learned Societies.

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, *88*, 97–100.

Traunmüller, H., & Eriksson, A. (1994). The frequency range of the voice fundamental in the speech of male and female adults. *Manuscript, Department of Linguistics, University of Stockholm*.

Traunmüller, H., & Eriksson, A. (1995). The perceptual evaluation of F excursions in speech as evidenced in liveliness estimations. *The Journal of the Acoustical Society of America*, *97(3)*, 1905–1915.

Trouvain, J. (2004). *Tempo Variation in Speech Production : Implications for Speech Synthesis*. Institut für Phonetik, Universität des Saarlandes.

Trouvain, J., & Grice, M. (1999). The effect of tempo on prosodic structure. *Proc. 14th Intern. Confer. Phonetic Sciences*, 1067–1070.

Truckenbrodt, H. (2002). Upstep and embedded register levels. *Phonology*, *19*(01), 77–120.

Tsao, Y., & Weismer, G. (1997). Interspeaker Variation in Habitual Speaking Rate Evidence for a Neuromuscular Component. *Journal of Speech, Language and Hearing Research*, *40*(4), 858–866.

Tseng, C., Chang, C.-H., & Su, Z. (2005). Investigating F0 reset and range in Relation to

Fluent Speech Prosody Hierarchy. *Technical Acoustics, vol 24*, 279–284.

Turk, A., & Sawusch, J. (1997). The domain of accentual lengthening in American English. *Journal of Phonetics*, *25*(1), 25–41.

Turk, A., & Shattuck-Hufnagel, S. (2000). Word-boundary-related duration patterns in English. *Journal of Phonetics*, *28*(4), 397–440.

Turk, A., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics*, *35*(4), 445–472.

Turk, A., & White, L. (1999). Structural influences on accentual lengthening in English. *Journal of Phonetics*, *27*(2), 171–206.

Umeda, N. (1982). Boundary : perceptual and acoustic properties and syntactic and statistical determinants. *Speech and Language*, *7*, 333–371.

Vaane, E. (1982). Subjective Estimation of Speech Rate. *Phonetica Basel*, *39*(2-3), 136–149.

Vainio, M., Hirst, D., Suni, A., & De Looze, C. (2009). Using functional annotation for high quality multilingual, multidialectal and multistyle speech synthesis.

Vaissiere, J. (1988). The use of prosodic parameters in automatic speech recognition. *Recent Advances in Speech Understanding and Dialog Systems (eds) Niemann, H., Lang, M. and Sagerer, G*.

Vaissière, J. (sous presse). Les universaux de substance prosodiques. *Sophie Wauquier (éd.), Les universaux sonores. Rennes : Presses Universitaires de Rennes*.

Vaissière, J. (1983). Language-independent prosodic features. *Prosody : Models and measurements*, 53–66.

Vaissière, J. (1991). Rhythm, accentuation and final lengthening in French. *J.Sundberg, L. Nord and R. Carlson [Ed], Music, language, speech and brain, 59, Stockholm : Wennergren, International Symposium series*, 108–120.

Vaissière, J. (2002). Cross-linguistic prosodic transcription : French versus English. *Problems and methods of experimental phonetics*, 147–164.

Vaissière, J. (2005). Perception of Intonation. *Pisoni, David B. ; Remez, Robert E. (eds), The Handbook of Speech Perception*.

Van Bezooijen, R. (1995). Sociocultural aspects of pitch differences between Japanese and Dutch women. *Language and Speech*, *38*(3), 253–265.

Van Den Berg, R., Gussenhoven, C., & Rietveld, T. (1992). Downstep in Dutch : Implications for a model. *Papers in Laboratory Phonology II : Gesture, Segment, Prosody*, 335–359.

Verhasselt, J., & Martens, J. (1996). A fast and reliable rate of speech detector. *Fourth International Conference on Spoken Language Processing*.

Verhoeven, J. (2002). The communicative setting and markers in speech. *Phonetic work in progress*, 177–190.

Verhoeven, J., De Pauw, G., & Kloots, H. (2004). Speech rate in a pluricentric language : A comparison between Dutch in Belgium and the Netherlands. *Language and Speech*, *47*(3), 297–308.

Walker, J., Archibald, L., Cherniak, S., & Fish, V. (1992). Articulation rate in 3-and 5-year-old children. *Journal of Speech and Hearing Research*, *35*(1), 4–13.

Wang, B., & Xu, Y. (2006). Prosodic encoding of topic and focus in Mandarin. *Proceedings of Speech Prosody 2006*, 1–4.

White, L. (2002). English speech timing : a domain and locus approach. *University of Edinburgh PhD dissertation*.

Whiteside, S., & Hodgson, C. (2000). Speech patterns of children and adults elicited via a picture-naming task : an acoustic study. *Speech Communication*, *32*(4), 267–285.

Whitmore, J., & Fisher, S. (1996). Speech during sustained operations. *Speech Communication*, *20*(1-2), 55–70.

Wichmann, A. (2000). The attitudinal effects of prosody, and how they relate to emotion. *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, *91*(3), 1707–1717.

Williams, B. (1996). The formulation of an intonation transcription system for British English. *Working with Speech : perspectives on research into the Lancaster/IBM Spoken English Corpus*, 38–57.

Williams, C., & Stevens, K. (1972). Emotions and speech : Some acoustical correlates. *The Journal of the Acoustical Society of America*, *52*, 1238–1250.

Woo, N. (1969). *Prosody and phonology*. Massachusetts Institute of Technology.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f0contours. *Journal of Phonetics*, *27*(1), 55–105.

Xu, Y., & Xu, C. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics*, *33*(2), 159–197.

Xu, Y., Xu, C., & Sun, X. (2004). On the temporal domain of focus. *Speech Prosody 2004, International Conference*, 81–84.

Yamazawa, H., & Hollien, H. (1992). Speaking fundamental frequency patterns of Japanese women. *Phonetica*, *49*(2), 128–140.

Yoon, T., Cole, J., & Hasegawa-Johnson, M. (2007). On the edge : Acoustic cues to layered prosodic domains. *Proceedings of ICPhS*.

Yule, G. (1980). Speakers'topics and major paratones. *Lingua Amsterdam*, *52*(1-2), 33–47.

Zellner, B. (1994). Pauses and the temporal structure of speech. *Fundamentals of speech synthesis and speech recognition*, 41–62.

Zellner, B. (1996). Structures temporelles et structures prosodiques en français lu. *Revue Française de Linguistique Appliquée : La communication parlée*, *1*, 7–23.

Zellner, B. (1998). Caractérisation et prédiction du débit de parole en français. Une étude de cas. *Thèse présentée à la Faculté de Lettres de l'Université de Lausanne*.

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (Frequenz-

gruppen). *The Journal of the Acoustical Society of America*, *33*, 248.

Zwicker, E., Flottorp, G., & Stevens, S. (1957). Critical band width in loudness summation. *The Journal of the Acoustical Society of America*, *29*, 548–557.

Résumé: Cette thèse a pour objectif l'étude de l'empan temporel des variations de registre et de tempo et des fonctions qu'elles revêtent, en anglais et en français. Elle s'ancre dans une des problématiques majeures de l'étude des variations prosodiques, celle de leur délimitation, à plus ou moins long terme, de leur interaction et de leur chevauchement et par là, de la difficulté à les analyser séparément. L'auteur y défend une structure emboîtée des variations de registre et de tempo et ainsi l'idée de variations à plusieurs niveaux et qui opèrent sur divers domaines. Y est aussi examinée la façon dont ces variations informent de l'identité du locuteur ou encore la façon dont elles indiquent la structure intentionnelle du discours.

Mots-clés: *Variations prosodiques - Empan temporel - Registre - Tempo - Mesure - Fonctions - Français - Anglais*

Abstract: This thesis aims at studying the temporal span of register and tempo variations and the functions they convey, in English and French. It tackles one of the major issues in the study of the sources of prosodic variability, that of their more or less long term delimitation, their interaction and their overlapping, hence the difficulty of their distinction. The author argues for an embedded structure of variations of register and tempo, hence the idea of multi-level and multi-domain variations. The author also examines the way these variations inform about a speaker's identity and the way they indicate the intentional structure of discourse.

Keywords: *Prosodic variations - Temporal span - Register - Tempo - Measure - Functions - French - English*

UFR LACS (Lettres, Arts, Communication et Sciences du Langage)

Laboratoire Parole et Langage

CNRS UMR 6057 / Université de provence

5, Avenue Pasteur 13100

Aix-en-provence, France